Lipreading With Local Spatiotemporal Descriptors

Guoying Zhao, Mark Barnard, and Matti Pietikäinen, *Senior Member, IEEE*

***Abstract*—Visual speech information plays an important role in lipreading under noisy conditions or for listeners with a hearing impairment. In this paper, we present local spatiotemporal descriptors to represent and recognize spoken isolated phrases based solely on visual input. Spatiotemporal local binary patterns extracted from mouth regions are used for describing isolated phrase sequences. In our experiments with 817 sequences from ten phrases and 20 speakers, promising accuracies of 62% and 70% were obtained in speaker-independent and speaker-dependent recognition, respectively. In comparison with other methods on AVLetters database, the accuracy, 62.8%, of our method clearly outperforms the others. Analysis of the confusion matrix for 26 English letters shows the good clustering characteristics of visemes for the proposed descriptors. The advantages of our approach include local processing and robustness to monotonic gray-scale changes. Moreover, no error prone segmentation of moving lips is needed.**

***Index Terms*—Lipreading, local binary patterns, spatiotemporal descriptors, visual speech recognition.**

I. Introduction

IT is well known that human speech perception is a multimodal process. Visual observation of the lips, teeth, and tongue offers important information about the place of pronunciation articulation. A human listener can use visual cues, such as lip and tongue movements, to enhance the level of speech understanding. The process of using visual modality is often referred to as lipreading which is to make sense of what someone is saying by watching the movement of his lips. In some research, lipreading combined with face and voice is studied to help biometric identification [4], [12], [13], [21]. There is also a lot of work focusing on audio-visual speech recognition (AVSR) [2], [3], [5]–[7], [11], [14], [15], [18], [26], [27], [29], [32], trying to find effective ways of combining visual information with existing audio-only speech recognition systems (ASR). The McGurk effect [23] demonstrates that inconsistency between audio and visual information can result in perceptual confusion. Visual information plays an important role especially in noisy environments or for the listeners with hearing impairment.

Most of the research focuses on using visual information to improve speech recognition. Audio features are still the main contribution and play a more important role than visual features. However, in some cases, it is difficult to extract useful information from the audio. There are many applications in which it is necessary to recognize speech under extremely adverse acoustic environments. Detecting a person's speech from a distance or through a glass window, understanding a person speaking among a very noisy crowd of people, and monitoring a speech over a TV broadcast when the audio link is weak or corrupted are some examples. Furthermore, for people with hearing impairments, visual information is the only source of information from TV broadcasts or speeches, if there is no assisting sign language. In these applications, the performance of traditional speech recognition is very limited. There are a few works focusing on the lip movement representations for speech recognition solely with visual information [9], [24], [34], [35]. Saenko *et al.* [34], [35] use articulatory features and dynamic Bayesian network for recognizing spoken phrases with multiple loosely synchronized streams. Chiou and Hwang [9] utilize snakes to extract visual features from geometric space, Karhunen-Loeve transform to extract principal components in the color eigenspace and HMMs to recognize the isolated words. Matthews *et al.* [24] present two top-down approaches that fit a model of the inner and outer lip contours and derive lipreading features from a PCA of shape, or shape and appearance, respectively, and as well a bottom-up method which uses a nonlinear scale-space analysis to form features directly from the pixel intensity.

Comprehensive reviews of automatic audio-visual speech recognition can be found in [32] and [33]. Extraction of a discriminative set of visual observation vectors is the key element of an AVSR system. Geometric features, appearance features, and combined features are commonly used for representing visual information. Geometry-based representations include fiducial points like facial animation parameters [3], contours of lips [2], [26], [29], shape of jaw and cheek [2], [26], and mouth width, mouth opening, oral cavity area, and oral cavity perimeter [7]. 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. Appearance features are based on observing the whole mouth region-of-interest (ROI) as visually informative about the spoken utterance. The feature vectors are computed using all the video pixels within the ROI. The proposed approaches include principal component analysis (PCA) [5],

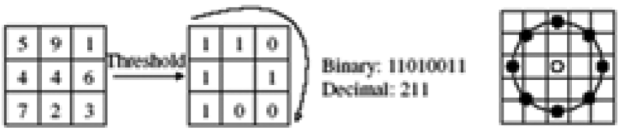
Manuscript received February 03, 2009; revised April 21, 2009. First published August 18, 2009; current version published October 16, 2009. This work was supported by the Academy of Finland. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shrikanth Narayanan.

G. Zhao and M. Pietikäinen are with the Machine Vision Group, Infotech Oulu and Department of Electrical and Information Engineering, University of Oulu, Oulu FI-90014, Finland (e-mail: gyzhao@ee.oulu.fi; mkp@ee.oulu.fi).

M. Barnard is with the Machine Vision Group, University of Oulu, Oulu FI-90014, Finland, and also with the Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: Mark.Barnard@surrey.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2009.2030637



(a) (b)

Fig. 1. (a) Basic LBP operator. (b) Circular (8,2) neighborhood.

[6], the discrete cosine transform (DCT) [31], or a combination of these transforms [14], [34], [35].

In addition, features from both categories can be combined for lip localization and visual feature extraction [9], [26], [27]. It appears that most of the research on visual speech recognition based on the appearance features has considered global features of lip or mouth images but omitted the local features. Local features can describe the local changes of images in space and time. In this paper, we propose an approach for lipreading, i.e., visual speech recognition, which could improve the human-computer interaction and understanding especially in noisy environments or for listeners with hearing impairments. A preliminary version of this work was presented in [40]. We focus on the recognition of isolated phrases using only visual information. A new appearance feature representation based on spatiotemporal local binary patterns is proposed, taking into account the motion of mouth region and time order in pronunciation. A support vector machine (SVM) classifier is utilized for recognition. Spatiotemporal multiresolution descriptors are introduced, and feature selection using AdaBoost to select more important slices (principal appearance and motion) is also presented. Experiments on different databases are carried out for performance analysis. Section II presents the spatiotemporal descriptors for mouth movement, and the multiresolution features and feature selection method are described in Section III. In Section IV, the whole system is introduced, and experiments are presented in Section V. Section VI concludes the paper.

II. Local Spatiotemporal Descriptors for Visual Information

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 [30]. For each pixel in an image, a binary code is produced by thresholding its neighborhood with the value of the center pixel [Fig. 1(a) and (1)]:

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

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, as shown in Fig. 1(b). In this way, one can collect larger-scale texture primitives or micro-patterns, like lines, spots, and corners [30].

"Uniform patterns" [30] are usually used to shorten the length of the feature vector of LBP. Here, a pattern is considered uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is considered circular (e.g., 11110011. However, 1000100 is not a uniform pattern since it contains four bitwise transitions). When using the uniform patterns, all non-uniform LBP patterns are collected into a single bin during the histogram computation. In the following sections, "u2" is utilized to refer to uniform patterns.

Local texture descriptors have gained increasing attention in facial image analysis due to their robustness to challenges such as pose and illumination changes. Ahonen *et al.* proposed LBP-based facial representation for face recognition from static images [1].

