

Local Spatiotemporal Descriptors for Visual Recognition of Spoken Phrases

Guoying Zhao, Matti Pietikäinen, Abdenour Hadid

Machine Vision Group, Infotech Oulu and Department of Electrical and Information Engineering,
P. O. Box 4500 FI-90014 University of Oulu, Finland

E-mail:{gyzhao, mkp, hadid}@ee.oulu.fi

ABSTRACT

Visual speech information plays an important role in speech recognition under noisy conditions or for listeners with hearing impairment. In this paper, we propose local spatiotemporal descriptors to represent and recognize spoken isolated phrases based solely on visual input. Positions of the eyes determined by a robust face and eye detector are used for localizing the mouth regions in face images. Spatiotemporal local binary patterns extracted from these regions are used for describing 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 the Tulips1 audio-visual database, the accuracy 92.7% of our method clearly outperforms the others. Advantages of our approach include local processing and robustness to monotonic gray-scale changes. Moreover, no error prone segmentation of moving lips is needed.

Categories and Subject Descriptors

H.5.2 [Information Interfaces and Presentation]: User Interfaces - *Evaluation/methodology, Theory and methods.*

General Terms

Algorithms

Keywords

Visual speech recognition, local spatiotemporal descriptors, mouth region localization, face and eye detection.

1. INTRODUCTION

Human-centered computing (HCC) is an interdisciplinary field which is about the computing technology research related to humans. Human condition can be easily recorded in the form of multimedia, including audio and video by cameras, which as the main equipment to acquire data, are cheap and of high quality. They can be used everywhere which increases their applicability.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
HCM'07, September 28, 2007, Augsburg, Bavaria, Germany.
Copyright 2007 ACM 978-1-59593-781-0/07/0009...\$5.00.

Machine vision contributes to the HCC mainly by enabling machines to recognize people and understand their actions, and thus provide smarter, better and more adaptive service, response, help and interaction with the computers. In this paper, we propose an approach for visual speech recognition, which could improve the human-computer interaction especially in noisy environments or for listeners with hearing impairment.

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. In some researches, lipreading combined with face and voice is studied to help biometric identification [1-3]. There is also much work focusing on audio-visual speech recognition (AVSR) [4-16], trying to find effective ways of combining visual information with existing audio-only speech recognition systems (ASR). McGurk effect [17] 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. 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.

Comprehensive reviews of automatic audio-visual speech recognition can be found in [4,18]. 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 [9], contours of lips [10,11,14], shape of jaw and cheek [10,11], and mouth width, mouth opening, oral cavity area and oral cavity perimeter [15]. 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. The appearance features are 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 Principal Component Analysis (PCA) [5,13], the discrete cosine transform (DCT) [19], or a combination of these transforms [8,20,21].

In addition, features from both categories can be combined for lip localization and visual feature extraction [10,12,22].

Most of the researches focus on using visual information to improve speech recognition. 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. 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 TV broadcast when the audio link is weak or corrupted, are some examples. Furthermore, for the persons with hearing impairment, visual information is the only source of information from TV broadcast 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 [20-23]. So addressing this problem could improve the quality of human-computer interaction (HCI). Saenko et al. [20,21] use articulatory features and dynamic Bayesian network for recognizing spoken phrases with multiple loosely synchronized streams. Chiou and Hwang [22] utilize snakes to extract visual features of geometric space, Karhunen-Loeve transform to extract principal components in the color eigenspace and HMMs to recognize the isolated words. Matthews et al. [23] presented 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 non-linear scale-space analysis to form features directly from the pixel intensity.

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 omitting the local features. Local features can describe the local changes of images in space and time. In this paper, 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 robust face and eye detector is used to get the mouth region automatically. A Support Vector Machine (SVM) classifier is utilized for recognition.

2. SYSTEM OVERVIEW

Our system consists of three stages, as shown in Fig. 1. 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 last stage is to recognize the input utterance using SVM classifier.

