Geometrical Feature Extraction for Robust Speech Recognition

Xiaokun Li and Chiman Kwan
Signal/Image Processing and Control Group
Intelligent Automation, Inc, Rockville, MD, 20855, USA
Email: {xli, ckwan}@i-a-i.com

***Abstract*— Visual information from lip contour has been successfully shown to improve the robustness of automatic speech recognition especially in noisy environments. In this paper, a novel method for lip reading is presented. In the method, hue information of input images is used for lip area detection. Then, a set of morphological operations is applied to detect lip contour. Polynomial fitting is designed for geometrical feature extraction. With the extracted features, Hidden Markov Models and Gaussian Mixture Models are trained to recognize speech. The experimental results demonstrated that the proposed method improved speech recognition rates in noisy environment. Another advantage of the method is its robustness to lighting variances.**

I. INTRODUCTION

The visual information of speaker's mouth provides helpful information for automatic speech recognition (ASR), especially in environments corrupted by acoustic noise and multiple talkers. Such scenarios can be in battlefield, conference room, public transportation places, etc. Motivated by the complementary nature of visual information, many research efforts have been made in the area of audio-visual speech recognition (AVSR). In the area, researchers use images analysis-based methods to detect the speaker's lip movement to recognize speech (called lip-reading) from image sequence or video. Among them, one popular approach called pixel-based feature method is based on the feature extracted from DCT domain. But, it has been found the DCT-based feature is sensitive to the lighting variations. For the same geometrical shape, different DCT coefficients will be obtained with different lighting positions. One example is demonstrated in Fig. 1. Another important approach is to detect shape change of the speaker's mouth, called lip detection based method. The lip detection based methods fit human behavior as people often guess what the speaker is talking about according to the mouth change of the speaker. Many lip detection based methods have been proposed, but most of them are relatively sophisticated, such as the methods of dynamic contour, active shape, and deformable templates [1]-[4], which prohibit real-time analysis. Another drawback of them is the sensitivity to the intensity distribution caused by lighting variation. In recent years, the approach in [5] and [6] that uses color information for lip detection is gaining more interest as color information increases the efficiency and robustness on lip detection. For completed literature review on AVSR, the papers [7] and [8] are highly recommended. In this paper, a new and robust lip detection and feature extraction method is proposed and tested. The method includes color decomposition, morphological operations, polynomial fitting for fast lip contour detection and feature extraction and Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) for speech recognition.

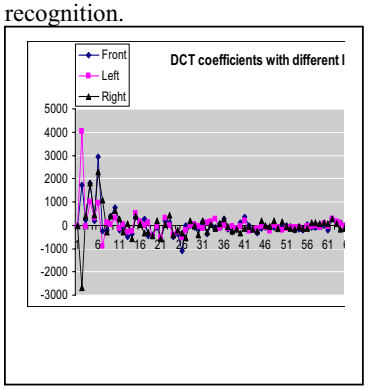


Fig. 1 DCT coefficients of a close mouth with different lighting positions

Note that in the test shown in Fig. 1 the closed mouth was captured under three different lighting settings. In case 1 the light was set on the front of the speaker as shown in the top image of the Fig. 1. In case 2 the light was set on the left side of the speaker as shown in the middle one of the Fig. 1. In case 3 the light was set on the right side of the speaker as shown in the bottom one of the Fig. 1. Three DCT coefficient matrices with size of 64 by 64 are computed from the three lip areas shown in the right part of Fig. 1. The first 101 major coefficients of the DCT Transformation with Zigzag scan scheme are drawn in the left part of Fig. 1. It can be seen that the DCT features have big differences on the same geometrical shape with different lighting settings.