Recently, a method for temporal texture recognition using spatiotemporal local binary patterns extracted from three orthogonal planes (LBP-TOP) was proposed [39]. With this approach, the ordinary LBP for static images was extended to the spatiotemporal domain. For LBP-TOP, the radii in spatial and temporal axes X, Y, and T, and the number of neighboring points in the XY, XT, and YT planes can also be different, which can be marked as R_X, R_Y and R_T, P_{XY}, P_{XT} and P_{YT}; the corresponding LBP-TOP feature is then denoted as $LBP - TOP_{P_{XY},P_{XT},P_{YT},R_X,R_Y,R_T}$. Suppose the coordinates of the center pixel $g_{t_c,c}$ are (x_c, y_c, t_c), the coordinates of local neighborhood in XY plane $g_{XY,p}$ are given by $(x_c - R_X \sin(2\pi p/P_{XY}), y_c + R_Y \cos(2\pi p/P_{XY}), t_c)$, the coordinates of local neighborhood in XT plane $g_{XT,p}$ are given by $(x_c - R_X \sin(2\pi p/P_{XT}), y_c, t_c - R_T \cos(2\pi p/P_{XT}))$, and the coordinates of local neighborhood in YT plane $g_{YT,p}$ $(x_c, y_c - R_Y \cos(2\pi p/P_{YT}), t_c - R_T \sin(2\pi p/P_{YT}))$. This is different from the ordinary LBP widely used in many papers, and it extends the definition of LBP. A histogram is created to represent the occurrences of different binary patterns in these three planes. Spatial information such as appearance is captured in the XY plane and temporal information such as horizontal or vertical motion is captured in the XT and YT planes, respectively. Sometimes, the radii in three axes are the same and so do the number of neighboring points in XY, XT, and YT planes. In that case, we use $LBP - TOP_{P,R}$ for abbreviation where $P = P_{XY} = P_{XT} = P_{YT}$ and $R = R_X = R_Y = R_T$. The length or dimension of the $LBP - TOP_{P,R}$ features is 3×2^P. Moreover, region-concatenated descriptors using LBP-TOP features were developed for facial expression recognition. The results obtained with the Cohn–Kanade facial expression database outperformed the state-of-the-art.

Due to its ability to describe spatiotemporal signals, robustness to monotonic gray-scale changes caused, e.g., by illumination variations, the LBP-TOP is utilized to represent the mouth movements in this paper. Considering the motion of the mouth region, the descriptors are obtained by concatenating local binary patterns on three orthogonal planes from the utterance sequence: XY, XT, and YT, considering only the co-occurrence statistics in these three directions. Fig. 2(a) demonstrates the

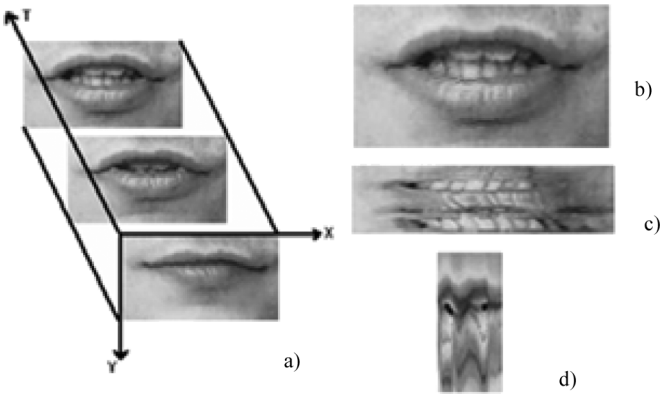


Fig. 2. (a) Volume of utterance sequence. (b) Image in XY plane (147 × 81). (c) Image in XT PLANE (147 × 38) in $y = 40$ (last row is pixels of $y = 40$ in first image). (d) Image in TY plane (38 × 81) in $x = 70$ (first column is the pixels of $x = 70$ in first frame).

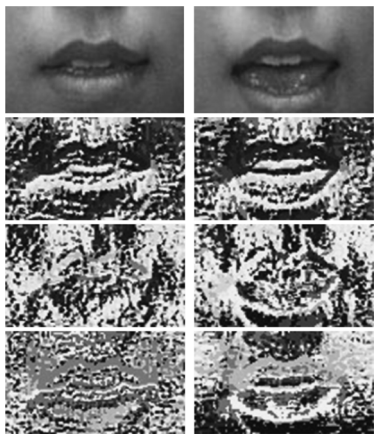


Fig. 3. Mouth region images (first row), LBP-XY images (second row), LBP-XT images (third row), and LBP-YT images (last row) from one utterance.

volume of utterance sequence. Fig. 2(b) shows image in the XY plane. Fig. 2(c) is an image in the XT plane providing a visual impression of one row changing in time, while Fig. 2(d) describes the motion of one column in temporal space. An LBP 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. 3 also gives some examples of the LBP images. The second, third, and fourth rows show the LBP images which are drawn using LBP code of every pixel from XY (second row), XT (third row), and YT (fourth row) planes, respectively, corresponding to mouth images in the first row. From this figure, the change in appearance and motion during utterance can be seen.

However, taking only into account the locations of micro-patterns is not enough. When a person utters a command phrase, the words are pronounced in order, for instance "you-see" or "see-you". If we do not consider the time order, these two phrases would generate almost the same features. To overcome

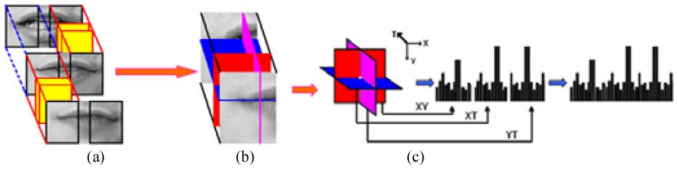


Fig. 4. Features in each block volume. (a) Block volumes. (b) LBP features from three orthogonal planes. (c) Concatenated features for one block volume with the appearance and motion.

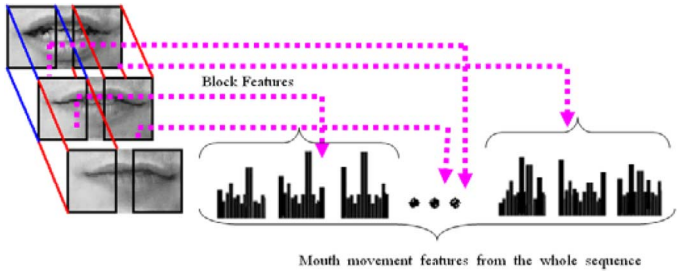


Fig. 5. Mouth movement representation.

this effect, the whole sequence is not only divided into block volumes according to spatial regions but also in time order, as Fig. 4(a) shows. The LBP-TOP histograms in each block volume are computed and concatenated into a single histogram, as Fig. 4 shows. All features extracted from each block volume are connected to represent the appearance and motion of the mouth region sequence, as shown in Fig. 5.

In this way, we effectively have a description of the phrase utterance on three different levels of locality. The labels (bins) in the histogram contain information from three orthogonal 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 of the appearance and motion in specific locations and time segments, and all information from the regional level is concatenated to build a global description of the mouth region motion. Moreover, even though different utterances have different length, they are divided into the same number of block volumes, so the lengths of their feature vectors are the same to compare.

A histogram of the mouth movements can be defined as

$$H_{b,c,d,j,i} = \sum_{x,y,t} I\{f_j(x,y,t) = i\}, \quad i = 0, \cdots, n_j - 1; j = 0, 1, 2 \tag{2}$$

in which n_j is the number of different labels produced by the LBP operator in the jth plane ($j = 0$: XY, 1: XT, *and* 2: YT), $f_j(x, y, t)$ expresses the LBP code of central pixel (x, y, t) in the jth plane, $x \in \{R_X, \cdots, X - 1 - R_X\}$, $y \in \{R_Y, \cdots, Y - 1 - R_Y\}$, $t \in \{R_T, \cdots, T - 1 - R_T\}$ (X and Y are width and height of image and T is the utterance length). b is the index of rows, c is of columns, and d is of time of block volume:

$$I\{A\} = \begin{cases} 1, & \text{if } A \text{ is true} \\ 0, & \text{if } A \text{ is false.} \end{cases} \tag{3}$$

The histograms must be normalized to get a coherent description:

$$N_{b,c,d,j,i} = \frac{H_{b,c,d,j,i}}{\sum_{k=0}^{n_j-1} H_{b,c,d,j,k}}. \quad (4)$$

III. Multiresolution Features and Feature Selection

Multiresolution features can provide more information and improve the analysis of dynamic events. Using multiresolution features, however, will also greatly increase the number of features available. If the features from different resolutions were concatenated directly, the feature vector would become very long, making the computational complexity too high. It is obvious that all multiresolution spatiotemporal features do not contribute equally, either. Therefore, it is necessary to find out what features (in which location, with what resolutions, and more importantly, what types: appearance, horizontal motion, or vertical motion) are more important. Feature selection is needed for this purpose.

