A New Lip Feature Representation Method for Video-based Bimodal Authentication

Hua Ouyang **Tan Lee**

Department of Electronic Engineering
The Chinese University of Hong Kong,
Shatin, N.T., Hong Kong,
Email: {houyang, tanlee}@ee.cuhk.edu.hk

Abstract

As the low-cost video transmission becomes popular, video based bimodal (audio and visual) authentication has great potential in various applications that require access control. It is especially useful for handheld terminals, which are often used under adverse environments, where the signal quality is rather poor. When human voice is used for authentication, one of the most relevant visual features is the dynamic movement of lips. In this research, we investigate on the use of static and dynamic features of speaking lips in the context of voice based authentication. A new feature representation that preserves both appearance and motion pattern of speaking lips is proposed. The dimension of extracted features is reduced by multiple discriminant analysis (MDA) and the method of nearest neighbor is used for classification. Our method can achieve an identification rate of 98% with only lips features for 200 clients of the XM2VTS database. Experiments on speaker verification using fused audio and visual features are on-going.

Keywords: multimodal, biometric authentication, lips, feature representation

1 Introduction

Automatic face recognition has been an active area of research for over 30 years. The major focus has been on the use of still images. Recently, video-based face recognition has attracted much attention (Zhao 2003). The intended applications are access control or restricted login through handheld terminals, on which low-cost video acquisition and transmission become increasingly popular. This research addresses the problem of person identification using short video clips of face images. In particular, we attempt to exploit the speaker-specific characteristics of talking mouth for bimodal (visual and audio) identification.

Video-based person identification technology enjoys some obvious advantages over still-image based methods. Firstly, the video signal is a sequence of correlated images. It is generally expected to contain more information, which stems from the spatial and temporal changes across frames. Secondly, video clips contain visual and audio information recorded in parallel, which provide complementary biometric features for person identification. In (Yacoub 1999) frontal face images and speech are fused using a linear weighted classifier as well as Support Vector Machine (SVM). In (Nefian 2003), a coupled hidden Markov model was used to combine lip, face and speech features for person identification.

While the spatial-temporal changes in a video sequence provide a lot more information than snap-shot images, there exists great redundancy due to that adjacent images are highly correlated. For person identification, we need to extract only the features that reflect the physical and/or behavioral traits of a person. In (B.Li 2001), a tracking-for-verification system was proposed. The study showed that the trajectory pattern of tracked points from the same face is more coherent than those from different faces. A similar work is found in (Zhou 2002), in which both the tracking motion vector and the identity variables are modeled to offers more degrees of freedom. In (Y.Li 2001), a 3-D Point Distribution Model and a shape-and-pose-free texture model were proposed to represent the depth information. The 3-D shape vector of the face is estimated from 2-D face images with a known pose. For recognition purpose, it proposed an Identity Surface to contain both pose information and textural features extracted by Kernel Discriminant Analysis(KDA). In (Mok 2004) and (Cetingul 2005), lip motion features during speech have been exploited based on extracted lip contours or motion vectors. In addition to the exploitation of temporal features from video signals, it is important to make the features invariant for the same person, i.e. to minimize the within-class variation. In (Mckenna 1998), a voting method was used for frame selection from video. Only those frames that are believed with high confidence to be in good quality are selected for recognition.

On handheld devices, the captured face image sequences are usually of low quality and low resolution. This makes tracking and feature localization more difficult and less accurate. Many discriminative details may be lost when the image size becomes small. In adverse environment, illumination and pose variations would be problematic. Face recognition is vulnerable to the variations of hair style, make-up, wearing glasses and facial expressions. Therefore, for person identification, we may rely on the lip area instead of the complete face. Lip, as one of the most important articulators, can be shaped differently to produce different sounds. Different styles of articulatory gesture are also reflected by lip movement. Since most salient movement of a talking face belongs to the lip area, we believe that robust and discriminative speaker-specific visual features can be extracted from this area. In this study, an effective representation of lip features in video sequence is developed for bimodal person identification.

2 Lip Feature Representation

2.1 Mouth Corner Detection

Before feature extraction and pattern classification, the ROI of lips should be localized first. There are many existing approaches of lip or face localization utilizing active contour models, intensity projection, pseudo hue or statistical models, e.g. active shape models. State-of-the-art techniques of lip localization can achieve good performance if there is no substantial illumination variation. In this research, we assume that the ROI for feature extraction can be reliably detected.