II. ALGORITHM DESCRIPTION

A fast and simple method is proposed, which consists of four steps: lip segmentation and outer contour extraction, ellipse fitting, invariant feature extraction, and speech recognition with HMM/GMM. In this method, the captured video of the speaker will be processed frame by frame. In each frame, the speaker's lip area of the images is firstly segmented by using color decomposition and a double-peak thresholding method. Then, the lip contour is detected and extracted by using a set of morphological operations. Reliable feature points are chosen from the extracted lip contour and a mathematical model (an ellipse) is fitted to these points with least square method (LSM). The invariant geometrical features such as

normalized lip width, height, and the ratio of height and width, are then computed to construct feature vector for each word or command. Frame interpolation is applied to produce the same size of feature vector array (feature matrix) to each word or command. In the off-line stage, the feature matrix will be used to train Hidden Markov Model /Gaussian Mixture Model based recognizer. In on-line stage, the feature matrix will be sent to the recognition part which uses the trained HMMs or GMMs for speech recognition. The algorithm flowchart is given in Fig. 2.

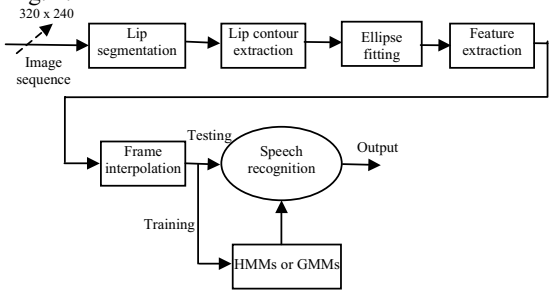


Fig. 2 Algorithm flowchart

A. *Lip Segmentation and Outer Contour Extraction*

To reduce the computational cost, the input images are firstly down-sampled from the original size of 320 by 240 to the size of 160 by 120. Then, the RGB color image (Fig. 3(a)) is transformed to HSI color image which includes three channels, hue (Fig. 3(c)), saturation, and intensity (Fig. 3(b)). Compared with intensity image (gray image), the hue image is relatively constant to lighting variations and different skin/lip colors. Therefore, hue image is chosen for lip segmentation in our research.

To get the outer contour of the lip, we need to segment the lip region from the hue image. Unfortunately, the current image segmentation methods, such as thresholding, region growing, and active/deformable model method, can not be used directly in this case as the lip region contains both very dark and bright pixels as shown in Fig. 3 (c). Therefore, before the segmentation, it is necessary to translate the value of the dark and the bright pixels to a similar value region. Here, an automatic double-peak thresholding method is used for this purpose. Firstly, the pixels of the hue image are classified into two categories by a predefined threshold value, such as 128 in our research to an image with the gray value in the range of 0 - 255. Then, to each category, OTSU thresholding method [9] is implemented to obtain an optimal threshold with between-class variance maximum (BCVM). According to the two new obtained thresholds, we classify the pixels of hue image into three categories, the dark, the middle, and the bright. Then, we assign "1" to the pixels in the dark category or the bright category, and "0" to the pixels in the middle category. In this way, a binary image is generated and the lip area is successfully detected. One example is given in Fig. 3(d).

To extract the outer contour of the lip from the generated binary image, a set of morphological operations is designed and applied in sequential order. The details are given as the followings:

(i) An operation of morphological opening is performed to fill the holes in the lip region and smooth the lip boundary as demonstrated in Fig. 3(e). In our implementation, the shape of the structuring element of the morphological operation is a disk with the size of 5. The opening operation can be described as the following:

$$A \circ B = (A - B) \oplus B \qquad (1)$$

which says that the opening of A by B is the erosion of A by B, and then, by a dilation of the result of B. The detail of the all morphological operations can be found in [10].

(ii) The eroding operation of morphology is applied to get another smaller lip area and saved into another image file which is smaller than the image shown in Fig. 3 (e).

(iii) The lip boundary is obtained by subtracting the new image from the previous image (Fig. 3 (e)). The result is given in Fig. 3 (f).

(iv) The skeleton operation of morphology is used to thin the boundaries of the lip contour to one-pixel level.

(v) Contour following method [11] is employed to track the close contours from the one-pixel boundary image. In our case, the longest closed contour is the outer contour of the lip. In the more general case, preprocessing of the face area is needed. Since the real-time implementation of the above operations are available in many API libraries. Our morphological-operation based lip-detection method can be executed in real-time speed.