In this section, we consider the use of spatiotemporal local binary patterns computed at multiple resolutions for describing dynamic events, combining static and dynamic information from different spatiotemporal resolutions. For a more complete description of this approach, see [41]. The whole video sequence can be divided into $B \times C \times D$ sub-volumes, and inside each sub-volume, the LBP-TOP features are computed to describe the characteristic of the sub-volume, and finally are connected together to represent the videos. In changing the parameters, three different types of spatiotemporal resolution are presented: 1) Use of a different number of neighboring points when computing the features in XY (appearance), XT (horizontal motion), and YT (vertical motion) slices; 2) Use of different radii that can capture the occurrences in different space and time scales; 3) Use of blocks of different sizes to create global and local statistical features. The first two resolutions focus on the pixel level in feature computation, providing different local spatiotemporal information, while the third one focuses on the block or volume level, giving more global information in the space and time dimensions.

Appearance and motion are the key components for visual speech analysis. The AdaBoost algorithm is utilized for learning the principal appearance and motion from spatiotemporal descriptors derived from three orthogonal slices (slice-based method), providing important information about the locations and types of features for further analysis. Our approach is unlike earlier work [16], [38] (block-based method), in which just the importance of block or location was considered, missing the detailed appearance and motion information. To keep the global description with histograms, and at the same time, to separate the appearance and motions, every slice histogram is thought as an element. To get the slice similarity within class and diversity between classes, we compare every slice histogram from different samples with same multiresolution parameters. The similarity values are used as the new features.

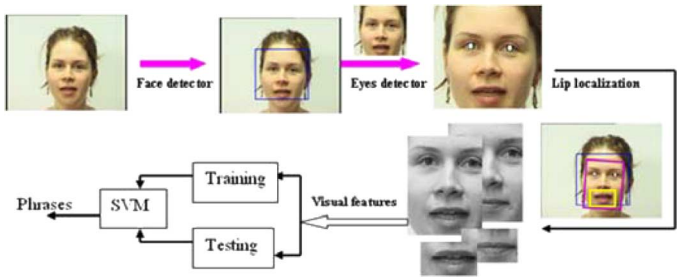


Fig. 6. System diagram.

Several possible dissimilarity measures are available. In this work, Chi square statistic (χ^2) defined below is adopted:

$$\chi^2(S,M) = \sum_{ii=0}^{n-1} \frac{(S_{ii} - M_{ii})^2}{(S_{ii} + M_{ii})} \quad (5)$$

where S and M are two slice histograms and n is the bin number of the histogram. The whole feature pool contains an enormous amount of possible features because of the highly overcomplete representation (each feature prototype can appear at different position, scale, and in any type). $\{\chi^2_{a,b}(XY), \chi^2_{a,b}(XT), \chi^2_{a,b}(YT)\}$ are the similarity of the LBP-TOP features in three slices from samples a and b, and used as the new features fed into learners. Here, a and b are the indexes of samples. They could come from the same class which would be the intra-class features, or different classes which would be the extra-class features. In this way, the dissimilarity for three kinds of slices are obtained, which can further be used to describe the importance of appearance, horizontal motion, and vertical motion.

In addition, learners are designed for selecting the most important features for each specific pair of speech classes [41]. Previous work [16], [38] employed an all-against-all (All-All) approach to AdaBoost learning. This approach determined the global variations between all classes. However, if we could determine the specific differences between each pair, it would be helpful to improve the further analysis. To deal with this problem, we propose to use the class-pair learning, also called one-against-one (One-One) learning. That means the learners are designed for every pair of two classes and the aim is to learn more specific and discriminative features for each pair.

IV. Our System

Our system consists of three stages, as shown in Fig. 6. The first stage is a combination of discriminative classifiers that first detects 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 the input utterance using an SVM classifier.

Boosted Haar features [37] are used for automatic coarse face detection and 2-D Cascaded AdaBoost [28] 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 predefined ratio parameters [40] for the whole sequence.

For recognition, an 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. Since the SVM is only used for separating two sets of points, the n-phrase classification problem is decomposed into $n(n-1)/2$ two-class problems, then a voting scheme is used to accomplish recognition. Here, after the comparison of linear, polynomial, and RBF kernels in experiments, we use the second degree polynomial kernel function, which provided the best results. Sometimes more than one class gets the highest number of votes; in this case, 1-NN template matching is applied to these classes to reach the final result. This means that in training, the spatiotemporal LBP histograms of utterance sequences belonging to a given class are averaged to generate a histogram template for that class. In recognition, a nearest-neighbor classifier is adopted.

V. Experiments

A. Databases

1) OuluVS Database: In contrast to the abundance of audio-only corpora, there exist only a few databases suitable for visual or audio-visual ASR research. The audio-visual datasets commonly used in literature include [17], [18], [25], [27], [32], and [36].

A variety of audio-visual corpora have been created in order to obtain experimental results for specific tasks. Many of the them contain recordings of only one subject, e.g., [3] and [34]. Even those with multiple subjects are usually limited to small tasks such as isolated digits [5], or a short list of fixed phrases [25]. The M2VTS database and the expanded XM2VTSDB [25] are geared more toward person authentication, even though they consist of 37 and 295 subjects, respectively. Only two of the audio-visual corpora published so far (including English, French, German, and Japanese) contain both a large vocabulary and a significant number of subjects. One of these is IBM's proprietary, 290-subject, large-vocabulary AV-ViaVoice database of approximately 50 h in duration [27]. The other one is the VidTIMIT database [36], which consists of 43 subjects each reciting the ten different TIMIT sentences. It has been used in multimodal person verification research.

There are few datasets providing phrase data [17], [25], [34], [36], and in those, the number of speakers is pretty small [34]. Though AVTIMIT [17], XM2VTSDB [25], and VidTIMIT [36] include many speakers, the speakers utter different sentences or phrases [17], [36] or small number of sentences [25]. Due to the lack of publicly available databases suitable for our needs, we collected our own visual speech dataset, i.e., OuluVS database, for performance evaluation.

A SONY DSR-200AP 3CCD-camera with a frame rate 25 fps was used to collect the data. The image resolution was 720 × 576 pixels. Our dataset includes 20 persons, each uttering ten everyday greetings one to five times. These short phrases are listed in Table I.

TABLE I
Phrases Included in the Dataset

C1	"Excuse me"	C6	"See you"
C2	"Goodbye"	C7	"I am sorry"
C3	"Hello"	C8	"Thank you"
C4	"How are you"	C9	"Have a good time"
C5	"Nice to meet you"	C10	"You are welcome"

Fig. 7. Mouth regions from the dataset.

The subjects were asked to sit on a chair. The distance between the speaker and the camera was 160 cm. He/she was then asked to read ten phrases which were written on a paper, each phrase one to five times. The data collection was done in two parts: the first from ten persons and four days later from the ten remaining ones. Seventeen males and three females are included, nine of whom wear glasses. Speakers are from four different countries, so they have different pronunciation habits including different speaking rates.

In total, 817 sequences from 20 speakers were used in the experiments.

Fig. 7 gives some examples of the mouth localization. The average size of the mouth image is around 120 × 70. We know that using a fixed ratio perfect mouth regions cannot always be obtained, so in the future, a combination of eye positions and mouth detection will be considered to get more accurate mouth regions.

