LIP FEATURES SELECTION WITH APPLICATION TO PERSON AUTHENTICATION

L.L. Mok†, W.H. Lau†, S.H. Leung, S.L. Wang† and H. Yan†*

†Department of Computer Engineering and Information Technology
*Department of Electronic Engineering
City University of Hong Kong, 83 Tat Chee Avenue, Hong Kong

ABSTRACT

Authentication system solely based on visual lip information is of advantage since the uttering characteristics/manner is unique to individual. This paper presents the results of studying for the most appropriate features extracted from the lip motion for person authentication application. The geometric-based, shape-based and inner lip features are derived from a 14-point Active Shape Model (ASM) lip model. The dynamic features reflecting the differential changes of the feature parameters are also considered in the experiment. A database consists of 40 speakers and each of their utterance last for 3 seconds. Various combinations of features obtained by analyzing this database are fed into a Hidden Markov Model (HMM) classifier for processing. It is observed that the best result has been obtained when both shape-based features and inner lip features are used.

1. INTRODUCTION

Person authentication is a process to match the claimed person's identity with the information stored in the accessing system and the candidate will either be accepted or rejected depending on the matching scores. Speech signal is one of the information to achieve this purpose. However, it may not be easy to have a suitable environment to obtain clean speech signal for processing and the accessing system will have to deal with the noise problems. Visual lip information, on the other hand, is an alternative choice since (i) individual will have different uttering characteristics/manner which is reflected in the lip motion, and (ii) the noise problems will automatically be eliminated. The main focus of this research is to examine the appropriateness of various visual lip features for the application of person authentication.

Image-based and model-based approaches [1,2] are commonly used for lip segmentation and feature extraction. However, model-based approach will be of advantage since only a small set of parameters is required to represent the lip features which are normally invariant to translation, rotation, scale and illumination. The lip features may be classified into geometric-based and shape-based features and they will use different kinds of parameters to describe the lip shape. In this study, the parameters extracted from a 14-point Active Shape Model (ASM) lip model [3,4] will be used to form the geometric- and shape-based features. The inner lip dimensions and teeth area are derived to provide additional information. The dynamic changes of the relevant parameters are also considered in the experiment. Various combinations of these features are fed into a Hidden Markov Model (HMM) classifier for processing. The flowchart of the authentication system is shown in Fig. 1. The candidate will be accepted if the scores exceed a certain threshold. The equal error rate (EER) which is defined as the point that the false acceptance rate (FAR) is equal to the false rejected rate (FRR) for various feature combinations will be examined to justify their suitability for authentication application.

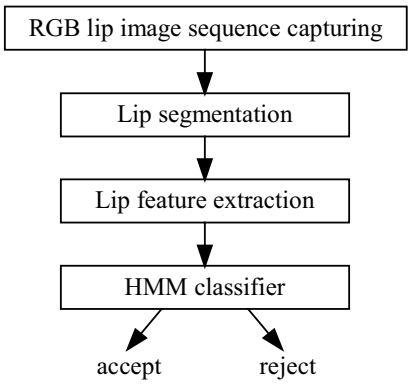


Fig.1 Flowchart of the speaker authentication system

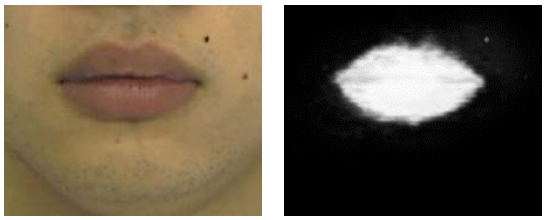
2. LIP MODELING

2.1 Lip Segmentation

In our system, the RGB lip images are captured by a video camera. Since the color distance between the two pixels in the non-uniform RGB color space is not proportional to their color difference, the lip image is required to be transformed to the approximately uniform CIELAB color space [5]. In CIELAB color space, each pixel is

The work described in this paper is fully supported by a research grant (CityU 1215/01E) from the RGC of the HKSAR, China.

represented by a color feature vector $\{L, a, b\}$ with luminance and chrominance components denoted by L, a, b, respectively. We then apply our recently developed Fuzzy Clustering Method incorporating with Shape function (FCMS) [3] to segment the lip image. This method takes both the color and shape information into account and thus gives a more accurate probability map, as shown in Fig. 2, for subsequent modeling.