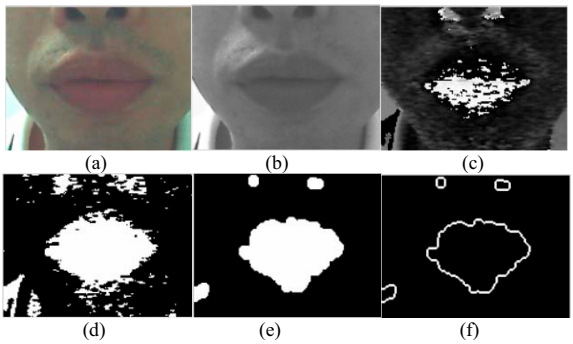


Fig. 3 Example of lip contour detection

B. *Ellipse Fitting*

In automatic speech recognition, the human lipreaders do not care about the exact geometric shape of the lip. They are only interested in the shape changes of the speaker's lip. In other words, they want to know the changes of the lip height and width. For the purpose of measuring the changes, a geometric model, ellipse, is used here to fit the lip boundary to obtain the height and

width of the lip. To a fitted ellipse, its length of long axis is lip width and the length of short axis is lip height. These values will be used as features for speech recognition. One important advantage of the features is they are invariant to general translation and rotation of the speaker's mouth. After scale normalization, they are also invariant on the size change of the lip. The entire process of the ellipse fitting can be described as the following steps:

(i) The boundary points close to the right and the left corners are chosen for ellipse fitting as shown in Fig. 4(a). The reason why only the right and the left part of lip contour are used for fitting is based on the observation that a shadow usually happens at the middle part of the bottom lip as shown in Fig. 4(a). Therefore, the contour information in this area is not reliable and could not be used for ellipse fitting.

(ii) Assume the general ellipse equation is given in (2).

$$ax^2 + bxy + cy^2 + dx + ey = f \qquad (2)$$

where x, y, is the coordinate values of one pixel. Its simple form can be expressed as (3)

$$TA = 1 \qquad (3)$$

where A is the vector of the parameters to be estimated $A = (\frac{a}{f}, \frac{b}{f}, \frac{c}{f}, \frac{d}{f}, \frac{e}{f})'$, T is the vector of (x^2, xy, y^2, x, y).

We put all chosen points (size = N) into (3) and use (4) to get the direct solution or use SVD method [12] to obtain the solution of A.

$$A = (T'T)^{-1}T'B \qquad (4)$$

where T is a matrix of N *by* 5.

(iii) If b is not equal to zero, the ellipse has an orientation to the camera coordinate system. To get the long axis value and the short axis value of the ellipse, conic representation of the ellipse is needed, which is given in (5).

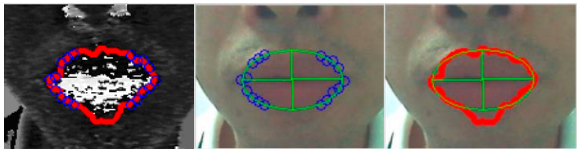
$$a'x^2 + c'y^2 + d'x + e'y = f' \qquad (5)$$

Assume the orientation angle of the ellipse to the camera coordinate system is θ. By replacing x with $\cos(\theta)x + \sin(\theta)y$ and y with $-\cos(\theta)x + \sin(\theta)y$ to the fitted ellipse equation and letting the new b equal to 0, we can get the value of θ.

(iv) By using the square completion method [13] to build the standard representation as (6) from the conic representation,

$$\frac{(x - x_0)^2}{a"^2} + \frac{(y - y_0)^2}{b"^2} = 1 \qquad (6)$$

we can get all coefficients of the standard representation. If $a" > b"$, $a"$ is the length of the long axis, otherwise, $b"$ is the length of the long axis. (x_0, y_0) is the center of the ellipse.