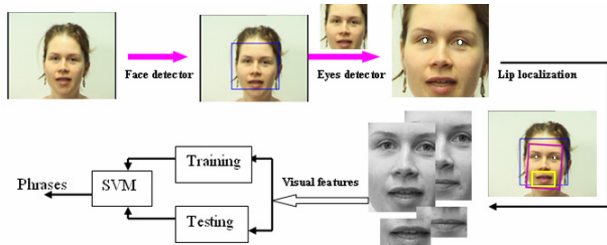


Figure 1. System diagram.

3. LOCAL SPATIOTEMPORAL DESCRIPTORES 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 [24]. For each pixel in an image, a binary code is produced by thresholding its neighborhood with the value of the center pixel (Fig. 2 (a) and Eq. (1)).

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

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.2 (b). In this way, one can collect larger-scale texture primitives.

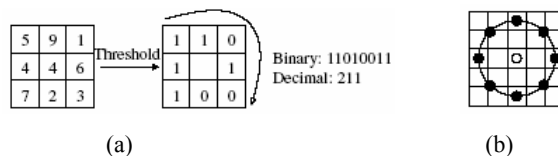


Figure 2. (a) Basic LBP operator. (b) The circular (8,2) neighborhood.

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 [25].

Recently, a method for temporal texture recognition using spatiotemporal local binary patterns extracted from three orthogonal planes (LBP-TOP) was proposed [26]. With this approach the ordinary LBP for static images was extended to 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 $LBP-TOP_{P_{XY}, P_{XT}, P_{YT}, R_X, R_Y, R_T}$. 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.

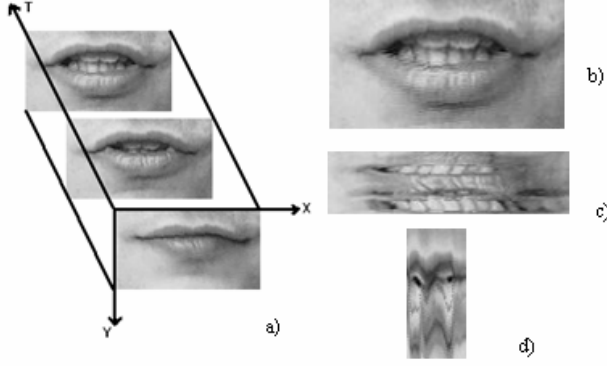


Figure 3. (a) Volume of utterance sequence (b) Image in XY plane (147x81) (c) Image in XT plane (147x38) in $y=40$ (last row is pixels of $y=40$ in first image) (d) Image in TY plane (38x81) in $x=70$ (first column is the pixels of $x=70$ in first frame).

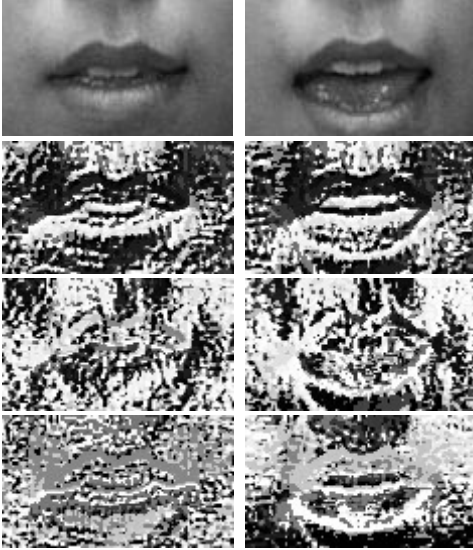


Figure 4. Mouth region images (first row), LBP-XY images (second row), LBP-XT images (third row) and LBP-YT images (last row) from one utterance.

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. 3 (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. 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. 4 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 get almost the same features. To overcome this effect, the whole sequence is not only divided into block volumes according to spatial regions but also in time order, as Fig. 5 (a) shows.

The LBP-TOP histograms in each block volume are computed and concatenated into a single histogram, as Fig. 5 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. 6.

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 segment, and all information from the regional level is concatenated to build a global description of the mouth region motion.

