PERSON AUTHENTICATION USING ASM BASED LIP SHAPE AND INTENSITY INFORMATION

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 and difficult to imitate. This paper will present the study of using lip shape-based and intensity features in person authentication. These features are derived from a 14-point Active Shape Model (ASM) lip model with the use of Principal Component Analysis (PCA). The differential change of the feature parameters reflecting the uttering characteristics are also considered in the study. A database containing the *visual utterance* of 40 speakers has been generated and each of the utterance is of duration 3 seconds. The visual features are then extracted from this database and a Hidden Markov Model (HMM) classifier is used to perform the analysis. It is observed that the best authentication result is obtained when the first 8 modes of intensity profile is used together with the lip shaped-based parameters.

1. INTRODUCTION

In recent year, the use of biometrics in person authentication has become an important area for research. Apart from iris, face, voice and hand geometry, fingerprint is perhaps the most popular feature being used [1,2]. However, human beings do possess certain habits that are unique to themselves and difficult to be imitated. Gait and speaking characteristics are potential candidates for authentication purpose. Lip features and its associated intensity information have been successfully demonstrated for speaker identification [3,4]. In this paper, we will present the study of using visual lip features, which are closely associated with speaking characteristics, for person authentication.

The geometric width/height of the lip and the shape parameters of the lip are commonly used to describe the lip features for authentication. These features can be extracted from a lip model and a 14-point Active Shape Model (ASM) lip model is used in our study. In this model, points are used to represent the outer lip contour which can be adjusted by varying a few main modes of shape variation. Thus only a few parameters are required to describe the lip shape. In order to explore the merit of the intensity information in authentication, the intensity information about the horizontal and vertical lip lines, which are the lines for measuring the width and height of the lip, are sampled. These information are used to form an global intensity vector and the Principal Component Analysis (PCA) is then employed to obtain an intensity feature vector with reduced dimension.

The flowchart of a Hidden Markov Model (HMM) classifier based authentication system is shown in Fig. 1. The candidate will be accepted if the similarity score (the log likelihood-ratio between the scores of the speaker model and the speaker-independent model) exceeds a certain threshold. The performance of various feature combinations will be examined for their suitability in authentication application by measuring the equal error rate (EER), which is defined as the point that the false acceptance rate is equal to the false rejected rate.

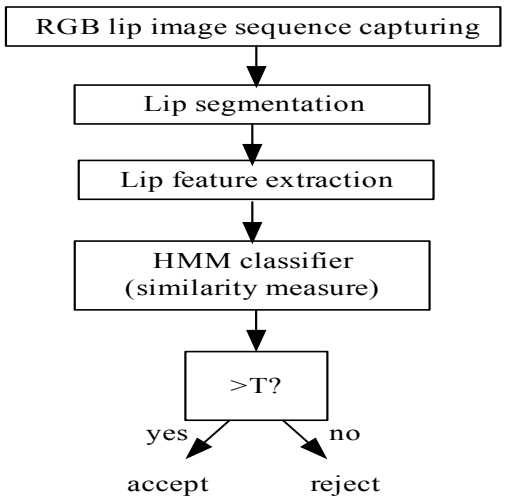


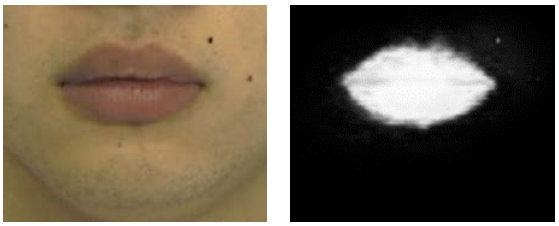
Fig.1 Flowchart of the speaker authentication system

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

2. LIP MODELING

2.1 Lip Segmentation

The RGB lip image is captured by a video camera and transformed to the CIELAB color space [5]. In CIELAB color space, each pixel is 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) [6] to segment the lip image. This method takes both the color and shape information into account and thus gives an 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 shown in Fig. 3 is used to describe the lip shape. ASM is a shape-constrained iterative fitting approach [7]. 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 PCA. 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 $\mathbf{x}$ represented by ASM can be approximated by (1).