(a) (b) (c)

Fig. 4 Example of ellipse fitting

C. *Feature Extraction*

Assume the length of image sequence of a speech unit is N frames. To save the computational time, we first down-sample the N frames to the size of $N/3$. Then, a feature vector to each down-sampled frame is constructed. The feature vector consists of the values of lip width, height, and the ratio of height/width. To each speech unit, a feature vector matrix with the size of $N/3$ by 3 is constructed. Fig. 5 gives an example of feature extraction. The command in the example is "Call for fire". Eight sample images shown in Fig. 5(a) are chosen from 161 image frames in time sequential order and the constructed feature matrix is drawn in Fig. 5(b) wherein the x-axis is the frame order.

To make the feature matrix invariant to lip scale, the feature vector array is normalized by dividing every feature vector of the vector matrix with the first feature vector.

D. *Speech Recognition with HMMs and GMMs*

Since speech is a temporal continues process and HMM has been proven to be a powerful mathematical model to recognize a process (pattern) which has temporal continuity, we use HMM to model each speech unit. For each speech unit which can be a word, command, or complete sentence, its corresponding Hidden Markov Model can be trained to construct a recognizer in the training (off-line) stage. In our implementation, the speech unit is a command and the left-to-right topology is used for modeling. At the testing stage (on-line stage), the image sequence will be sent to the trained recognizers for speech recognition.

The HMM consists of a finite set of states, each of them associated with a probability distribution that models the distribution of the feature vector. Transitions among the states are governed by a set of probabilities called transition probabilities. To each state, the sum of outcome probabilities is equal to 1. In our work, we set up a HMM which has ten states in a left-to-right topology. We use a single Gaussian function to describe the state-conditional probability distribution.

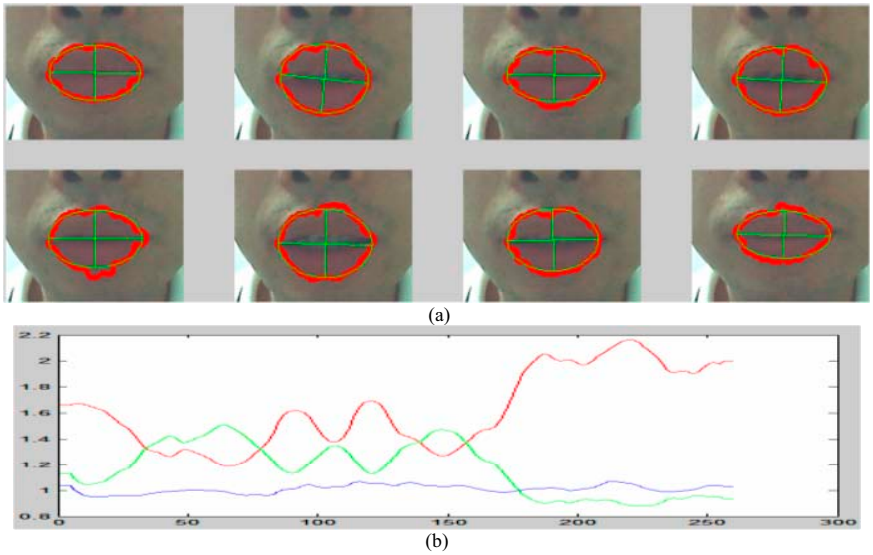


Fig. 5 Example of feature vector array for a command

An expectation-maximization (EM) method called Baum-Welsh algorithm is used to compute the unknown HMM parameters (probabilities). EM algorithms perform an iterative computation of maximum likelihood estimation when the observed data are fed in. The aim of parameter learning is to find the model parameter λ which maximizes $\lambda = \arg\max(\log[p(V \mid \lambda)])$ for a given set V of observed data. The learning process produces a sequence of estimates for λ. Given a set of observed data V, the estimate λ^i has a greater value of $\log[p(V \mid \lambda)]$ than the previous estimate λ^{i-1}. The EM includes two parts:

- Preliminaries

$$\zeta_t(i,j) = P(q_t = i, q_{t+1} = j \mid V, \lambda) \quad (7)$$

$$\gamma_t(i) = P(q_t = i \mid V, \lambda) \quad (8)$$

where $V = \{v_1, ..., v_T\}$ is training sequence and T is the length of training sequence. $\zeta_t(i,j)$ and $\gamma_t(i)$ can be efficiently calculated by the forward-backward algorithm [14].