2) AVLetters Database: The AVletters database [24] consists of three repetitions by each of ten speakers, five male, two of whom have moustaches, and five female, of the isolated letters A-Z, a total of 78 utterances. Speakers were prompted using an autocue that presented each of three repetitions of the alphabet in nonsequential, nonrepeating order. Each speaker was requested to begin and end each letter utterance with their mouth in the closed position. No head restraint was used, but speakers were provided with a close-up view of their mouth and asked not to move out of frame. The full face images were further cropped to a region of 80 × 60 pixels after manually locating the center of the mouth in the middle frame of each utterance. Each utterance was temporally segmented by hand using the visual data so that each utterance began and ended with the speaker's mouth in the closed position. Fig. 8 shows example images from the ten speakers. To make an unbiased comparison, we also carry out the experiments on this public database.

B. Experimental Protocol and Results

For comprehensive evaluation of our proposed method, we design different experiments, including speaker-independent, speaker-dependent, multiresolution, and one-against-one versus one-against-rest experiments on these two databases. We

Fig. 8. Example images from ten speakers.

also analyze the viseme confusion matrix to see the clustering ability of the proposed method.

1) Speaker-Independent Experiments: For the speaker-independent experiments, leave-one-speaker-out is utilized. In the testing procedure on OuluVS database, in each run, training was done on 19 speakers in the data set, while testing was performed on the remaining one. The same procedure was repeated for each speaker, and the overall results were obtained using M/N (M is the total number of correctly recognized sequences and N is the total number of testing sequences).

When extracting the local patterns, we take into account not only locations of micro-patterns but also the time order in articulation, so the whole sequence is divided into block volumes according to not only spatial regions but also time order.

According to tests, parameter values $P_{XY} = P_{XT} = P_{YT} = 8$, $R_X = R_Y = R_T = 3$, and an overlap ratio of 70% of the original non-overlapping block size were selected empirically. After experimenting with different block sizes, we chose to use $1 \times 5 \times 3$ (rows by columns by time segments) blocks in our experiments.

Fig. 9 shows the recognition results using three different features on OuluVS database. As expected, the result of the features from three planes is better than that just from the appearance (XY) plane which justifies the effectiveness of the feature combining appearance with motion. The features with $1\times5\times1$ block volumes omitted the pronunciation order, providing a lower performance than those with $1 \times 5 \times 3$ block volumes for almost all the tested phrases. It can be seen from Fig. 9 that the recognition rates of phrases "See you" (C6) and "Thank you" (C8) are lower than others because the utterances of these two phrases are quite similar, just different in the tongue's position. If we take those two phrases as one class, the recognition rate would be 4% higher.

We compared the recognition performance for automatic mouth localization to that obtained with hand-marked eye positions. The results are given in Table II, showing that automatic eye detection gave similar performance to the manual approach. The second row demonstrates the results from the combined features of two kinds of block features, which are a little higher than those from one kind of block features (first row). We also used the temporal derivatives [15] which means pixel-by-pixel differences between consecutive frames, optical flow features [22], and DCT features [27], [31], [34] which have been exploited in early research. The DCT features are first computed for every frame, while the temporal derivatives and optical flow features are computed for every two frames to get the frame-level features. The whole utterance sequence can also be divided into D segments in time axis, and the final features

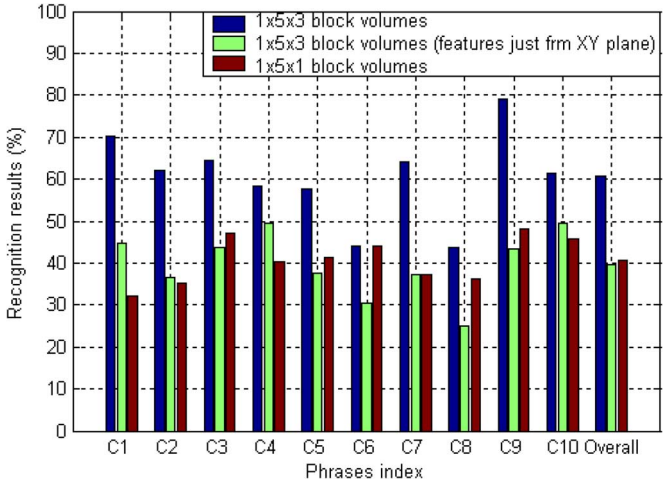


Fig. 9. Phrases recognition comparison of different features on OuluVS database.

TABLE II
RESULTS OF SPEAKER-INDEPENDENT EXPERIMENTS ON OULUVS DATABASE ("$S5$" FOR DCT: S MEANS SQUARE SUBLATTICES AND "5" MEANS THE LAYERS OF THE COEFFICIENTS SELECTION. SIMILAR MEANINGS FOR "$T10$", "$C6$", AND "$H6$". FOR DETAILS, PLEASE REFER TO [31])

Eye detection	Manual	Automatic
Blocks $(1 \times 5 \times 3)$	60.6%	58.6%
Blocks $(1 \times 5 \times 3 + 1 \times 5 \times 2)$	**62.4%**	**59.6%**
Temporal Derivatives: $D = 4$	———	28.03%
Optical Flow: $D = 5$	———	27.17%
DCT: $S5, D = 5$	———	37.09%
DCT: $T10, D = 5$	———	35.37%
DCT: $C6, D = 5$	———	34.88%
DCT: $H6, D = 5$	———	36.96%

are obtained by averaging the frame-level features through the segment. This is to keep the pronunciation order. The results for automatically localized mouth regions are listed in Table II. We have experimented using different parameters, and here we only list the best accuracy for temporal derivatives and optical flow features, and some DCT results with different number of coefficients, the lattice selection (S: square, T: triangular, C: circular, or H: hyperbolic sublattices), and the number of time segments. (For DCT, here we list the best accuracies for respective sublattice in our experiment.) From Table II, we can see our features perform much better than these features.

On AVLetters database, in each run, training was done on nine speakers in the data set, while testing was performed on the remaining one. The same procedure was repeated for each individual test speaker.

Fig. 10 demonstrates the performance for every speaker. As we can see, the results from the second speaker are the worst, mainly because the big moustache of that speaker (as shown in Fig. 8) really influences the appearance and motion in the mouth region.

Table III lists the accuracy from different parameters. The uniform features with neighborhood samples number eight and radius three extracted from blocks $2 \times 5 \times 3$ got the best result 43.46%, which is even comparable to the best accuracy 44.6% from the semi-speaker-dependent evaluation in [24]. Normal features even with longer feature vectors do not work as well

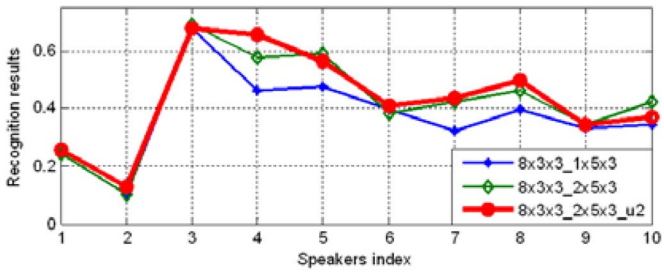


Fig. 10. Recognition performance for every speaker.