A histogram of the mouth movements can be defined as

$$H_{r,c,d,j,i} = \sum_{x,y,t} I\{f_j(x,y,t) = i\}, \quad (2)$$

$$i = 0, \dots, n_j - 1; j = 0, 1, 2.$$

in which n_j is the number of different labels produced by the LBP operator in the j th 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 j th plane, r is the index of rows, c is of columns and d is of time of block volume.

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

The histograms must be normalized to get a coherent description:

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

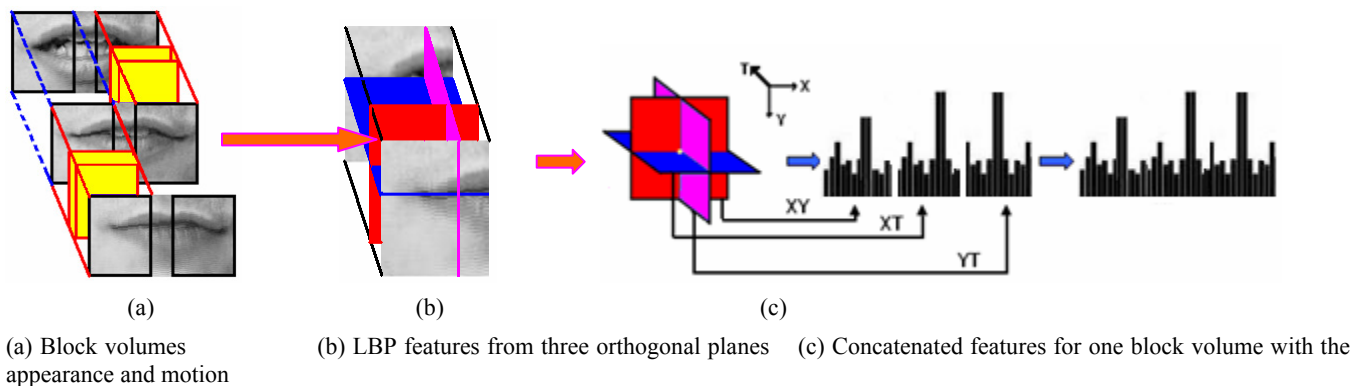


Figure 5. Features in each block volume.

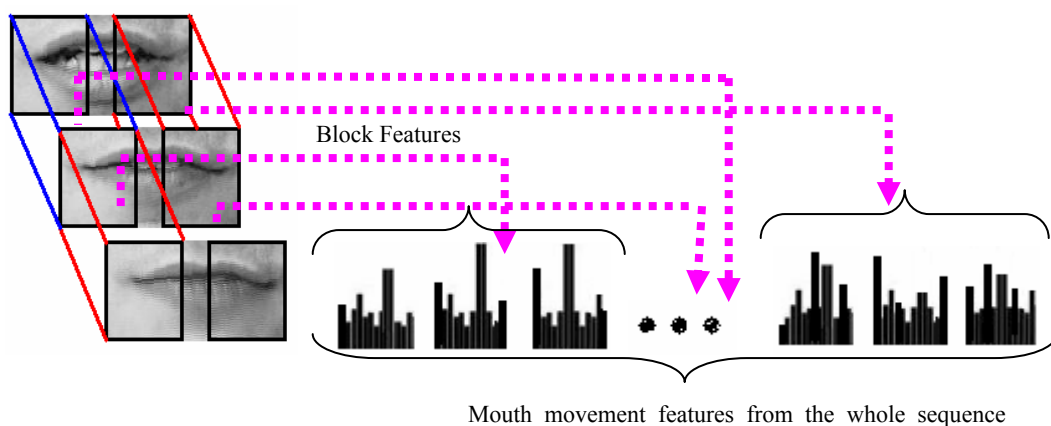


Figure 6. Mouth movement representation.