Without tackling the localization problem in depth, we adopt a semi-automatic method based on Lam's mouth corner detection (Lam 1996) to localize the ROI. The basic idea of this approach is that the most salient feature for locating eye or mouth is the corner.

The first step for corner detection is to detect the edges of an image. A gradient based Canny edge detector has been used. Every detected edge is regarded as a candidate for corner lines. A window is then

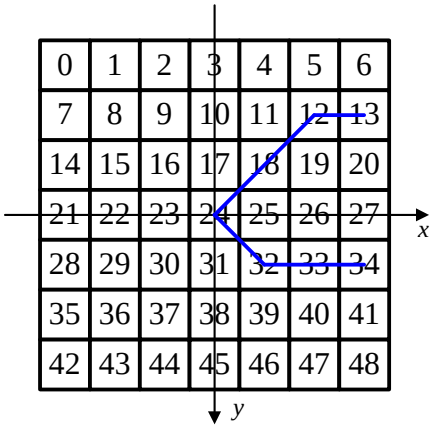

Figure 1: Detect lines of a corner. Each grid stands for a pixel in edge image

moved along the edge image to find candidate pairs of corner lines. As shown in Figure 1, a line is assumed to start from point 24, center of the window. There are a total of 48 lines in the window that can be defined, e.g. 24-32-33-34, 24-18-12-13. If more than two lines exist inside the window, the center of the window will be treated as a corner. Due to noise in images, there may be many corners being detected, thus a clustering is performed to group nearby points. Two points $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ will be grouped if

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} < TH \qquad (1)$$

where TH is a threshold. The point which represents all points in one cluster is chosen to be the one with maximum region dissimilarity $|D|$:

$$D(R_1, R_2) = \overline{I}_{R_1} - \overline{I}_{R_2} = \frac{1}{n_1}\sum_{R_1} I(x,y) - \frac{1}{n_2}\sum_{R_2} I(x,y) \qquad (2)$$

where $I(x, y)$ is the gray level intensity value of point $P(x, y)$; n_1 and n_2 are the number of pixels in region R_1 and R_2, which stands for the detected inner-corner region and outer-corner region respectively. $\overline{I}_{R_1}$ and $\overline{I}_{R_2}$ are the averaged intensity values. As shown in Figure 2, mouth corner areas are shadowed(R_2). Suppose that the true position for the left mouth corner C_0 locates in the center of window (a). If, for instance, the corner lines are detected to be l_{11} and

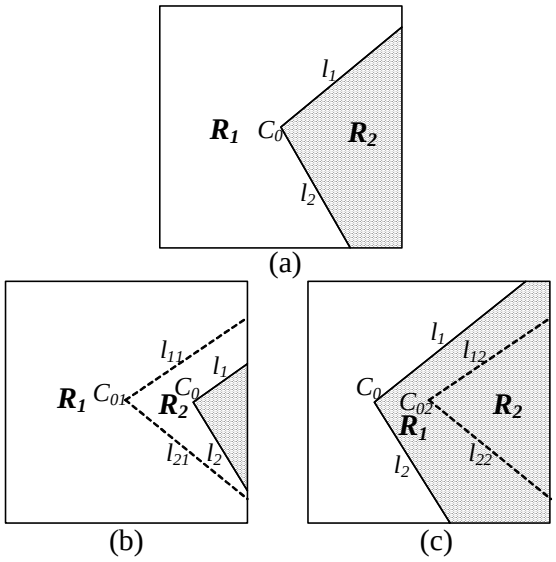

Figure 2: Locate the left mouth corner by window searching

l_{12}(Figure2(b)) because of noise, and mouth corner is detected to be C_{01}, then $\overline{I}_{R_1}$ remains unchanged, but $\overline{I}_{R_2}$ gets larger, thus $|D|$ will decrease. Similarly, if C_{02} is detected(Figure2(c)), then $\overline{I}_{R_1}$ gets smaller and $\overline{I}_{R_2}$ remains unchanged, thus $|D|$ will also decrease. Above analysis shows that by maximizing $|D|$, the most salient corner on the face can be properly locked at the center of the window.