TABLE III
RESULTS OF SPEAKER-INDEPENDENT EXPERIMENTS ON AVLETTERS DATABASE (U2 REFERS TO UNIFORM PATTERNS)

Features:u2	Blocks	Results(%)
$LBP-TOP_{8,3}$	$2\times5\times3$	**43.46**
$LBP-TOP_{8,3}$	$1\times5\times3$	40.26
$LBP-TOP_{8,3}$	$2\times5\times2$	40.77
$LBP-TOP_{8,3}$	$2\times5\times4$	41.15
$LBP-TOP_{8,1}$	$1\times5\times3$	32.44
$LBP-TOP_{8,1}$	$2\times5\times3$	36.41
$LBP-TOP_{4,3}$	$1\times5\times3$	32.95
$LBP-TOP_{4,3}$	$2\times5\times3$	37.69

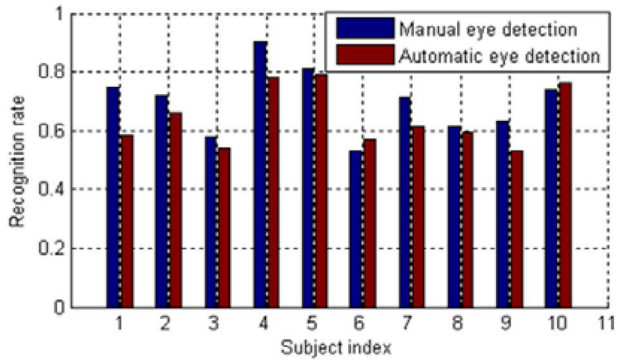


Fig. 11. Speaker-dependent recognition results for every subject on OuluVS database.

as the uniform patterns. The radius with three and neighboring points with eight outperform the radius with one and neighboring points four, which is consistent to the results from facial expression recognition.

2) Speaker-Dependent Experiments: For speaker-dependent experiments, the leave-one utterance-out is utilized for cross validation on OuluVS database because there are not abundant samples for each phrase of each speaker. In total, ten speakers with at least three training samples for each phrase are selected for this experiment, because too few training samples, for instance, one or two, could bias the recognition rate. In our experiments, every utterance is left out, and the remaining utterances are trained for every speaker. Fig. 11 presents a detailed comparison of the results for every subject. Table IV shows the overall recognition results. The block parameters used here are also $(1\times5\times3+1\times5\times2)$. We can see there is no significant difference in performance between automatic eye detection and manual eye positioning.

On the basis of AVLetters database, Matthews *et al.* [24] presented two top-down approaches that fits a model of the inner and outer lip contours and derive lipreading features from a PCA of shape—Active Shape Model (ASM)—or shape and appearance—Active Appearance Model (AAM)—respectively, and as well a bottom-up method which uses a nonlinear scale-space analysis—multiscale spatial analysis (MSA)—to form features directly from the pixel intensity.

TABLE IV
RESULTS OF SPEAKER-DEPENDENT EXPERIMENTS ON OULUVS DATABASE

Features	Eye detection	Results
$LBP-TOP_{8,3}$	Manual	70.2%
	Automatic	64.2%

In their experiments, their training set was the first two utterances of each of the letters from all speakers (520 utterances) and the test set was the third utterance from all speakers (260 utterances). In this way, the training set includes the utterances from all speakers, so it is not speaker-independent. But it is not trained and tested for individual speakers, so it is also not completely speaker-dependent. We call this evaluation setup "semi-speaker-dependent".

We did the same evaluation using the same training set and test set, i.e., using the first two utterances of each of the letters from all speakers (520 utterances) as training set and the third utterance from all speakers (260 utterances) as test set. The results are listed in the second column (third-test) in Table V. As well, the three-fold-cross-validation is also made by using every one from three repetitions as test set and the other two repetitions as training set. In this way, the overall performance could be evaluated, seeing the third column (three-fold) in Table V. Comparing to the best results from ASM, AAM, and MSA proposed in [24], our accuracy (fifth row) from same classifier HMM, but with our own proposed LBP-TOP features, is 12.7% higher than MSA, 30.4% higher than ASM. Table V also gives the results from different parameters of LBP-TOP features with SVM classifiers. The best result is 58.85% for the third-fold test and 62.82% for the three-fold test. The performance of commonly used features: temporal derivatives, optical flow, and DCT with same SVM classifiers are also provided. Even though they work better than ASM, and the accuracy from DCT is even better than MSA, our method obtained the best recognition results on the same evaluation setup.

Moreover, we also carry out experiments on continuous speech segmentation. The letters are combined into longer sequences to be used for segmentation and classification. Every one from three repetitions for each letter is put into one sequence in random order for all subjects, so we have 30 long sequences each containing 26 spoken letters in random order.

The groundtruth for each sequence is provided by the labeling of the letters in the AVLetters database.

For training, these 30 sequences are divided into three groups, and every time, one group is used as test set and the other two as training set. That is to say, the ith ($i=1, 2, 3$) test sequence includes ten speakers with their ith utterances for 26 letters. This is repeated three times. In this way, the training set includes the utterances from all speakers, not for every speaker, so it is not speaker-dependent. This evaluation setup is semi-speaker-independent. HMM has been used successfully in many different sequence recognition applications. In speech recognition, HMM is

TABLE V
RESULTS OF SEMI-SPEAKER-DEPENDENT EXPERIMENTS ON AVLETTERS DATABASE

Features	Classifier	Third-test	Three-fold
ASM [24]	HMM	26.9	———
AAM [24]	HMM	41.9	———
MSA [24]	HMM	44.6	———
$LBP-TOP^{u2}_{8,3}:2\times5$:	HMM	**57.3**	**59.6**
Temporal Derivatives: $D=5$	SVM	30.00	31.54
Optical Flow: $D=5$	SVM	32.31	32.18
DCT: $S5, D=5$	SVM	53.46	53.33
DCT: $T10, D=5$	SVM	48.08	52.56
DCT: $C7, D=5$	SVM	48.85	52.31
DCT: $H7, D=5$	SVM	51.15	53.59
$LBP-TOP_{8,3}:1\times5\times3$	SVM	54.62	58.85
$LBP-TOP^{u2}_{8,3}:1\times5\times3$	SVM	55.00	59.23
$LBP-TOP^{u2}_{8,3}:2\times5\times3$	SVM	**58.85**	**62.82**

the most common method of modeling. Here, based on $LBP-TOP^{u2}_{8,3}$ features with 2×5 blocks, an HMM is utilized to recognize and segment these longer sequences. This is done by first training an HMM for each letter in the AVLetters database; these HMMs form the states of a larger HMM used to model the transitions between the letters and decode the long sequence of letters.

Frame recognition rate (FRR) is used as measure. FRR is defined as

$$FRR=\frac{N_c}{N} \tag{6}$$

where N_c is the number of frames classified correctly and N is the total number of frames in the sequence. This is a measure of segmentation accuracy as well as classification accuracy.

With this approach, an accuracy of 56.09% is obtained by averaging the results from three rounds of evaluation, which shows promising performance for continuous speech segmentation.

3) Experiments With Feature Selection: For OuluVS database, we use $LBP-TOP_{8,3}$ with $1\times5\times3$ block volumes in two-fold cross-validation for the following unbiased comparison, from which the baseline result is 54.22%. To learn more effective multiresolution features, the proposed feature selection method is utilized and a comparison is made. To get the multiresolution features, eight groups of $LBP-TOP$ features from different neighboring points, radii, and block volume sizes with 339 slices in total, as shown in Table VI, were extracted and exploited for selection. In the experimental results from separate resolution features, the best accuracy is from $LBP-TOP_{8,3}$ with $2\times5\times3$ block volumes. The highest number of features is also selected from $LBP-TOP_{8,3}$ with $2\times5\times3$ block volumes, which proves the consistency of the selected effective features. To give a concise presentation, in the following parts, Figs. 13 and 14 just show the selected features in $LBP-TOP_{8,3}$ with $2\times5\times3$ block volumes while the results in comparisons shown in Fig. 12 are from the multiresolution features.

Fig. 12 shows that the slice-based feature selection algorithm works much better than the block-based one. It also demonstrates that when the number of selected slices is quite high, e.g., 60 slices, the All-All strategy provided better results than the one-one approach. This is perhaps because the use of too many slices will weaken the discrimination among the pairs of classes. More importantly, when a smaller number of selected slices is used, the One-One strategy will learn more discriminative features for the pairs of spoken phrases achieving better results than the All-All approach, for example 56.18% versus 55.57% from just 15 slices, as well as around 2% better than obtained from separate $LBP-TOP_{8,3}$ with 45 slices, 54.22%.

