A Novel Motion Based Lip Feature Extraction for Lip-reading

Meng Li and Yiu-ming Cheung
Department of Computer Science, Hong Kong Baptist University
mli@comp.hkbu.edu.hk, ymc@comp.hkbu.edu.hk

Abstract

In a lip-reading system, one key issue is how to extract the visual features, which greatly impact on the lip-reading recognition accuracy and efficiency. In this paper, we propose a novel motion based visual feature representation. Compared with the existing methods, our approach focuses on the crucial part of lip movement, but not all pixels around lip contours for different utterance, and captures the motion tracks of each part. Accordingly, distinctive feature vectors are built to represent the whole lip motion process for the specified utterance, rather than the separate frame images. Experimental result shows the efficacy of the proposed approach.

1. Introduction

It is well known that the useful information about speech content can be obtained through analyzing the lip movements of speakers [1]. In 1984, the first automatic lip-reading system was presented by Petajan [2]. From then on, lip-reading has received considerable attention from the community because of its potential attractive applications in speech recognition, secret communication, and so forth.

So far, several methods have been proposed to enhance the performance of an automatic lip-reading system. Although some progress has been made, the recognition rate of lip-reading is still far behind our expectation. One primary reason is that the visual features used in the existing systems cannot represent the entire information conveyed by the lip movement of a specified utterance [3][4].

Actually, the existing methods of visual feature representation can be classified into two categories in view of spatio-temporal relativity — the stationary based methods, and the motion based ones. In the former one, a video of lip movements is split into a sequence of images, on which the shape, appearance or texture of lip in each frame can be used for visual feature representation [6][7][8]. The recent reviews of the existing typical automatic lip-reading system utilizing the "static feature representation" can be found in [4] and [9]. Although a lot of efforts have been spent on improving the performance of system based on "static feature representation", the performance of almost all of them is still quite poor.

On the other hand, along with the progress in motion perception and interpretation field, the motion based feature representation has recently played a more and more important role in automatic lip-reading system [10]. Compared to the "static feature representation", the motion based one can keep more crucial information for recognition. Many of the existing motion based feature representation methods use the optical flow field as the foundation in general. Optical flow arises from relative motion of objects and the viewer is distribution of the apparent velocity of the brightness patterns in an image. It can. Consequently, optical flow can give crucial information regarding the spatial arrangement of the viewed objects and the rate of change of this arrangement [11]. Nevertheless, the existing optical flow based feature representations have some limitations. For instance, they are sensitive to the translation, scaling, rotation, and in particular the change of optical condition. Also, from the viewpoint of algorithm efficiency, the feature vector calculated by optical flow based methods always has high dimension and redundancy. Hence, the speed and efficiency of such an algorithm are relatively low [10].

In this paper, we propose a novel motion based feature representation. Different from the traditional optical flow methods, the proposed one dose not utilize all pixels in each image, but just some landmarks with the obvious lip movement. Since the lip movement for each utterance is different, the positions of landmarks are various in each image. Further, the number of these landmarks may not be constant. When the landmarks are chosen, the corresponding movement tracks will be also captured, whereby the feature is built to represent the lip movements.

978-0-7695-3508-1/08 $25.00 © 2008 IEEE
DOI 10.1109/CIS.2008.214

2. Visual Feature

2.1. Feature Extraction

To extract the features exactly, a pre-processing is performed. We split the video of a number of utterances into an image sequence. The sample of a frame is shown in Figure 1(a), and the contours of nostrils and lips are shown in Figure 1(b). We calculate the center points of both nostrils as the datum mark, and adjust (translation and rotation) all images into the same co-ordinate.

It is known that the most visual information can be conveyed from the lip contours. Hence, the four crucial curves involving upper outer, lower outer, upper inner and lower inner contours of lips are extracted, respectively, as shown in Figure 2. The four curves are numbered as 1 to 4 in order.

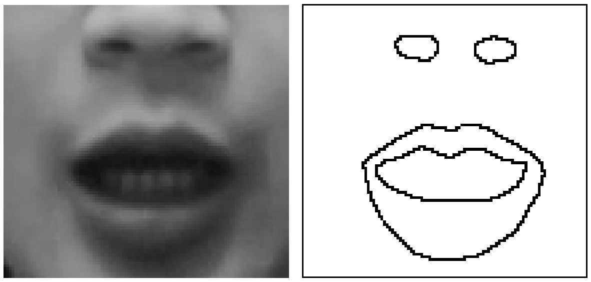

(a) (b)

Figure 1. (a) Lip area in an example frame; (b) the contours of nostrils and lips.