(a) (b)

Fig.2 (a) A RGB lip image, **(b)** Probability map of (a)

2.2 Lip Model

A 14-point ASM lip model as shown in Fig. 3 is used to describe the lip shape for our system. ASM is a shape-constrained iterative fitting approach [6]. The valid lip shape is allowed to deform within the main deformation modes which are obtained from a training data set of lip images via Principle Component Analysis. One major advantage of using ASM is that no heuristic assumptions are made to the legal shape deformation. In addition, the ASM is flexible enough to capture the shape details with the use of a linear combination of a small set of deformation modes. In general, the coordinates of the contour points of an arbitrary shape **x** represented by ASM can be approximated by (1).

$$\mathrm{x} = \bar{\mathrm{x}} + \mathbf{Pb} \qquad (1)$$

where $\bar{\mathrm{x}}$ is the mean shape, **P** is the matrix of eigenvectors of covariance matrix and **b** is the weight vector for each eigenvector. After the optimization process, an optimal parameter set given in (2) for describing the lip shape can be obtained.

$$\lambda = \{s, \theta, x_c, y_c, \mathbf{b}\} \qquad (2)$$

where (x_c, y_c) is the center point of the lip model, s is a scaling factor and θ is the rotation angle. Examples of lip contour fitting results are given in Fig. 4.

2.3 Inner lip key points detection

In order to evaluate the importance of the inner lip information for authentication, 4 key points as shown in Fig. 3 are located to help establish the inner lip parameters. 2 key points are located on the line joining points 3 and 10. The pixels marked with **A** and **B** are with gradient change in L exceeding a preset value and they represent the boundary points for the upper and lower inner lip, respectively. Similarly, the left and right inner lip corner points, **C** and **D**, are located by detecting the gradient change along the line joining points 6 and 14. An example of the four inner lip key points is shown in Fig. 5.

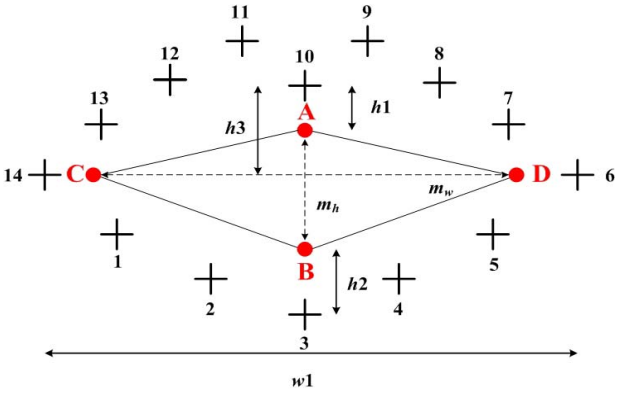


Fig. 3 The 14-point lip model and the inner lip key points

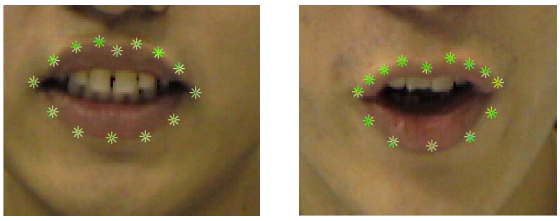


Fig. 4 Examples of outer lip contour fitting results for different speakers

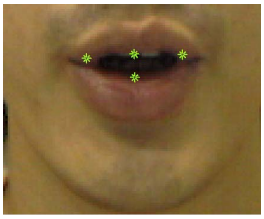


Fig. 5 The four inner lip key points

3. LIP FEATURE EXTRACTION

Two kinds of features can be used to characterize the shape of a lip, namely the geometric-based and shape-based features. The geometric-based features generally include the height and width of both the inner and outer lip to describe the geometric dimension of the lip. Whereas the shape-based features include the information related to the shape deformation modes to describe the shape of a lip. The kind of features used for implementing the authentication system is much dependent on the information available from the features extraction algorithm. With the lip features obtained from our extraction algorithm using the ASM lip model, both the geometric-based and shape-based parameters are available for examining their suitability in authentication. Furthermore, the dynamic features and teeth information will also be considered.

3.1 Geometric-based lip features

X. Zhang [7] has shown that geometric-based lip features can be used to achieve good speaker authentication results. With the 14-point ASM lip model, the geometric-based feature parameters can easily be obtained. As shown in Fig. 3, *w1* represents the width of the lip and is the distance between points 6 and 14. The thickness of the upper lip *h1* is the distance between points **A** and 10. Likewise, the thickness of the lower lip *h2* is the distance between points **B** and 3. *h3* represents the distance between the horizontal lip line and the upper outer lip and is the distance between the lip model center and point 10. Since these distance features are greatly influenced by the object-camera distance, simply incorporating these parameters in the feature vector is unable to provide reliable performance. By normalizing these parameters with respect to the first image of the lip image sequence, $h1_n$, $h2_n$, $h3_n$ and $w1_n$ will be independent to the object-camera distance and become useful and reliable information. Nevertheless, these parameters are static information for individual image, we will also incorporate their dynamic changes in the feature vector in order to closely reflect the uttering characteristics/manners of individuals. The dynamic features $h1_n'$, $h2_n'$, $h3_n'$ and $w1_n'$ represent their corresponding differential changes between adjacent frames. Finally, the visual feature vector f_{GEO}' incorporating the geometric-based lip features and their corresponding dynamic features is given in (3). It should be noted that from this point onward differential visual feature parameter is designated with an apostrophe.