TABLE VI
MULTIRESOLUTION FEATURES

Features	Slices	Accuracy(%)
$LBP-TOP_{8,3}$:$1\times5\times3$ blocks	45	54.22
$LBP-TOP_{8,3}$:$2\times5\times3$ blocks	90	55.57
$LBP-TOP_{8,3}$:$1\times5\times2$ blocks	30	49.33
$LBP-TOP_{8,1}$:$1\times5\times3$ blocks	45	48.59
$LBP-TOP_{4,1}$:$1\times5\times3$ blocks	45	44.43
$LBP-TOP_{4,3}$:$1\times5\times3$ blocks	45	44.31
$LBP-TOP_{8,3}$:$2\times4\times1$ blocks	24	33.05
$LBP-TOP_{8,3}$:$1\times5\times1$ blocks	15	31.09

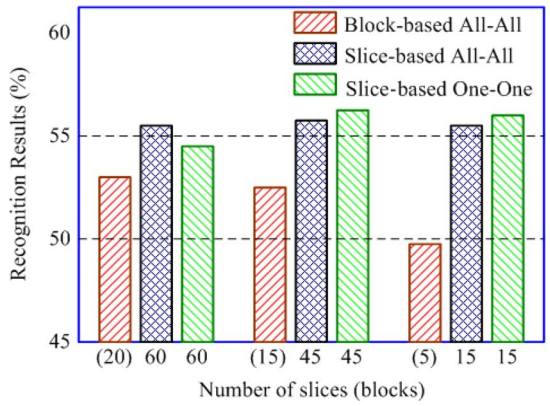


Fig. 12. Comparative results for slice-based and block-based methods on OuluVS database.

Fig. 13 shows the selected slices for similar phrases "see you" and "thank you". These phrases were the most difficult to recognize because they are quite similar in the latter part containing the same word "you". The selected slices are mainly in the first and second part of the phrase; just one vertical slice is from the last part. The selected features are consistent with the human intuition. The phrases "excuse me" and "I am sorry" shown in Fig. 14 are different throughout the whole utterance, and the selected features also come from the whole pronunciation. With the proposed feature selection strategy, more specific and adaptive features are selected for different pairs of phrase classes, as shown in Figs. 13 and 14, providing more discriminative features.

4) One-One versus One-Rest Recognition: We use the SVMs as the classifiers. Since SVMs are only used for separating two classes, when we have multiple classes, there could be different strategies. In the previous experiments on our own dataset, the ten-phrase classification problem is decomposed into 45 two-class problems ("Hello"- "Excuse me", "I am sorry"- "Thank you", "You are welcome"- "Have a good time", etc.). But using this multiple two-class strategy, the number of classifiers grows quadratically with the number of classes to be recognized like in AVLetters database. When the class number is N, the number of the SVM classifiers would be $N(N-1)/2$. The other option

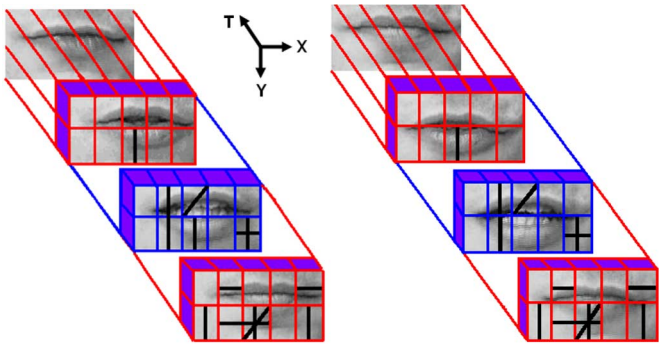


Fig. 13. Selected 15 slices for phrases "See you" and "Thank you". "|" in the blocks means the YT slice (vertical motion) is selected, and "−" the XT slice (horizontal motion), "/" means the appearance XY slice.

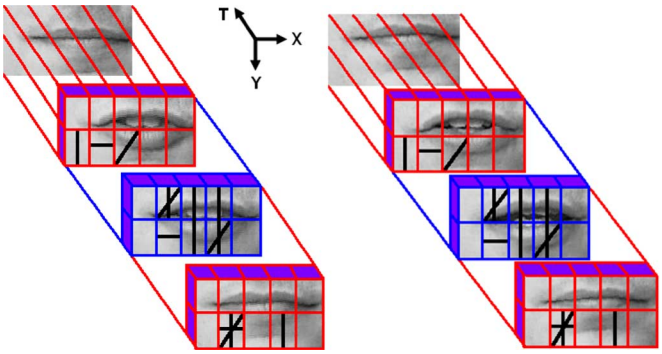


Fig. 14. Selected 15 slices for phrases "Excuse me" and "I am sorry". "|" in the blocks means the YT slice is selected, and "−" the XT slice, "/" means the appearance XY slice.

TABLE VII
RESULTS FROM ONE-TO-ONE AND ONE-TO-REST CLASSIFIERS ON SEMI-SPEAKER-DEPENDENT EXPERIMENTS (RESULTS IN THE PARENTHESES ARE FROM ONE-TO-REST STRATEGY)

Features	Blocks	Third-test (O-R)	Three-fold (O-R)
$LBP-TOP_{8,3}$	$1\times5\times3$	**54.62**(50.77)	**58.85**(54.36)
$LBP-TOP^{u2}_{8,3}$	$1\times5\times3$	**55.00**(52.31)	**59.23**(55.77)
$LBP-TOP^{u2}_{8,3}$	$2\times5\times3$	**58.85**(54.23)	**62.82**(57.18)

is one-to-rest strategy, to decompose the N-class problem into N one-to-rest problems. Here we give the results from one-to-one and one-to-rest strategies for semi-speaker-dependent evaluation on AVLetters database.

It can be seen from Tables V and VII that the results from one-to-rest using the proposed features are better than those from the ASM, AAM, MSA, and other commonly-used features. However, compared with one-to-one, the results from one-to-rest are much lower. So the decision of which strategy to use depends on the application. If the number of classes is not too high, and the recognition accuracy is much more important than the time consumed, the one-to-one strategy could be utilized. Otherwise, one-to-rest can be a good option.

5) Confusion Matrix: In visual speech recognition, a viseme is defined as the smallest visibly distinguishable unit of speech [8]. The viseme is analogous to the phoneme in audio speech, as words are composed of phonemes, so the visual sequences used here are composed of visemes. There is currently no agreement on the mapping of phonemes to visemes, for example, [8] groups the audio consonants into nine viseme groups, whereas [20] and [19] group audio phonemes into five consonant visemes and six vowel visemes, as shown in Table VIII.

TABLE VIII
PHONEMES USED IN THE AVLetters DATABASE WITH THE CORRESPONDING VISEMES. IN THIS PAPER, WE USE THE ARPABET PHONETIC ALPHABET NOTATION, COMMONLY USED IN THE SPEECH RECOGNITION COMMUNITY, TO REPRESENT PHONEMES. A MAPPING OF ARPABET NOTATION TO IPA PHONEME SYMBOLS CAN BE FOUND AT www.cs.cmu.edu/laura/pages/arpabet.ps

Visemes	Phonemes	Visemes	Phonemes
/p/	P, B, M	/iy/	IY
/f/	F	/aa/	AA
/t/	T, D, S, Z	/ah/	AY
/ch/	CH, JH, ZH	/ow/	OW
/w/	W, R	/uw/	UW
/k/	K, N, L	/ey/	EH, EY

It is interesting to note that the distribution of errors in our experiments on the AVLetters database is not random. In Table IX, showing the confusion matrices for subjects pronouncing the letters of the alphabet, we can see that the majority of confusion is between sequences consisting of the same visemes, for example the words B and P, composed of phonemes [B + IY] and [P + IY], respectively. If we take the mapping of phonemes to visemes from Table VIII, we can see that these words are visually the same and composed of the visemes /p + iy/. Similarly the words C, D and T, [S + IY], [D + IY] and [T + IY], are composed of the same sequence of visemes, /t + iy/.