4. MOUTH REGION LOCALIZATION

An automatic approach was adopted for mouth region localization. First, the face region is localized based on a robust method using LBP features in a coarse-to-fine detection strategy (pyramid architecture) embedded in a fast classification scheme based on AdaBoost learning. The obtained results are comparable to the state-of-the-art.

Once the face is localized, the eyes are then searched for in the upper part of the face region. Our eye detector is inspired by the works of Viola and Jones on the use of Haar-like features with integral images [27] and that of Heusch et al. on the use of LBP as a preprocessing step for handling illumination changes [28]. It uses Haar-like features extracted from LBP images. Thus, the images are first filtered by LBP operator ($LBP_{8,1}$) and then Haar-like features are extracted and used with AdaBoost for building a cascade of classifiers. Once the eyes are localized, we then estimate the location and extract the mouth area as shown in Figure 7. In this figure d represents the distance between eyes, and $d1$ and $d2$ are set to $0.4d$ and $0.7d$, respectively. In this way accurate location of the face can be determined, as the inside square shows. The width F_w is $1.8d$ and height F_h $2.4d$. This approach can handle the in-plane rotation. Next, the mouth center (M_x, M_y) is determined using

$M_x = F_w/2$ (midpoint of the face in width) and $M_y = 0.8F_h$. Then the width and height of the mouth are set as $0.3F_w$ and $2F_h/3$, respectively. In this way, the mouth region is obtained.

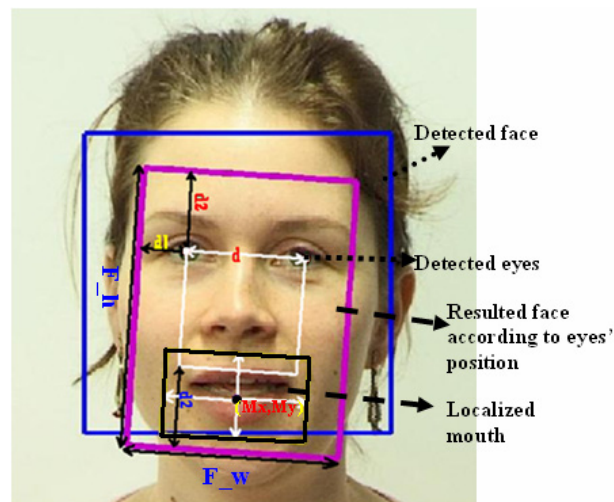


Figure 7. Example of face, eye and mouth localization.

5. EXPERIMENTS

5.1 Dataset description

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 [4,7,12,29-31].

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. [9,20]. Even those with multiple subjects are usually limited to small tasks such as isolated letters [23], digits [5], or a short list of fixed phrases [29]. The M2VTS database and the expanded XM2VTSDB [29] 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 hours in duration [12]. The other one is the VidTIMIT database [30], which consists of 43 subjects reciting different 10 TIMIT sentences each. It has been used in multi-modal person verification research.

There are just few datasets providing phrase data [20,29-31], and in those the number of speakers is pretty small [20]. Though AVTIMIT [31], XM2VTSDB [29] and VidTIMIT [30] include quite many speakers, but the speakers utter different sentences or phrases [30,31] or small number of sentences [29]. Due to the lack of a publicly available database suitable for our needs, we collected our own visual speech dataset for performance evaluation.

A SONY DSR-200AP 3CCD-camera with frame rate 25 fps was used to collect the data. The image resolution was 720 by 576 pixels. Our dataset includes twenty persons, each uttering ten everyday's greetings one to five times. These short phrases are listed in Table 1.

Table 1. Phrases included in the dataset.

C1	"Excuse me"	C6	"See you"
C2	"Good bye"	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"

The subjects were asked to sit on a chair. The distance between the speaker and the camera was 160cm. 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: at 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 speeds.