- Update Rules

$$\overline{\pi_i} = \gamma_1(i) \quad (9)$$

$$\overline{\mu_i} = \frac{\Sigma_{t=1}^{T} v_t \gamma_t(i)}{\Sigma_{t=1}^{T} \gamma_t(i)} \quad (10)$$

$$\overline{\Sigma i} = \frac{\Sigma_{t=1}^{T} \gamma_t(i)(v_t - \overline{\mu_i})(v_t - \overline{\mu_i})^t}{\Sigma_{t=1}^{T} \gamma_t(i)} \quad (11)$$

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T} \zeta_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (12)$$

After setting the initial value to λ, the parameter estimation process will repeat (7)-(12) until the $\log[p(V \mid \lambda)]$ reaches to a local maximum. One advantage of the EM algorithm is the convergence is guaranteed and the convergence time is short. Also, the local maximum is usually an adequate model for the data.

- Recognition

Once the HMM has been trained, speech detection is performed by using the Viterbi algorithm [14], which is a standard technique for pattern recognition. Given a sequence of feature vectors from one input speech unit, the Viterbi algorithm produces the sequence of states which are most likely to have generated these feature vectors with a probability value. We try each trained model with the unit to obtain a probability array and get the recognition result by picking up the one with the highest probability value.

Another technical approach, Gaussian Mixture Model (GMM), was also explored by us. Due to the space limitation, the implementation details can be found in [15]. The testing result of GMM is given in Section III.

III. Experimental Results

In our study, the speech unit is a command. 50 potential battlefield commands were chosen for testing. Each command was repeated 20 times in heavy noisy environments with different lighting situations. (Note that since the noise is non-stationary, the speech recognition system without visual information support completely crashed even with noise cancellation and compensation. Our tests showed the recognition rate decreased from 95% to 10% or even lower.)

In our test, the cross validation method was used. That is, at each test time we used all but one sample for training, then tested with the remaining one. The procedure was then repeated for all samples, which is called Leave-one-out method. To HMM-based recognizer, the average recognition rate was close to 60%. To GMM-based recognizer, the recognition rate was close 40%. With the optimal parameter setting, the recognition rate is expected to have a 10% - 20% improvement. In Table 1, we list the first 10 commands

used in our test and show the recognition results in Fig. 6 and Fig. 7.

Table 1 Command List

1	Call for fire
2	Right 50 down 10 fire for effect
3	I am an American Army officer
4	Use of deadly force is authorized
5	What is your name
6	Where are you going
7	Have you seen any military forces
8	Taking fire, snipers
9	Combined operations
10	Four six seven five