While most confusions in visual speech recognition are caused by the phonemes of two words being mapped to the same viseme, it is possible for different visemes to appear the same due to their context. These confusions are caused by the phenomena of co-articulation [10], where the mouth shape of a particular phoneme can cause nearby phonemes to have a similar mouth shape. This is particularly true in cases where the phonemes have little visible effect on the shape of the lips. In the SVM confusion matrix, Table IX, we can see that the words Q and U, [K + W + UW] and [J + UW], are confused due to the rounded vowel [U] causing the [K] in Q to also be rounded. Similarly in the words X and S, [EH + K + S] and [EH + S] the initial vowel [EH] governs the lip shape of the whole word. Consonants can also cause co-articulation effects. In the case of G and J, although the final vowel is mapped to a different viseme, [JH + IY] and [JH + EY], the sequence is dominated by the rounded lip shape of the consonant [JH] causing the confusion between the two sequences.

If we put (B,P), (C,D,T), (Q,U), (S,X), (G,J) into viseme groups, the recognition accuracy just from the visual features is up to 75.77%.

VI. CONCLUSIONS

A novel local spatiotemporal descriptor for visual speech recognition was proposed, considering the spatial region and pronunciation order in the utterance. The movements of mouth regions are described using local binary patterns from XY, XT, and YT planes, combining local features from pixel, block, and volume levels. Reliable lip segmentation and tracking is a major problem in automatic visual speech recognition, especially in poor imaging conditions. Our approach avoids this

TABLE IX
CONFUSION MATRIX FROM SVMS ($LBP-TOP^{u2}_{8,3}$ FEATURES WITH $2 \times 5 \times 3$ BLOCKS)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	**17**	1			1				1		1	3		5					1							
B	1	**18**														10						1				
C			**11**	8	1		1				1					1				5		1				1
D			6	**12**			1				1									9						1
E	3			1	**19**						2								2	3						
F						**25**					1	1	1						1					1		
G			1				**16**			6										3		2			1	1
H						1		**20**	1		1	2		1					2		1			1		
I	3					1			**18**		1	1		2	1			3								
J			1				3			**22**	1									1		2				
K	2		1	2	1					1	**17**	1		2						3						
L	4				1			1			1	**13**	1	5				1		1					2		
M	1					1					1	1	**23**	1						1						1
N	5			1	2				2			3	1	**11**		1		1	2					1		
O		1													**24**		2	1			2					
P	2	10				1								1		**15**							1			
Q								1							1		**17**	1			10					
R	1						1		1		1				2			**23**							1	
S								1						2					**19**					8		
T			4	4	2						1									**19**						
U				1					1						3		12				**13**					
V			3							1	1											**23**				2
W																					1		**28**		1	
X						1		1			2			2					6	1				**17**		
Y		1															1								**28**	
Z			3	1						2									1	1						**22**

using local spatiotemporal descriptors computed from mouth regions which are much easier to extract than lips. Automatic face and eye detection are exploited to extract mouth regions. With our approach, no error prone segmentation of moving lips is needed.

Experiments on a dataset collected from 20 persons show very promising results. For ten spoken phrases, the obtained speaker-independent recognition rate is around 62% and speaker-dependent result around 70%. Moreover, 62.8% accuracy is obtained for AVLetters database, which is much better than the other methods. Especially, when using the same classifier, our accuracy is 12.7% higher than [24] under the same test setup, which obviously shows the effectiveness of our proposed features. Multiresolution features and feature selection approach are presented and the preliminary experiments are carried out on OuluVS database. Results show the effectiveness of selecting principal appearance and motion for specific class pairs. OuluVS database includes ten phrases from 20 people, while AVLetters database has 26 letters from ten people. So with these two databases, we evaluate and report the performance for data with phrase variations and as well the diversities from different speakers. We also carried out continuous speech segmentation experiments on AVLetters database. The obtained accuracy 56.09% is promising for this challenging task using solely visual information.

From the analysis of confusion matrix with 26 English letters, we can see that the clustering of errors in the word recognition actually shows that this method is accurately recognizing visemes by capturing the shape of the mouth.

Compared with the state-of-the-art, our method does not need to 1) segment lip contours [2], [26]; 2) track lips in the subsequent frames; 3) select constant illumination or perform illumination correction [34]; and 4) align lip features with respect to the canonical template [2], [3] or normalize the mouth images to a fixed size as done by most of the papers [5], [26], [34]. Furthermore, our method shows stability for low-resolution sequences. In this way, our experimental setup is more realistic.

Our future plan is to research not only isolated phrases but also the continuous speech, e.g., using viseme models for recognition, to improve the quality of lipreading. Moreover, it is of interest to combine visual and audio information to promote speech recognition, and to apply our methodology to human-robot interaction in a smart environment.

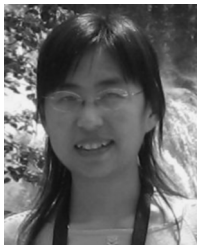
ACKNOWLEDGMENT

The authors would like to thank Mr. J. Kontinen for helping collect the OuluVS database used in the paper and Dr. R. Harvey for providing the AVLetters database used in experiments.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, 2006.
[2] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audio-visual speech recognition using MPEG-4 compliant visual features," *EURASIP J. Appl. Signal Process.*, vol. 11, pp. 1213–1227, 2002.
[3] P. S. Aleksic and A. K. Katsaggelos, "Product HMMs for audio-visual continuous speech recognition using facial animation parameters," in *Proc. Int. Conf. Multimedia and Expo (ICME)*, 2003, pp. 481–484.
[4] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *Proc. IEEE*, vol. 94, no. 11, pp. 2025–2044, Nov. 2006.
[5] I. Arsic and J. P. Thiran, "Mutual information engenlips for audio-visual speech," in *Proc. 14th Eur. Signal Processing Conf.*, Italy, 2006.
[6] S. Basu, C. Neti, N. Rajput, A. Senior, L. Subramaniam, and A. Verma, "Audio-visual large vocabulary continuous speech recognition in the broadcast domain," in *Proc. IEEE 3rd Workshop Multimedia Signal Processing*, 1999, pp. 475–481.
[7] N. M. Brooke, "Using the visual component in automatic speech recognition," in *Proc. Int. Conf. Spoken Language*, 1996, pp. 1656–1659.
[8] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proc. IEEE*, vol. 86, no. 5, pp. 837–852, May 1998.