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

5.2 Detection of face and eyes

Extraction of the visual features starts with the detection of the face and eyes of the speaker. The face and eye detection methods were briefly described in Section 4. The training of the face and eye detectors was done in a similar way, using a bootstrap strategy to collect negative examples. In case of eye detection, we randomly extracted non-eye samples from a set of natural images which do not contain eyes. Then, we trained the system, run the eye detector, and collected all those non-eye patterns that were wrongly classified as eyes and used them for training. Additionally, we considered negative training samples extracted also from the facial regions because it has been shown that this can enhance the performance of the system. In total, we trained the system using 3,116 eye patterns (positive samples) and 2,461 non-eye patterns (negative samples). Then, we tested our system on a database containing over 30,000 frontal face images and compared the results to those obtained by using Haar-like features and LBP features separately. Detection rates of 86.7%, 81.3% and 80.8% were obtained when considering LBP/Haar-like features, LBP only and Haar-like features only, respectively. From the speed point of view, LBP-Haar is four times faster than LBP only because no histogram computation is needed. Besides, it is worth noting that, unlike most of other eye detectors, the presence of glasses did not affect the performance of our detector. Fig. 8 shows some examples of eye detections performed by the system.

Once the positions of the speaker's eyes are obtained, we make use of the distance between the speaker's eyes to estimate the position of the speaker's mouth region (as shown in Figure 7). The face and eye detectors were applied on the first frame of each sequence to locate the mouth region. In the subsequent frames, the same region position was used to get the mouth regions since the position of the speaker's head is our database does not change significantly during articulation. Fig. 9 gives some examples of the mouth localization. The average size of mouth image is around 120 by 70. We know that using a fixed ratio perfect mouth regions cannot be obtained always, so in the future a combination of eye positions and mouth detection will be considered to get more accurate mouth regions.

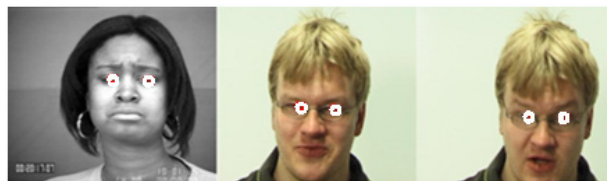


Figure 8. Eye detection examples.



Figure 9. Mouth regions from the dataset.

5.3 Experiments with our dataset

Compared to [26], our extension and application of LBP-TOP to visual speech recognition is original. 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. Also, a novel face and eye detector is used to get the mouth region automatically.

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. Since SVM is only used for separating two sets of points, the 10-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.), 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.

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. After experimenting with different block sizes, we chose to use 1x5x3 (rows by columns by time segments) blocks in our experiments.

For the speaker-independent experiments, leave-one-speaker-out is utilized in the testing procedure. 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).

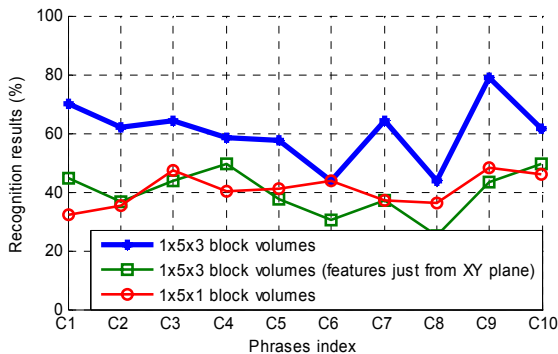


Figure 10. Phrases recognition comparison of different features.

Fig. 10 shows the recognition results using three different features. 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 1x5x1 block volumes omitted the pronunciation order, providing a lower performance than those with 1x5x3 block volumes for almost all the tested phrases. It can be seen from Fig. 10 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 taking 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 2, showing that automatic eye detection gave similar performance as 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).

Table 2. Results of speaker-independent experiments.

Eye detection	Manual	Automatic
Blocks (1x5x3)	60.6%	58.6%
Blocks (1x5x3+1x5x2)	62.4%	59.6%

For speaker-dependent experiments, the leave-one utterance-out is utilized for cross validation because there are not abundant samples for each phrase of each speaker. Totally 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 rest utterances are trained for every speaker. Fig. 11 presents a detailed comparison of the results for every subject. Table 3 shows the overall recognition results. We can see there is no significant difference in performance between automatic eye detection and manual eye positioning.