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b} \qquad (1)$$

where $\mathbf{x} = [x_1 y_1 x_2 y_2 \cdots x_{14} y_{14}]^T$, $\bar{\mathbf{x}}$ is the mean shape, $\mathbf{P}$ is the matrix of eigenvectors of covariance matrix and $\mathbf{b}$ is the weight vector for each eigenvector. In order to better describe the lip contour, more points are used to describe the upper lip and lesser points for the lower lip. More points could be used for the lip modeling, however it is not considered because too much detail to the lip contour will eventually reduce the modeling flexibility. Also it will require more training samples than it is actually required. After the optimization process [8], 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.

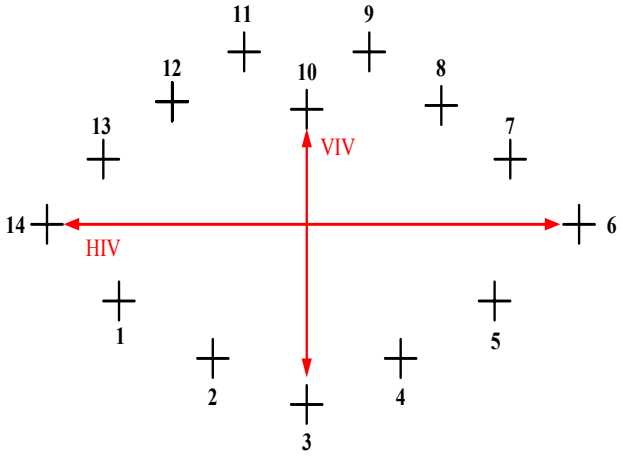


Fig.3 The 14-point ASM lip model with the horizontal and vertical intensity feature lines

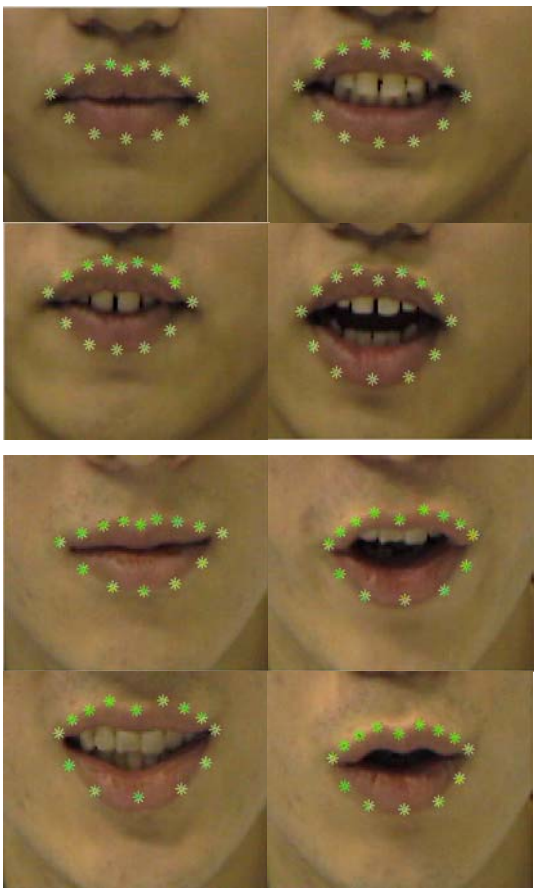


Fig. 4 Examples of outer lip contour fitting results for different speakers and different lip shapes