In order to choose two clusters to represent the left and right mouth corner C_0 and C_1, a cost function (3) has been proposed.

$$(C_0, C_1) = \underset{corners}{\operatorname{argmin}} \left(K_1 \frac{|dy|}{|dy|_{max}} + K_2 \frac{|D|_{max}}{|D|}\right) \qquad (3)$$

where K_1 and K_2 are weightings, $|dy|_{max}$ is the maximum value of vertical distances of two candidate clusters, i.e. $|dy|_{max} = |y_{C_1} - y_{C_0}|_{max}$, and $|D|_{max}$ is the maximum value of $|D| = |D_{C_0} + D_{C_1}|$

When using the window searching scheme, a lip area position has to be coarsely estimated before fine searching of mouth corners. The estimation is based on some assumptions on human face geometries. However, in our experiments we found it to be not robust due to the variations of face geometry. In our applications, a video clip may consists of hundreds of frames, thus it is not practical to perform a good estimation on each frame. To avoid failures in coarse estimation stage, we proposed a quasi-automatic tracking method: label the two mouth corners manually for the first frame, then for the successive frames, search the neighboring regions base on corner positions of last frame to find new mouth corners.

Figure 3 shows some results of mouth corner detection. It should be pointed out that this method is not robust to faces with lips covered by bushy beard. But for these cases, the shape of beard can be a distinctive feature for recognition.

2.2 Lip Feature Extraction

Like most features used in biometric systems, lip features for the sake of speaker recognition can be categorized into physical and behavioral traits.

The information conveyed by a sequence of talking mouth can be partitioned into those that affect identity and those that describe the residual situations,

Figure 3: Some results of mouth corner detection

such as pose, ambient lighting, expressions and linguistic contents. When talking about physical traits, one of our motivations is to discard these variations and only keep the speaker dependent geometric features by exploring all the contents in a video clip. To extract speech-independent geometric features of lips, the non-rigid deformation of lip area will be regarded as "noise" and can be removed by a method of averaging over all the frames.

Except for the physiological information, behavioral characteristics is also conveyed by the image sequences of lips. A lot of research works have shown that lip movement can offer improved performance to speaker recognition systems when it is used together with acoustic-only or still-image-only features (Frischholz 2000, Yemez 2003). An intuitive understanding of features related to lip movement is that different people have different articulatory styles when speaking the same utterance. For instance, when uttering the same phoneme or word, some speakers tend to widely open the mouth while others may only keep slightly opened; some speakers tend to move the lips in a specific direction, and for some speakers, the teeth are always visible when uttering. To model these features resulting from movement, we propose to calculate the variances of pixel intensities in the lip image.

When no prior knowledge is known about a group of data, the frequently used method to represent the data set is to model these samples with a probabilistic distribution, e.g., normal densities. And the problem of parameter estimation can be solved by statistical methods, such as maximum likelihood estimation(ML) or Bayesian estimation(MAP). However, when the number of data sets is not adequate, or when the data dimensionality becomes problematic, the generalization ability of statistical methods will be greatly decreased. In practical applications, for example in our database, the frame rate is 25fps, and the length of a typical utterance is less than 10s, thus the total number of frames is less than 250, which is always not enough for statistical training. Based on these considerations, for each pixel in a specific position, we propose to regard the intensity values over all the frames as a random variable and find their means and variances (see Figure 4).

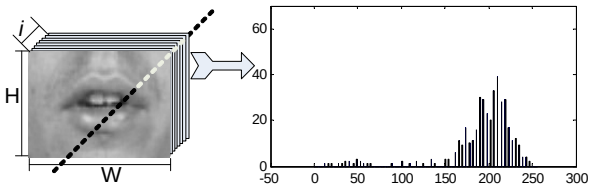

Figure 4: Take the intensity values of each pixel as a random variable.

The reconstruction process for "mean image" and "variance image" can be formulated by:

$$M_{i,j} = E\{\mathbf{x}_{i,j}\} = \frac{1}{k}\sum_{m=1}^{k} x_{i,j,m} \tag{4}$$

$$V_{i,j} = \sigma^2_{x_{i,j}} = \frac{1}{k}\sum_{m=1}^{k} (x_{i,j,m} - M_{i,j})^2 \tag{5}$$

where $M_{i,j}$ and $V_{i,j}$ are the intensity values of pixel (i,j) in the "mean image" and "variance image" respectively. $x_{i,j,m}$ is the intensity value of pixel (i,j) in frame m, and k is the total number of frames of a video clip.