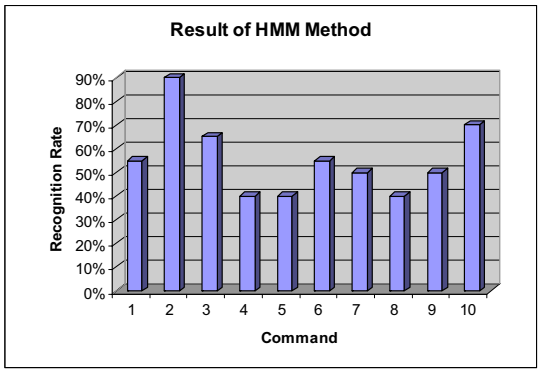


Fig. 6 Recognition rate using HMM method

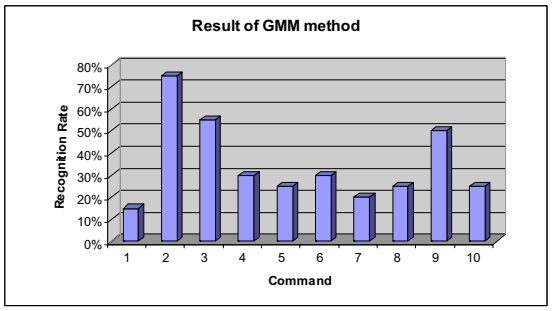


Fig. 7 Recognition rate using GMM method

IV. DISCUSSION

From the testing results, it can be seen that the contour of outer lip can provide a strong support for speech recognition. For the HMM method, the average recognition rate is close to 60%. For the GMM method, the recognition rate is around 40%. With the optimal parameter setting, the recognition rate is expected higher than the current. When the traditional speech recognition system crashes in a heavy noisy environment, the speech recognition system with our proposed method can still provide an acceptable recognition result. Integrating the visual and audio feature together to construct a huge feature vector for speech recognition is another approach to provide robust speech recognition in noisy environment.

It is known from perceptual studies that human lipreaders rely heavily on the information about the presence/absence of the teeth and the tongue inside the mouth along with the information of outer lip [16]. Therefore, the information of inner area of mouth is another important feature for speech recognition. The inner information includes the inner lip contour, teeth position, and tongue potion. But, as we can see from the example images, it is difficult to obtain the reliable inner information with the current available image segmentation methods. More research efforts are needed in the area to develop efficient methods for the inner information detection.

ACKNOWLEDGMENT

The authors would like to thank the support from Army Research Lab under the contract DAAD17-01-C-0075 and Mr. Pete Fisher for his helpful suggestions.

REFERENCES

[1] C. Bregler and Y. Konig, "Eignelips for Robust Speech Recognition," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 669-672, 1994

[2] T. Coianiz, L. Torresani, and b. Caprile, "2D Deformable Models for Visual Speech Analysis," *IEEE Trans. Speech and Audio Processing*, pp. 391-398, 1996

[3] T. F. Cootes, G. J. Edwards, and C.J. Taylor, "Active Appearance Models," *Proc. European Conf. Computer Vision*, pp. 484-498, 1998.

[4] A. Liew, S. H. Leung, W. H. Lau, "Lip Contour Extraction from Color Images Using a Deformable Model," *Pattern Recognition*, Vol. 35, pp. 2949-2962, 2002

[5] M. Lievin and F. Luthon, "Unsupervised Lip Segmentation under Natural conditions," *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 3065-3068, 1999

[6] X. Zhang, R. M. Mersereau, "Lip Feature Extraction Towards an Automatic Speechreading System," *IEEE Int. Conf. on Image Processing,* pp. 226-229, 2000.

[7] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, R. Harvey, "Extraction of Visual Features for Lipreading," *IEEE Trans. PAMI*, Vol. 24, No.2, pp. 198-213, 2002.

[8] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech", *In Proc. IEEE*, Vol. 91, No.9, 2003.

[9] Otsu N, "A threshold selection method from gray-level histograms," *IEEE Trans. Sys. Man Cyber.* Vol. 9(1), pp. 62-66, 1979

[10] R. C. Gonzalez, R. E. woods, *Digital Image Processing*, 2nd , Addison-Wesley, 1998.

[11] Anil K. Jain, *Fundamentals of digital image processing*, Prentice Hall, 1989.

[12] M. E. Hochstenbach, "A Jacobi-Davidson Type SVD Method," *SIAM Journal on Scientific Computing,* Vol. 23, No. 2, pp. 606-628, 2001

[13] M. L. Bittinger, D. Ellenbogen, B. L. Johnson, D. J. Ellenbogen, *Elementary and Intermediate Algebra: concepts and Applications a Combined Approach*, Addison-Wesley, 2002.

[14] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *In Proc. IEEE*, Vol. 77, No.2, pp. 257-286, Feb. 1989.

[15] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker Inc., New York, 1988.

[16] Q. Summerfield, A. MacLeod, M. McGrath, and M. Brooke, "Lips, Teeth and the Benefits of Lipreading," *Handbook of Research on Face Processing*, pp. 223-233, Elsevier Science Publishers, 1989.