Suppose the deformation of lips in horizontal and vertical directions is linear. That is, the relative position of each macroblock in the entire lip shape is constant during the deformation process. Hence, we obtain the positions of mouth corners, and divide the distance between two corners equally into *n* scales in horizontal. Subsequently, we obtain the landmarks that are defined as the connection by the scale lines (i.e. the vertical lines in Figure 3) and the four crucial curves so as to capture the motion tracks. In some cases, the inner contour cannot be crossed by all scale lines, so we employ an assistant line connect the most left/right points of inner and outer contour relatively to make sure the number of landmark in each crucial curve are constant. Moreover, the landmarks on the assistant lines are belonged to both upper and lower inner contour. Figure 3 illustrates the status with $n=7$.

For a certain utterance, we calculate the landmarks in each frame and build two matrices $C_{i,x}$ and $C_{i,y}$ by utilizing the horizontal and vertical motion of landmark points on each crucial curve respectively.

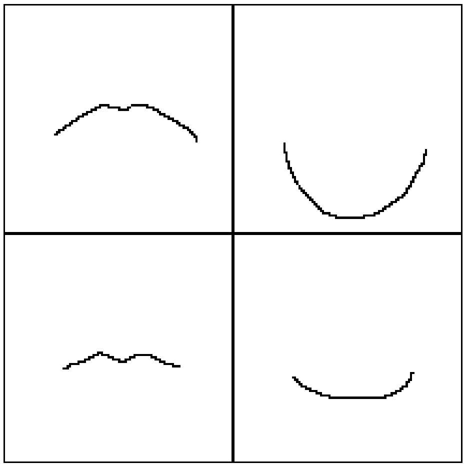


Figure 2. The four crucial curves of lip contour.

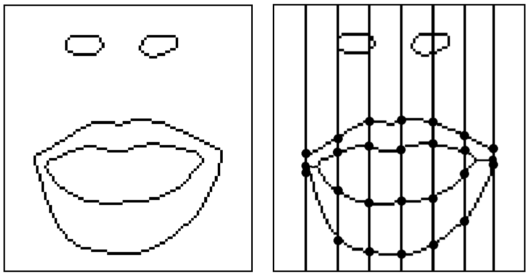


Figure 3. Sample images when the seven scale lines are employed.

$$C_{i,x} = \begin{bmatrix} x_{i,1}^{1} & x_{i,1}^{2} & \cdots & \cdots & x_{i,1}^{k-1} & x_{i,1}^{k} \\ x_{i,2}^{1} & x_{i,2}^{2} & \cdots & \cdots & x_{i,2}^{k-1} & x_{i,2}^{k} \\ & & \vdots & \vdots & & \\ x_{i,n-1}^{1} & x_{i,n-1}^{2} & \cdots & \cdots & x_{i,n-1}^{k-1} & x_{i,n-1}^{k} \\ x_{i,n}^{1} & x_{i,n}^{2} & \cdots & \cdots & x_{i,n}^{k-1} & x_{i,n}^{k} \end{bmatrix}$$

$$C_{i,y} = \begin{bmatrix} y_{i,1}^{1} & y_{i,1}^{2} & \cdots & \cdots & y_{i,1}^{k-1} & y_{i,1}^{k} \\ y_{i,2}^{1} & y_{i,2}^{2} & \cdots & \cdots & y_{i,2}^{k-1} & y_{i,2}^{k} \\ & & \vdots & \vdots & & \\ y_{i,n-1}^{1} & y_{i,n-1}^{2} & \cdots & \cdots & y_{i,n-1}^{k-1} & y_{i,n-1}^{k} \\ y_{i,n}^{1} & y_{i,n}^{2} & \cdots & \cdots & y_{i,n}^{k-1} & y_{i,n}^{k} \end{bmatrix}$$

where $x_{i,j}^{k}$ is the horizontal position of the *j*th point (the most left point is the first) on the *i*th crucial curve extracted from the *k*th frame in the video squence, and

$y_{i,j}^{k}$ is the corresponding vertical one. Hence, for each utterance, we obtain two matrices: $C_{i,x}$ and $C_{i,y}$ (i = 1, 2, 3, and 4), each row vector of $C_{i,x}$ represent a motion track curve of corresponding point in horizontal direction, and row vector of $C_{i,y}$ represent the vertical one. So, each matrix is composed of *n* curves.

Ideally, the motion of each landmark point should be consecutive. Hence, the stochastic reciprocate motion can be regarded as noise. In order to avoid this kind of noise, a low pass filter is therefore employed.