3. LIP FEATURE EXTRACTION

3.1 Shape-based lip features

The outer lip contour 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. 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, s_n and θ_n will be independent to the object-camera distance and become useful and reliable information. To further provide the dynamic speaking characteristics of the speaker to the authentication system, the differential changes of all these parameters are incorporated into the visual lip feature vector f_{ASM} as shown in (3). It should be noted that the differential parameters are designated with an apostrophe from this point onward.

$$f_{ASM} = \{s_n, \theta_n, \mathbf{b}_3, s_n', \theta_n', \mathbf{b}_3'\} \quad (3)$$

3.2 Lip intensity features

The lip height and width are widely used in visual speech recognition and speaker identification since these features can describe the geometric information during speaker uttering. However, experimental results show that recognition accuracy based on these geometric parameters is lower than that of the ASM shape-based features [9]. In fact, together with the scaling factor s_n, the ASM shape-based features not only provides the shape information, but also includes the geometric information. For our approach, instead of using the geometric information directly, we will use the intensity profile along the horizontal line measuring the width and the vertical line measuring the height of the outer lip contour to enhance the authentication performance. By sampling the line joining points 3 and 10, a 15 points vertical intensity vector (VIV) is formed, as shown in Fig 3. Likewise, a 35 points horizontal intensity vector (HIV) is formed by sampling the line joining points 6 and 14. These intensity profile vectors provide critical information on speaking characteristics.

The VIV and HIV are then concatenated together to form a global intensity vector **h.** Since the order of **h** is 50 and is too high for processing in practice, a reduced dimension feature vector is then derived based on PCA over a set of global intensity vectors obtained from a representative training set covering all speakers in the database. In fact, the same approach for obtaining **b** is used to derive a new intensity profile weighting factor $\mathbf{c}_n$ as follows:

$$\mathbf{h} = \bar{\mathbf{h}} + \mathbf{P}\mathbf{c}_n \quad (4)$$

where $\bar{\mathbf{h}}$ is the mean intensity profile of the training set, **P** is a 50 × n matrix whose n columns are the eigenvectors corresponding to the largest n eigenvalues of the covariance matrix. In order to incorporate the dynamic information for authentication, the differential changes of $\mathbf{c}_n$ is also incorporated into the intensity feature vector $f_{I(n)}$ as follows:

$$f_{I(n)} = \{\mathbf{c}_n, \mathbf{c}_n'\} \quad (5)$$

4. EXPERIMENTAL RESULTS

A video camera is used to capture the frontal view of the mouth region of a speaker with 25 fps and size of 110 x 90 for 3 seconds without data compression. It is assuming that the speaking pace of all speakers is more or less the same during the 3-second recording. The database consists of the *visual utterance* 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. 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 recognition and authentication experiment. In many related studies, the experimental result is usually obtained by analyzing a pre-assigned training set and testing set for a given database [3,4,10-12]. The same result is not guaranteed when using another training set and testing set from the same database. In order to obtain a more representative result, we randomize the data selection for each speaker and arbitrarily pick up a training set for each experiment. The experiment is repeated for a couple of times. The final result is obtained by averaging the rates of all the experiments.

4.1 Recognition experiment

To realize an authentication system, recognition accuracy of the selected features is of primarily concern. The recognition experiment in speaker dependent test has been carried out. Four sets of data have been randomly selected to train the speaker model for each speaker in the database. The remaining data is used for testing. The experiment is repeated for 20 times and the result presented in Table 1 is the average rate of these experiments.

It is observed from the experimental results that the recognition rate using the shape-based lip feature can be improved considerably by incorporating the intensity features. The recognition rate of using shape-based lip features alone is only 90.23%. A 5% improvement has been achieved when just adding the first mode of intensity

features. However, it has been noted that the improvement in recognition rate for adding more than 4 modes of intensity features is relatively insignificant. The main reason is that the change of the intensity profile within such a short duration of speech is not very large. The first few modes of intensity variation are sufficient to represent most of the characteristics of speaker utterance. The results of adding the first 5, 6 and 7 modes of intensity features are roughly the same. The accuracy is further increased to 98.09% when the first 8 modes of intensity features are used. Further increasing the modes of intensity features has practically made no improvement to the recognition rate.