$$f_{GEO}' = \{ h1_n, h2_n, h3_n, w1_n, h1_n', h2_n', h3_n', w1_n' \} \quad (3)$$

3.2 Shape-based lip feature

The outer lip shape is described by the ASM lip model parameters given in (2). The weight vector **b** containing the shape information plays an important role in distinguishing different lip shapes. Since the first few eigenvectors corresponding to the largest eigenvalues dominate the shape variation, the first three weights are included in the feature vector for our authentication system. In addition, the scaling factor s and rotation angle θ can provide useful information for the authentication. Same as the geometric-based features, normalized scaling factor s_n and rotation angle θ_n are required to eliminate the undesirable object-camera distance effects to the shape-based features. By incorporating the differential changes of these parameters to the basic shape-based visual feature vector, f_{ASM}' for describing the outer lip contour is given by:

$$f_{ASM}' = \{s_n, \theta_n, \mathbf{b}_3, s_n', \theta_n', \mathbf{b}_3' \} \quad (4)$$

3.3 Inner lip features

Tongue is known to provide important information for person authentication. However, it is difficult to obtain this information accurately since the color and the luminance are very similar to that of the lip in some circumstances. Instead, we will consider other inner lip information for the authentication. As shown in Fig. 3, the inner lip height m_h is defined as the distance between points **A** and **B**. The inner lip width m_w is the distance between points **C** and **D**. The m_h and m_w are normalized against the height (the distance between point 3 and point 10) and the width (the distance between point 6 and point 14) of the mouth region, respectively. To quantify the teeth information, we first locate the teeth pixels inside the inner lip region that is approximated by the diamond shape **ABCD** [8]. Pixels with luminance higher than t_L in (5) and chrominance lower than t_a in (6) are regarded as teeth.

$$t_L = \mu_L + \sigma_L \quad (5)$$

$$t_a = \mu_a \quad (6)$$

where σ_L is the standard deviation of L, μ_L and μ_a are the means of L and a inside the diamond shape, respectively.

Fig. 6 shows the result of the teeth detection method. The total number of teeth pixels represented by t_{area} is normalized against the total number of pixels bounded by **ABCD**. With the normalized and differential parameters $m_{h,n}$, $m_{w,n}$, $t_{area,n}$, $m_{h,n}'$, $m_{w,n}'$ and $t_{area,n}$', the final inner lip feature vector f_{inner}' is given in (7).

$$f_{inner}' = \{ m_{h,n}, m_{w,n}, t_{area,n}, m_{h,n}', m_{w,n}', t_{area,n}' \} \quad (7)$$

(a) (b)

Fig.6 (a) The original lip image, **(b)** teeth region marked with white color

4. EXPERIMENTAL RESULTS

A video camera is used to capture the frontal view of the mouth region of speakers in 25 fps with size 110 x 90 for 3 seconds without data compression. It is assuming that the speaking pace of all speakers in the database is more or less the same during the 3-second recording. The database consists of 40 speakers with 29 males and 11 females and each of them utters the same phrase three-seven-two-five (3725) ten times in English. Four sets of data are used for training and the remaining is used for testing. In the experiment, all the utterances not belonging to the subject are used as imposter data. Having extracted the feature vectors from these image sequences, a left to right, six states and continuous density HMM classifier with diagonal covariance matrix Gaussian modes associated with each state is used for the authentication.

For authentication system involves speech signal, the continuous speech will be subdivided into isolated digit and HMM models for each digit have to be built.

Whereas for our authentication system solely based on visual information, the one model approach is adopted to represent the utterance of each speaker and this makes the system design much simpler and realistic. In order to investigate the appropriateness of various visual features for speaker authentication through a short period of continuous speech, various combinations of these features will be examined. The equal error rate (EER) is used to quantify their performance. The feature combinations are categorized into geometric-based and shape-based features and the results are given in Tables 1 and 2.