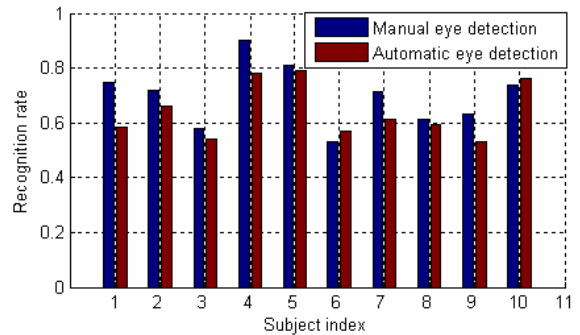


Figure 11. Speaker-dependent recognition results for every subject.

Table 3. Results of speaker-dependent experiments.

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

Most visual speech analysis systems attempt to recognize utterances from data collected in a highly controlled environment with very high resolution. However, in real environments, the obtained mouth image is often of lower resolution. Next, we focus on the effects of image resolution for speech recognition, and evaluate our algorithm over a range of image resolutions. To our knowledge, there is little work done to investigate visual speech recognition under different resolutions. We evaluate our algorithm on the mouth images with the half, one third, one fourth and one sixth (less than 20 by 12 pixels) of the original resolution. Fig. 12 shows the speaker-independent automatic speech recognition results under different resolutions. It can be seen that, with the decrease of resolution, the results do not reduce much.

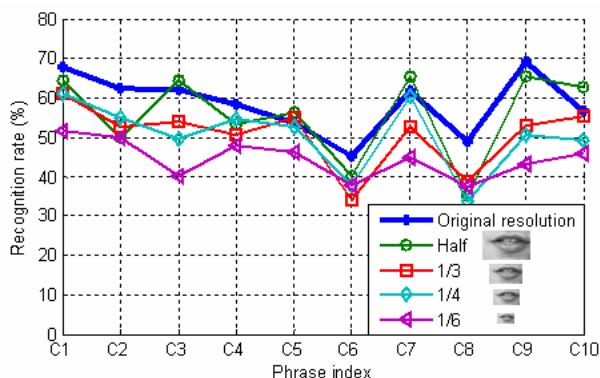


Figure 12. Speaker-independent recognition results under different resolutions.

Unfortunately, it is difficult to compare the performances of different visual speech recognition approaches due to the different speech datasets used and goals of the studies. PCA and other appearance based features focus on extracting global features which omit local properties. Our local patterns can catch the pronunciation transition information in space and time which is good for visual speech recognition. We will emphasize a comparison of our results to those obtained in [20], because in that paper they also try to recognize some isolated phrases using solely visual information. A set of such 20 commands was collected from two speakers which could be used to control an in-car stereo system, for example, “begin scanning”, “turn on the radio”. Each command was recorded three times. The speakers were required to clearly enunciate the phrases at three speeds: slow, medium and fast. Training was done with two speed conditions and testing with the remaining one. Finally, the accuracies over the three trials were averaged. An accuracy of 65.8% was achieved within a totally 120-sequence dataset. But their results are just from two speakers with an emphasis in speed robustness; therefore their evaluation is not speaker-independent. Using our method, we got 62% and 70% for the 817-sequence dataset in speaker-independent and speaker-dependent experiments, respectively, which are promising.

5.4 Experiments with a public database

To compare to some commonly used appearance features like PCA, we also experimented with the Tulips1 audio-visual database [32]. It is a small, publicly available database of 12 subjects, pronouncing the first four digits in English two times in repetition. The video part consists of 934 gray scale lip images of size 100 by 75, sampled at the rate of 30 fps. So we do not need to do the face and eye detection to localize the mouth region.