[9] G. I. Chiou and J. N. Hwang, "Lipreading from color video," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1192–1195, Aug. 1997.
[10] M. Cohen and D. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, D. Thalmann and N. Magnenat-Thalmann, Eds. New York: Springer-Verlag, 1993.
[11] P. Duchnowski, M. Hunke, D. Busching, U. Meier, and A. Waibel, "Toward movement-invariant automatic lipreading and speech recognition," in *Proc. Int. Conf. Spoken Language*, 1995, pp. 109–112.
[12] N. Fox, R. Gross, and P. Chazal, "Person identification using automatic integration of speech, lip and face experts," in *Proc. ACM SIGMM Workshop Biometrics Methods and Applications*, 2003, pp. 25–32.
[13] R. W. Frischholz and U. Dieckmann, "BioID: A multimodal biometric identification system," *Computer*, vol. 33, no. 2, pp. 64–68, 2000.
[14] J. N. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, "DBN based multi-stream models for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. 993–996.
[15] M. Gurban and J. P. Thiran, "Audio-visual speech recognition with a hybrid SVM-HMM system," in *Proc. 13th Eur. Signal Processing Conf.*, 2005.
[16] A. Hadid, M. Pietikäinen, and S. Li, "Learning personal specific facial dynamics for face recognition from videos," in *Proc. Workshop Analysis and Modeling of Faces and Gestures*, 2007, pp. 1–15.
[17] T. Hazen, K. Saenko, C. H. La, and J. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proc. ICMI*, 2005.
[18] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *Proc. Int. Conf. Spoken Language*, 2004, pp. 2489–2492.
[19] S. Lee and D. Yook, "Audio-to-visual conversion using Hidden Markov Models," in *Proc. 7th Pacific Rim Int. Conf. Artificial Intelligence*, 2002, pp. 563–570.
[20] P. Lucey, T. Martin, and S. Sridharan, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in *Proc. 10th Australian Int. Conf. Speech Science and Technology*, 2004.
[21] J. Luettin, N. A. Thacher, and S. W. Beet, "Speaker identification by lipreading," in *Proc. Int. Conf. Spoken Language (ICSLP)*, 1996, pp. 62–64.
[22] K. Mase and A. Pentland, "Automatic lipreading by optical flow analysis," *Syst. Comput. Jpn.*, vol. 22, no. 6, pp. 67–75, 1991.
[23] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
[24] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, 2002.
[25] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "Xm2vtsdb: The extended m2vts database," in *Proc. 2nd Int. Conf. Audio and Video-Based Biometric Person Authentication*, Washington, DC, 1999.
[26] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, vol. 11, pp. 1274–1288, 2002.
[27] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," in *Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, Final Workshop 2000 Report*, Oct. 2000.
[28] Z. Niu, S. Shan, S. Yan, X. Chen, and W. Gao, "2d cascaded adaboost for eye localization," in *Proc. Int. Conf. Pattern Recognition*, 2006, pp. 1216–1219.
[29] P. Niyogi, E. Petajan, and J. Zhong, "Feature based representation for audio-visual speech recognition," in *Proc. Audio Visual Speech Conf.*, 1999.
[30] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray scale and rotation invariant texture analysis with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.
[31] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. ICIP*, Chicago, IL, 1998, pp. 173–177.
[32] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, vol. 91, pp. 1306–1326, 2003.
[33] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-Visual Speech Processing*. Cambridge, MA: MIT Press, 2004.
[34] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, "Visual speech recognition with loosely synchronized feature streams," in *Proc. ICCV*, 2005, pp. 1424–1431.
[35] K. Saenko, K. Livescu, J. Glass, and T. Darrell, "Production domain modeling of pronunciation for visual speech recognition," in *Proc. ICASSP*, 2005, pp. 473–476.
[36] C. Sanderson, "The VidTIMIT database," in *IDIAP Communication 02-06*, Martigny, Switzerland, 2002.
[37] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, 2001, pp. 511–518.
[38] G. Zhang, X. Huang, S. Z. Li, Y. Wang, and X. Wu, "Boosting local binary pattern (LBP)-based face recognition," in *Sinobiometrics*, 2004, vol. 3338, LNCS, pp. 179–186.
[39] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, 2007.
[40] G. Zhao, M. Pietikäinen, and A. Hadid, "Local spatiotemporal descriptors for visual recognition of spoken phrases," in *Proc. 2nd Int. Workshop Human-Centered Multimedia (HCM2007)*, 2007, pp. 57–65.
[41] G. Zhao and M. Pietikäinen, "Boosted multi-resolution spatiotemporal descriptors for facial expression recognition," *Pattern Recognit. Lett.,* Special Issue on Image/Video-Based Pattern Analysis and HCI, 2009, accepted for publication.



Guoying Zhao received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005.

Since July 2005, she has been a Senior Researcher in the Machine Vision Group at the University of Oulu, Oulu, Finland. Her research interests include gait analysis, dynamic texture recognition, facial expression recognition, human motion analysis, and person identification. She has authored over 50 papers in journals and conferences and has served as a reviewer for many journals and conferences. She gave an invited talk "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions" in the Institute of Computing Technology, Chinese Academy of Sciences, July 2007. With Prof. Pietikäinen, she gave a tutorial: "Local Binary Pattern Approach to Computer Vision" in the 18th ICPR, August 2006, Hong Kong. With Prof. Matti Pietikäinen again, she will give a tutorial: Local Texture Descriptors in Computer Vision, in 12th ICCV, September 2009, Kyoto, Japan. She is authoring/editing three books: Springer book *Computer Vision Using Local Binary Patterns*; Springer book *Machine Learning for Vision-based Motion Analysis*; and IGI global book *Machine Learning for Human Motion Analysis: Theory and Practice*.

Dr. Zhao is a guest editor for the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS Special Issue on New Advances in Video-Based Gait Analysis and Applications: Challenges and Solutions. She was a co-chair of the ECCV 2008 Workshop on Machine Learning for Vision-Based Motion Analysis (MLVMA) and is a co-chair of the *MLVMA* workshop at ICCV 2009.

Mark Barnard received the Ph.D. (Docteur es Science) from Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland, in November 2005.

While completing his Ph.D., he worked at the IDIAP Research Institute as a Research Assistant. His thesis was entitled "Multimedia event modeling and recognition". In 2006, he joined the Machine Vision Group at the University of Oulu, Oulu, Finland, where he completed three years as a Postdoctoral Researcher. He is currently a research officer at the Centre for Vision, Speech and Signal processing at the University of Surrey, Guildford, U.K. His current research interests include feature tracking, recognition of human activity in video material, using motion recognition to control mobile devices, and also general sequence processing applications. He has published numerous papers in international journals and conferences and also served as a reviewer for many top level journals and conferences.

Dr. Barnard is a member of the program committee of the Machine Learning for Vision-Based Motion Analysis (MLVMA) at ICCV in 2009.

Matti Pietikäinen (SM'96) received the Doctor of Science in Technology degree from the University of Oulu, Oulu, Finland, in 1982.

In 1981, he established the Machine Vision Group at the University of Oulu. This group has achieved a highly respected position in its field, and its research results have been widely exploited in industry. Currently, he is a Professor of information engineering, Scientific Director of Infotech Oulu Research Center, and Leader of the Machine Vision Group at the University of Oulu. From 1980 to 1981 and from 1984 to 1985, he visited the Computer Vision Laboratory at the University of Maryland. His research interests include texture-based computer vision, face analysis, activity analysis, and their applications in human-computer/robot interaction, person identification, visual surveillance, and image/video retrieval. He has authored over 200 refereed papers in international journals, books, and conference proceedings and about 100 other publications or reports. His research on texture-based computer vision, local binary pattern (LBP) methodology, and facial image analysis, for example, is frequently cited, and its results are used in various applications around the world.

Dr. Pietikäinen was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and *Pattern Recognition* journals. He was guest editor (with L.F. Pau) of a two-part special issue on "Machine Vision for Advanced Production" for the *International Journal of Pattern Recognition and Artificial Intelligence* (also reprinted as a book by World Scientific in 1996). He was also the editor of the book *Texture Analysis in Machine Vision* (Singapore: World Scientific, 2000) and has served as a reviewer for numerous journals and conferences. He was the president of the Pattern Recognition Society of Finland from 1989 to 1992. From 1989 to 2007, he served as a member of the Governing Board of the International Association for Pattern Recognition (IAPR) and became one of the founding fellows of the IAPR in 1994. He regularly serves on program committees of the top conferences and workshops of his field. Recently, he was an area chair of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07), a co-chair of Workshops of International Conference on Pattern Recognition (ICPR '08), a co-chair of ECCV 2008 Workshop on Machine Learning for Vision-Based Motion Analysis (MLVMA), and is a co-chair of MLVMA workshop at ICCV 2009. He was the vice-chair of the IEEE Finland Section.