Figure 5 shows some examples of the feature representation results. The four rows are taken at four different time sessions with one-month intervals. For each video clip, lip areas are supposed to have been properly localized by methods of section 2, and all the frames have been rotated and aligned. Based on this preprocessing, the intra-class variations due to lip movement, random movement of head and slight pose changes can be removed. As can be seen from column (a) and (c), as long as the utterance is the same, good consistence can be achieved. For column (b) and (d), the dynamic range and directions of lip movement can be roughly represented by those areas with high intensity values.

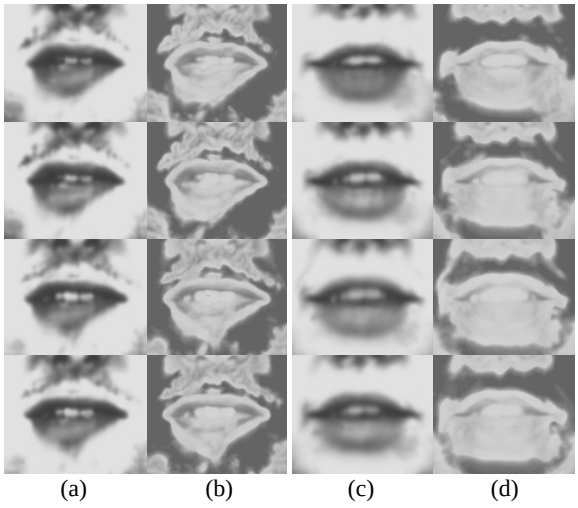

Figure 5: (a): mean images of speaker A; (b): variance images of speaker A; (c): mean images of speaker B; (d): variance images of speaker B;.

The represented lip features are essentially still images(although for variance "images", the intensity value may not follow the 8-bit or 16-bit range), thus a dimension reduction method should be imposed before classification. Principle Component Analysis(PCA) and Linear Discriminant Analysis(LDA) are two classic linear transformation methods which can project high-dimensional data to a low-dimensional space. PCA is often useful for representing data, while LDA is useful for discriminating between data in different classes. Since dissimilarities among lip geometries are not as significant as that among faces, holistic methods such as PCA may not be efficient. Thus we choose Multiple Discrimination Analysis(a multi-class generalization of LDA) to perform dimension reduction and classification.

The mean and variance images are processed by MDA separately, and the matching scores for the probe video and clients are calculated by a linear combination of Euclidian distances.

$$d = \beta d_{mean} + (1-\beta) d_{variance} \tag{6}$$

where β is a weighting factor within $[0, 1]$. d_{mean} and $d_{variance}$ are the normalized Euclidian distances for mean and variance images respectively.

3 Experiments

3.1 The XM2VTS Database

Our identification experiments are based on the XM2VTS database(Messer 1999). It is one of the largest audio-video databases designed for biometric research. Image sequences and speech data were synchronically recorded. Recordings consist of 4 sessions of 295 subjects pronouncing the English digits from zero to nine. The interval between two successive sessions is one month. Speeches were sampled at 32kHz. The video frame rate is 25fps, and all the video clips were captured in front of a blue background. Original resolution of these 24bit color images is 720*576 pixels.

As mentioned in Section 2, the mouth corner detection is not robust to those lips covered by bushy beard, thus we only chose 200 out of 295 subjects whose mouth corners are properly located for our experiments.

The high-resolution image sequences extracted from video clips are firstly downsampled to 360*288 pixels, and then converted into 8-bit gray-level images. After that, mouth corner detection was imposed to all the sequences. Based on the extracted corner locations, images are rotated to make sure that the lines connecting the two corners are horizontal. A lip ROI of 60*50 pixels was then cropped from each frame based on the aligned images. Geometric and behavior features of lips are then represented by the proposed method. Euclidean distances calculated from these two features are normalized and linearly combined with a weight β, and its value is empirically choose to be 0.9.