Toward the filtered data, we calculate the correlation coefficients between motion track curve of each nearby two points using the equations below:

$$\bar{x}_{i,j}^{l} = \frac{\sum_{l=1}^{k} x_{i,j}^{l}}{k}$$

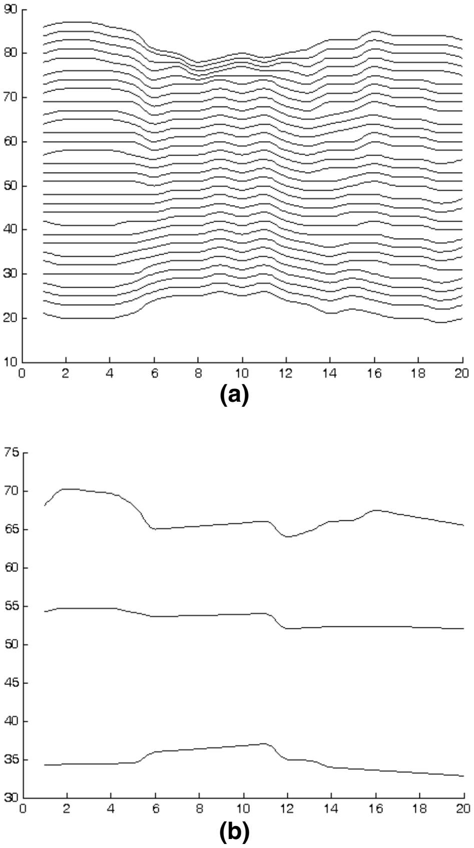
$$r_{x_{j,j+1}} = \frac{\sum_{l=1}^{k} (x_{i,j}^{l} - \bar{x}_{i,j}^{l})(x_{i,j+1}^{l} - \bar{x}_{i,j+1}^{l})}{\sqrt{\sum_{l=1}^{k} (x_{i,j}^{l} - \bar{x}_{i,j}^{l})^{2} \sum_{k=1}^{n} (x_{i,j+1}^{l} - \bar{x}_{i,j+1}^{l})^{2}}}$$

The detailed procedure is described as follows:

- Step 1: Get the motion track curves that should be operated.
- Step 2: Arrange the curves by the index of corresponding landmarks descend.
- Step 3: Suppose dimension of the curve set is *n*, we get the correlation coefficients between each adjacent two points.
- Step 4: For each correlation coefficient, if the value is greater than or equal to a given threshold, the two curves corresponding to the correlation coefficient are replaced by the mean curve of the two.
- Step 5: Update the curve set by the new one, and the dimension of the new curve set is *m*.
- Step 6: If *m* is equal to *n*, we let the curve set be the final one, otherwise go to Step 2. For the curves which are replaced by a “mean curve”, the index, namely *j*, of corresponding landmarks are recorded, and the median of them is marked as “middle point”.

For example, when we analyze the lip movement during the utterance “5” (*n*=30, *k*=20), the horizontal motion track curves of all 30 points on the upper outer contour ($C_{1,x}$) are shown in Figure 4(a). The new set (marked as $C_{1,x}^{'}$) is composed by the three feature curves as shown in Figure 4(b), and the corresponding middle points are 6, 17 and 25.

We use several samples to train our system, and obtain the scatter of middle points for each sample as shown in Figure 5. A set of intervals are utilized to describe these points (marked as $P_{1,x}$).



(a)

(b)

Figure 4. Horizontal motion tracks on the upper outer lip contour when speaking “5”.

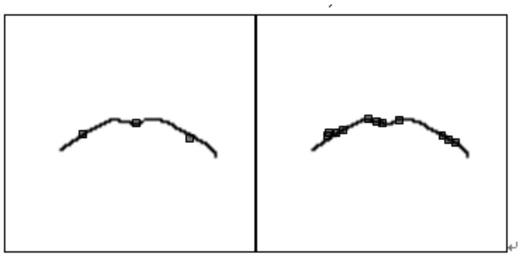


Figure 5. Scatter of middle points, with the initial lip shape overlaid.

The feature representation for each utterance is then shown below:

$$C=\left(C'_{1,x},C'_{1,y}\cdots,C'_{4,x},C'_{4,y}\right)$$
$$P=\left(P_{1,x},P_{1,y},\cdots,P_{4,x},P_{4,y}\right)$$

2.2. Matching

The testing data are processed through the method above, the feature curves and the corresponding middle points are extracted. Compared middle points and intervals of them in training result, several utterances that have the same motion parts on lip contours are screened out.

Since the speaking velocities are different for each people, a cubic spline interpolation algorithm is employed to make *k* in each curve is same.

Subsequently, we calculate the correlation coefficients between the feature curves of testing data and the corresponding curves in the same position for each utterance. The mean correlation coefficient is used to explain the comparability between the two curve sets. Since there are the eight curve sets for one utterance, namely $C'_{1,x},C'_{1,y}\cdots,C'_{4,x},C'_{4,y}$, the corresponding correlation coefficient between them and testing data are marked as $r_{1,x},r_{1,y}\cdots,r_{4,x},r_{4,y}$.