Features	**Recognition** Accuracy (%)	**Authentication** EER (%)
f_{ASM}	90.23	6.25
$f_{ASM} + f_{I(1)}$	95.39	3.99
$f_{ASM} + f_{I(2)}$	96.06	3.68
$f_{ASM} + f_{I(3)}$	97.08	3.14
$f_{ASM} + f_{I(4)}$	97.71	2.57
$f_{ASM} + f_{I(5)}$	97.80	2.54
$f_{ASM} + f_{I(6)}$	97.87	2.47
$f_{ASM} + f_{I(7)}$	97.93	2.27
$f_{ASM} + f_{I(8)}$	98.09	2.17
$f_{ASM} + f_{I(9)}$	98.04	2.18
$f_{ASM} + f_{I(10)}$	98.07	2.17

Table 1. Speaker recognition and authentication results

4.2 Authentication experiment

In the speaker authentication experiment, all the utterances in the database are used to train the speaker-independent model for the similarity measure. The similarity score is calculated by taking the log likelihood-ratio between the scores of the speaker model and the speaker-independent model. Four sets of data are randomly selected for training the speaker model and the remaining client data are used for testing. In the experiment, all the visual utterances not belonging to the subject are used as imposter data. The experiment is repeated for 20 times. The equal error rate (EER) is used to quantify the performance of different feature selections. The result presented in Table 1 is the average rate of these experiments.

It is observed from the experimental results that the authentication results follow a similar trend as the recognition results. The EER of using the shape-based lip feature alone is 6.25% and it is improved to 3.99% when the first mode of the intensity features is incorporated. Same as the recognition experiment, the improvement of the result is gradually decreased when adding the subsequent modes of intensity profile to the intensity feature vector. In fact, the EER cannot be further improved after adding the first 8 modes of the intensity profile and the best authentication result is 2.17%.

5. CONCLUSION

In this paper, we have presented the person authentication results using lip visual features derived from a 14-point ASM lip model. The experimental results have demonstrated that: (i) person authentication can be realized by solely using visual lip features; (ii) the use of shape-based lip features do not warrant acceptable performance; (iii) the EER can be substantially improved by incorporating the intensity information into the feature vector.

6. REFERENCES

[1] Xuejun Tan and Bir Bhanu, "On the fundamental performance for fingerprint matching", *Proc of IEEE CVPR, vol. 2 pp. 499-504, June 2003*

[2] J.L. Dugelay, J.C. Junqua, C. Kotropuolos, R. Kuhn, F. Perronnin and I. Pitas, "Recent advances in biometric person authentication", *Proc of IEEE ICASSP, vol. 4 pp. 4060-4063, May 2002*

[3] J. Luettin, Neil A. Thacker and Steve W. Beet, "Speechreading using shape and intensity information", *Proc of IEEE ICSLP, vol. 1, pp. 58-61, Oct 1996*

[4] T. Wark, D. Thambiratnam and S. Sridharan, "Person authentication using lip information", *Proc of IEEE TENCON, vol.1, pp.153-156, Dec1997*

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

[6] 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, May 2002

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

[8] 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, May 2001

[9] L.L.Mok, W.H.Lau, S.H.Leung, S.L.Wang and Hong Yan, "Lip features selection with application to person authentication", to appear in ICASSP'2004.

[10] 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

[11] T. Wark, Sridharan and Chandran, "An approach to statistical lip modeling for speaker identification via chromatic feature extraction", *Proc of IEEE Pattern Recognition, vol.1 pp. 123-125, Aug 1998*

[12] A. Kanak, E. Erzin, Y. Yemez and A.M. Tekalp, " Joint audio-video processing for biometric speaker identification", *Proc of IEEE ICASSP*, vol. 2, pp.377-380, April 2003