Four groups of experiments had been carried out. Features used in group 1 and 2 are geometric and behavioral features respectively, while linearly combined features are utilized in group 3. In group 4, only the first frame of each lip image sequence was used for identification.

3.2 Results and Discussion

Session		Identification Rate			
Train	Test	Geometric	Behavioral	Combined	1st Frame
s_1, s_2	s_4	97.00%	89.50%	98.00%	84.00%
s_1, s_2, s_3	s_4	99.25%	94.00%	99.50%	93.75%
s_1, s_2	s_3	96.75%	89.75%	97.25%	78.25%
s_1, s_2, s_4	s_3	98.75%	94.75%	99.25%	91.25%
s_1, s_3	s_2	96.25%	88.25%	96.75%	83.50%
s_1, s_3, s_4	s_2	98.50%	93.25%	98.75%	92.00%
s_2, s_3	s_1	96.25%	88.75%	97.00%	80.75%
s_2, s_3, s_4	s_1	98.50%	94.00%	98.75%	88.25%

The identification result is shown in the above table. It is clear that, the performance of geometric only features is better than behavior only features. The performance gain of combined features over geometric-only features is 0.25%-1% absolute. This may indicates that the proposed behavior features can provide some complementary information, however, they are still highly correlated with geometric features.

On the other hand, when the proposed features are compared with features based on still lip images, the improvement of performance is much greater. According to recent research works (Tang 2004), the identification rate of video-based face recognition using the same database is 98.6%(294 subjects were used).

4 Conclusion

A new lip feature representation method has been proposed for an audio-visual biometric authentication system which focuses on the lip area. Experimental results show that static and dynamic lip features have better performance than still lip images, and its performance is even comparable to face based methods. Behavior lip features can be a good complement to geometric lip features. The proposed lip features can be used in conjunction with speech data to improve the overall performance of multimodal system. Experiments on video-audio fusion is still on-going.

References

Wenyi Zhao, Rama Chellappa, P.J.Phillips, (2003), Face Recognition: A Literature Survey, ACM Computing Surveys **35**(4), 399–458.

Baoxin Li, Rama Chellappa (2001), Face verification through tracking facial features, Journal of the Optical Society of America **18**(12), 2969–2981.

Yongmin Li, Shaogang Gong, Heather Liddell (2004), Constructing Facial Identity Surfaces in a Nonlinear Discriminating Space, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR01).

Shaohua Zhou, Volker Kruger, Rama Chellappa (2003), Probabilistic Recognition of Human Faces from Video, Computer Vision and Image Understanding **91**, 214–245.

L.L. Mok, W.H. Lau, S.H. Leung, S.L. Wang, H. Yan (2004), Lip Features Selection with Application to Person Authentication , IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP04).

H. Ertan Cetingul, Yucel Yemez, Engin Erzin,A. Murat Tekalp (2004), Robust Lip-Motion Features for Speaker Identification , IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP05).

Stephen J. Mckenna, Shaogang Gong (1998), Recognising Moving Faces , Face Recognition: From Theory to Applications, Springer, 578–587.

S.Yacoub, J.Luettin, J.Kittler (1999), Audio-Visual Person Verification, Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR99).

A.V.Nefian, L.H.Liang, T.Fu, X.X.Liu (2003), A Bayesian Approach to Audio-Visual Speaker Identification, Proc. 4th Int. Conf. on Audio and Video-based Biometric Person Authentication (AVBPA).

K.M.Lam, H.Yan, (1996), Locating and Extracting the Eye in Human Face Images, Pattern Recognition **29**(5), 771–779.

Robert W. Frischholz, Ulrich Dieckmann (2000), BioID: A Multimodal Biometric Identification System, Computer **33**(2), 64–68.

Y.Yemez, A.Kanak, E.Erzin, A.M.Tekalp (2003), Multimodal speaker identification with audio-video processing, Proceedings of the 2003 IEEE Computer Society Conference on Image Processing (ICIP03).

K.Messer, J.Matas, J.Kittler, J.Luettin, and G.Maitre (1999), A Bayesian Approach to Audio-Visual Speaker Identification, Proc. 2nd Int. Conf. on Audio and Video-based Biometric Person Authentication (AVBPA).

Tang, Xiaoou, Li, Zhifeng(2004), Video Based Face Recognition Using Multiple Classifiers, Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR04).