All experiments and comparison were carried out using the leave-one-speaker-out testing procedure. Training was done on 11 of the 12 speakers in the dataset, while testing was performed on the last one. The procedure was repeated for each one of the 12 speakers, 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), like in the experiments presented in Section 5.3

In [5], PCA features were extracted from the normalized mouth region images and then mutual information was used to do the feature selection. The best results of 81.25% and 87.5% were obtained using HMMs with a mean-removal PCA (MRPCA) and the selected MRPCA from mutual information (MI MRPCA), respectively, as shown in table 4. In [6], the pixels of downsampled images of size 20 by 15 were coupled with their first temporal derivatives, pixel by pixel differences between consecutive frames. Such features were fed into SVM-HMM classifiers. Their result for the visual-only feature was around 80%. The audio and visual features used together gave result of 91% for 10 dB signal to noise ratio (SNR) level. It should be noted that the approaches in [5] and [6] used normalization with respect to head rotation, translation and scaling, as shown in Fig. 13, on the basis of the manually marked mouth images. Our methods do not make any normalization and give superior results, even better than that those from the combined audio and visual features in [6]. This demonstrates robustness to errors in alignment.

Compared with the state-of-the-art, our method does not need to 1) segment lip contours [10,11]; 2) track lips in the subsequent frames; 3) select constant illumination or perform illumination correction [20]; 4) align lip features with respect to the canonical template [9,11] or normalize the mouth images to a fixed size as done by most of the papers [5,10,20]. Furthermore, our method shows stability for low resolution sequences. In this way our experimental setup is more realistic.



Figure 13. Mouth images with translation, scaling and rotation from Tulips1 database.

Table 4. Comparison to other methods on Tulips1 audio-visual database.

	Features	Normalization	Results (%)
[5]	MRPCA	Y	81.25
[5]	MI MRPCA	Y	87.5
[6]	Temporal Derivatives Features	Y	80
Ours	$LBP - TOP_{8,8,8,1,1,1}$ Blocks: 3x6x2	N	92.71

6. 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 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%. The accuracies obtained with different resolutions show the stability of our method. Moreover, experiments on Tulips1 audio-visual database provided a 92.7% accuracy. This clearly outperforms the commonly used PCA features and HMM approach.

Our future plan is to use different classifiers, such as HMM, to deal with not only the isolated phrases, but also the continuous speech, to improve the quality of human-computer interaction. Besides, we would combine eye detection with mouth detection, and use these features for head tracking. With this approach the accuracy and robustness of the mouth region detection could be further improved. 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.