Geometric-based features	EER (%)
f_{GEO}'	16.7
$f_{GEO}' + m_{h,n} + m_{h,n}'$	14.3
$f_{GEO}' + m_{w,n} + m_{w,n}'$	14.6
$f_{GEO}' + m_{h,n} + m_{h,n}' + m_{w,n} + m_{w,n}'$	12.1
$f_{GEO}' + t_{area,n} + t_{area,n}'$	13.7
$f_{GEO}' + m_{h,n} + m_{h,n}' + t_{area,n} + t_{area,n}'$	11.2
$f_{GEO}' + m_{w,n} + m_{w,n}' + t_{area,n} + t_{area,n}'$	11.5
$f_{GEO}' + f_{inner}'$	10.8

Table 1. Speaker authentication results for various geometric-based features combinations

Shape-based features	EER (%)
$\mathbf{b}_3 + \mathbf{b}_3'$	8.8
f_{ASM}'	8.3
$f_{ASM}' + m_{h,n} + m_{h,n}'$	7.2
$f_{ASM}' + m_{w,n} + m_{w,n}'$	7.3
$f_{ASM}' + m_{h,n} + m_{h,n}' + m_{w,n} + m_{w,n}'$	6.1
$f_{ASM}' + t_{area,n} + t_{area,n}'$	6.7
$f_{ASM}' + m_{h,n} + m_{h,n}' + t_{area,n} + t_{area,n}'$	5.6
$f_{ASM}' + m_{w,n} + m_{w,n}' + t_{area,n} + t_{area,n}'$	5.7
$f_{ASM}' + f_{inner}'$	5.1

Table 2. Speaker authentication results for various shape-based features combinations

From the experimental results, it is observed that using the shape-based lip visual features alone in discriminating person identity has already outperformed that uses both the geometric-based and inner lip features. This is intuitively valid since lips having similar geometric dimensions may have very different shapes which is an important information for authentication. For the shape-based features approach, the weights associated with the first 3 deformation modes provide most of the information for authentication since they "capture" the lip shapes and the changes in speakers' utterance. The addition of scaling factor and the rotation angle can provide moderate improvement since all the images are captured with similar object-camera distance and the speakers have negligible movements during speaking.

Comparing the results in both tables, it is observed that the introduction of various inner lip features has produced similar trend of improvement for both the geometric-based and shape-based features. The improvement of only adding the inner lip height or width to the feature vector is significantly lower than that of adding both features to the feature vector. The results also show that the information of teeth area can improve the authentication substantially. However, the information of the inner lip dimensions is seemed to be more important than that of the teeth area. The best authentication results are achieved when both the inner lip dimensions and teeth area are added to the feature vector for both approaches. Since the lip shape features and the inner lip features characterize the lip shape and the uttering characteristics/manners of an individual, the shape-based visual feature vector composed of f_{ASM}' and f_{inner}' is the best choice for person authentication and a lowest EER of 5.1% has been achieved.

5. CONCLUSION

In this paper, we have presented the person authentication results solely based on lip visual features. With the aid of our previously developed lip extraction algorithm, the geometric-based, shape-based and inner lip features can be established. Various combinations of these features have been investigated for their appropriateness for person authentication. The results indicate that shape-based visual features are more suitable and the introduction of the inner lip features can significantly improve the result.

6. REFERENCES

[1] I. Mattews, T.F Cootes, J.A Bangham, S. Cox, R. Harvey, "Extraction of visual features for lipreading", *IEEE Trans on PAMI*, vol.24, pp.198-213, Feb. 2002

[2] T.A. Faruquie, A. Majumdar, N. Rajput, L.V. Subramaniam, "Large vocabulary audio-visual speech recognition using active shape models", *Proc of IEEE Int'l Conf on Pattern Recognition*, vol.3, pp.106-109, Barcelona, Sept. 2000

[3] S.L. Wang, S.H. Leung and W.H. Lau, "Lip segmentation by fuzzy clustering incorporating with shape function", *Proc of IEEE ICASSP*, vol.1, pp.1077-1080, Orlando, May 2002

[4] K.L. Sum, W.H. Lau, S.H. Leung, A.W.C Liew, K.W. Tse, "A new optimization procedure for extracting the point-based lip contour using active shape model", *Proc of IEEE ICASSP*, vol.3, pp.1485-1488, Salt Lake City, May 2001,

[5] R. W. G.. Hunt, *Measuring Color*, 2nd Ed., Ellis Horwood Series in Applied Science and Industrial Technology, Ellis Horwood Ltd., 1991.

[6] J. Luettin, Neil A. Thacker and Steve W. Beet, "Active Shape Models for Visual Speech Analysis", *Speechreading by Humans and Machines*, Springer, 1996.

[7] X.Zhang, R.M.Mersereau and M. Clements, "Automatic speechreading with application to speaker verification", *Proc of ICASSP*, vol.1 , pp.685-688, May 2002

[8] A. W. C. Liew, S. H. Leung and W. H. Lau, "Segmentation of Color Lip Images by Spatial Fuzzy Clustering,", *IEEE Trans on Fuzzy Systems*, vol.11, pp. 542-549, Aug 2003.