We use the weighted sum, written as *r*, with $r=r_{1,x}+r_{1,y}+r_{2,x}+r_{2,y}+w(r_{3,x}+r_{3,y}+r_{4,x}+r_{4,y})$ to explain the comparability between the training utterance and testing utterance. As it is difficult that the inner contours of lip can be extracted exactly, the weight *w* of $r_{3,x},r_{3,y},r_{4,x},r_{4,y}$ is suggested to use a small value.

When a testing is processed, *r* for each utterance in training set is calculated. The utterance that has the maximum *r* is considered as the class that the testing data belongs to.

4. Experiment Results

Since most of the existing database for lip-reading either are not available for public or do not have the appropriate language samples, a number of research groups develop their own visual-speech database [3][12]. Under the circumstances, we established a database for our experiments. The database consisted of nine speakers (5 males and 4 females). Each speaker uttered nine isolated digits from one to nine in Mandarin Chinese. Short pauses were inserted between each digit utterance for some speakers, but not all. In order to generate the normal lighting condition on speakers, two 36W fluorescent lamps were utilized, where they were placed at the left and right side in top of a speaker, respectively. Furthermore, to test the robustness of our system, several cameras with different resolutions and sample rates were used.

To investigate the efficacy of the proposed feature extraction in lip-reading, we have conducted several experiments. Because of the space limitation, we conducted an experiment to perform multi-speaker recognition task involving isolated digits (0 to 9). We utilized two sets in the database as training data, and the remaining seven data sets as the testing data. Table 1 shows the recognition rate for each digit. The overall recognition rate of our approach on the testing set was 72.9%. To the best of our knowledge, this rate is considerably higher than the existing results reported in the literature.

Table 1: Recognition rate for each digit

1	2	3	4	5	6	7	8	9
71.4%	57.1%	85.7%	57.1%	85.7%	57.1%	57.1%	100%	85.7%

5. Acknowledgement

The work described in this paper was supported by the Faculty Research Grant of Hong Kong Baptist University with the project code: FRG/07-08/II-88.

6. Conclusion

In this paper, a motion based lip feature extraction method has been proposed for a speaker independent lip-reading system. Such a method provides an effective way to extract the features from the videos of lip movement. The proposed approach has been empirically investigated on different speakers. The experiments have shown the promising results.

References

[1] J. Bulwer. *Philocopus, or the Deaf and Dumbe Mans Friend*. Humphrey and Moseley, 1648.

[2] E. D. Petajan. Automatic lipreading to enhance speech recognition. PhD thesis, University of Illinois, 1984.

[3] I. Matthews, T. F. Cootes, J. A. Bangham, et al. Extraction of Visual Features for Lipreading. *IEEE Tran.*

Pattern Analysis and Machine Intelligence, 24(2):198-213, February 2002.

[4] H. X. Yao, W. Gao, et al, A Survey of Lipreading – One of Visual Languages. *ACTA Electronica Sinica*, 29(2): 239-246, February 2001.

[5] M. S. Gray, J. R. Movellan, T. J. Sejnowski. A*dvances in Neural Information Processing Systems*. The MIT Press, 1997.

[6] J. Luettin, N. A. Thacker. Speechreading Using Probabilistic Models. *Computer Vision and Image Understanding*, 65(2):163-178, February 1997.

[7] G. I. Chiou, J. N. Hwang. Lipreading by Using Snakes, Principal Component Analysis and Hidden Markov Models to Recognize Color Motion Video. *IEEE Trans. on Image Processing*, 6(8):1192-1195, 1997.

[8] T. F. Cootes, C. J. Taylor. A Mixture Model for Representing Shape Variation. *Image and Vision Computing*, 17(8):567-574, 1999.

[9] T. Chen, R. R. Rao. Audio-Visual Integration in Multimodal Communication. *Proc. IEEE, Special Issue on Multimedia Signal Processing*, 86(5):837-852, May 1998.

[10] C. Cedras, M. Shah. Motion-Based Recognition: A Survey. *Image and Vision Computing*, 13(2):129-155, March 1995.

[11] J. Barron, D. J. Fleet, S. S. Beauchemin. Performance of Optical Flow Techniques. *International Journal of Computer Vision*, 12(1):43-77, February 1994.

[12] L. Liang, Y. Luo, F. Huang, A. V. Nefian. A Multi-Stream Audio-Video Large-Vocabulary Mandarin Chinese Speech Database. *IEEE International Conference on Multimedia and Expo*, 1787-1790, 2004.