REFERENCES

- [1] Fox N., Gross R. and Chazal P. Person identification using automatic integration of speech, lip and face experts. *ACM SIGMM workshop on Biometrics Methods and Applications*, 2003, 25-32.
- [2] Frischholz R.W. and Dieckmann U. BioID: a multimodal biometric identification system. *Computer*, 33(2), 2000, 64-68.
- [3] Luetttin J., Thacher N.A. and Beet S.W. Speaker identification by lipreading. *International Conference on Spoken Language Proceedings (ICSLP)*, 1996, 62-64.
- [4] Potamianos G., Neti C., Gravier G., Garg A., and Senior A. Recent advances in the automatic recognition of audio-visual speech. *Proc. IEEE*, 2003.
- [5] Arsic I. and Thiran J.P. Mutual information engenlips for audio-visual speech. *14th European Signal Processing Conference*, Italy, 2006.
- [6] Gurban M. and Thiran J.P. Audio-visual speech recognition with a hybrid SVM-HMM system. *13th European Signal Processing Conference*, 2005.
- [7] Lee B., Hasegawa-Johnson M., Goudeseune C., Kamdar S., Borys S., Liu M. and Huang T. AVICAR: Audio-visual speech corpus in a car environment. *ICSLP*, 2004, 2489-2492.
- [8] Gowdy J.N., Subramanya A., Bartels C. and Bilmes J. DBN based multi-stream models for audio-visual speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Canada, 2004, 993-996.
- [9] Aleksic P.S. and Katsaggelos A.K. Product HMMs for audio-visual continuous speech recognition using facial animation parameters. *International Conference on Multimedia and Expo (ICME)*, vol. 2, 2003, 481-484.
- [10] Nefian A.V., Liang L., Pi X., Liu X. and Murphy K. Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing* 2002, 11, 1-15.
- [11] Aleksic P.S., Williams J.J., Wu Z., and Katsaggelos A.K. Audio-visual speech recognition using MPEG-4 compliant visual features. *EURASIP Journal on Applied Signal Processing* 2002, 11, 1213-1227.
- [12] Neti C., Potamianos G., Luetttin J., Matthews I., Glotin H., Vergyri D., Sison J., Mashari A. and Zhou J. *Audio-visual speech recognition*. Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, Final workshop 2000 Report, Oct. 2000.
- [13] Basu S., Neti C., Rajput N., Senior A., Subramaniam L., Verma A. Audio-visual large vocabulary continuous speech recognition in the broadcast domain. *IEEE 3rd Workshop on Multimedia Signal Processing*, 475-481, 1999.
- [14] Niyogi P., Petajan E. and Zhong J. Feature based representation for audio-visual speech recognition. *Audio Visual Speech Conference*, 1999.
- [15] Brooke N.M. Using the visual component in automatic speech recognition. *ICSLP, Vol. 3*, 1996, 1656-1659.
- [16] Duchnowski P., Hunke M., Busching D., Meier U. and Waibel A. Toward movement-invariant automatic lipreading and speech recognition. *ICSLP*, 109-112, 1995.
- [17] McGurk H. and MacDonald J. Hearing lips and seeing voices. *Nature*, vol. 264, 1976, 746-748.
- [18] Potamianos G., Neti C., Luetttin J., and Matthews I. Audio-visual automatic speech recognition: an overview. *Issues in Visual and Audio-Visual Speech Processing*, MIT Press, 2004.
- [19] Potamianos G., Graf H. P. and Cosatto E. An image transform approach for HMM based automatic lipreading. *Proc. of ICIP* 1998, Chicago, Illinois, 1998, 173-177.
- [20] Saenko K., Livescu K., Siracusa M., Wilson K., Glass J. and Darrell T. Visual speech recognition with loosely synchronized feature streams. *ICCV*, 2005, 1424-1431.

- [21] Saenko K., Livescu K., Glass J., and Darrell T. Production domain modeling of pronunciation for visual speech recognition. *ICASSP*, vol. 5, 2005, 473-476.
- [22] Chiou G.I. and Hwang J.N. Lipreading from color video. *IEEE Transactions on Image Processing*, 6(8), 1997, 1192-1195.
- [23] Matthews I., Cootes T.F., Bangham J.A., Cox S., and Harvey R. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2), 2002, 198-213.
- [24] Ojala T., Pietikäinen M., and Mäenpää T. Multiresolution gray scale and rotation invariant texture analysis with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7), 971-987, 2002.
- [25] Ahonen T., Hadid A., and Pietikäinen M. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(12), 2006, 2037-2041.
- [26] Zhao G. and Pietikäinen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6), 2007, 915-928.
- [27] Viola P. and Jones M. Rapid object detection using a boosted cascade of simple features. *CVPR*, 2001, 511-518.
- [28] Heusch G., Rodriguez Y., and Marcel S. Local binary patterns as an image preprocessing for face authentication. *7th International Conference on Automatic Face and Gesture Recognition (FG2006)*, 2006, 9-14.
- [29] Messer K., Matas J., Kittler J., Luetttin J., and Maitre G. Xm2vtsdb: The extended m2vts database. *Second International Conference on Audio and Video-Based Biometric Person Authentication*, Washington, D.C., 1999.
- [30] Sanderson C. The VidTIMIT database. IDIAP Communication 02-06, Martigny, Switzerland, 2002.
- [31] Hazen T., Saenko K., La C. H., and Glass J. A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. *Proc. ICMI*, 2005.
- [32] Movellan J.R. Visual speech recognition with stochastic networks. *Advances in Neural Information Processing Systems*, vol. 7, 1995, pp. 851-858.