REAL TIME LIP MOTION ANALYSIS FOR A PERSON AUTHENTICATION SYSTEM USING NEAR INFRARED ILLUMINATION

Faisal Shafait, Ralph Kricke, Islam Shdaifat, Rolf-Rainer Grigat

Department of Vision Systems
Hamburg University of Technology, Germany
{faisal.shafait, kricke, shdaifat, grigat}@tu-harburg.de

ABSTRACT

In this paper we present an approach for lip motion analysis that can be used in conjunction with a person authentication system based on face recognition, to avoid attacks on the system using passive photographs. This work focuses on robustly tracking lips in gray scale images, which may be captured in the visible light or near infrared spectrum. We present an approach for locating the two lip corners in a face image. Then we extract suitable features from the mouth region to classify mouth states (visemes). The system shows a classification accuracy of above 85%. The temporal changes in the detected viseme classes can be used for detecting the imposter.

Index Terms— Infrared imaging, Discrete cosine transforms, Image analysis

1. INTRODUCTION

The science of biometrics provides means to identify persons by truly verifying non-transferable features of an individual. Face recognition is a particularly compelling biometric feature [1]. However, access control systems based on face recognition are easily decieved by a photograph placed in front of the camera. To overcome this issue we classify mouth states (visemes) in an image sequence and determine the liveliness of a person from the temporal changes of the detected viseme classes (see Fig. 1). This differs from other approaches, where the identification task is based solely on the lip information [2, 3] or on combined facial and speech features [4].

Robustly tracking lip motion in image sequences is especially difficult because lips are highly deformable, and they vary in shape and color across individuals. In addition, they are subject to both non-rigid (expression, utterance of speech) and rigid motion (head movement). [5, 6] show that color space transformation works very well for lip image segmentation. However, restricting the system to gray scale images eliminates the possibility to use color segmentation for lip tracking. Hence we focus on detecting the two lip corners instead. We describe this step in section 2. Then we extract features which we use for classification of different mouth shapes (section 3). The experiments and the obtained results are presented in section 4 followed by the conclusion in section 5.

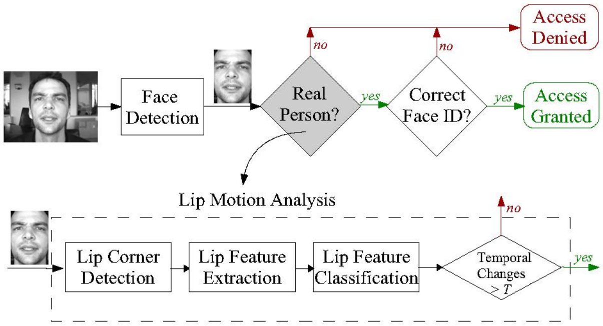

Fig. 1. Lip motion analysis for person authentication

2. LIP CORNER DETECTION

Given the large amount of research already carried out in face detection, we assume that the positions of both eyes in the image have been located. The lip corners in an image $\mathbf{I}$ can then be detected as described below.

2.1. Construction of the representative prototype

For modeling the corners of the lips, closed and open mouth are considered as two different classes represented by two different appearance models (templates) denoted by $\mathbf{T}_c$ and $\mathbf{T}_o$. These two classes are represented by their average templates $\mathbf{T}_{o,\text{avg}}$ and $\mathbf{T}_{c,\text{avg}}$ that are gathered from the training data of all individuals. The position of the two lip corners is manually extracted from the images used for training. Then a small region centered on a lip corner is used as a template for that lip corner.

This work is part of the project *KATO – Innovative Cabin Technologies* funded by the German Federal Ministry of Economics and Technology.

1-4244-0481-9/06/$20.00 ©2006 IEEE

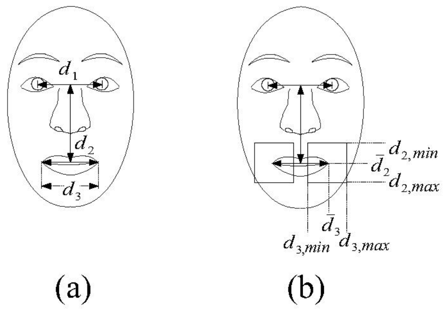

Fig. 2. Region-of-interest (ROI) estimation for locating the two lip corners.

2.2. Size, orientation, and region-of-interest (ROI) estimation

The position of the eyes is used to calculate the ROI to search for the lip corners by providing an estimate of the scale and orientation of lips in the image.

Consider the distances between different facial features (Fig. 2). The distances $d_i, i = 2, 3$ are modeled individually as univariate Gaussian distributions $d_i \sim \mathcal{N}(\mu_{d_i}, \sigma_{d_i})$. The boundaries of the region of interest are selected as the limits of the centered confidence interval $(1 - \gamma)$, where $0 \leq \gamma \leq 1$, such that $(1 - \gamma)$ is the probability that d_i lies within the interval $[d_{i,min}, d_{i,max}]$ as defined in Eq. 1.

$$P(d_{i,min} \leq d_i \leq d_{i,max}) = 1 - \gamma \quad i = 2, 3 \tag{1}$$

2.3. Similarity measurement between image and model

In order to locate the two lip corners in an image $\mathbf{I}$, we find the best matching position of the lip corner template in the ROI under a given similarity measure. The similarity measure used in this work is

$$s(\mathbf{I}, \mathbf{T}) = \alpha\rho_{(\mathbf{IT})} + (1 - \alpha)\rho_{(|\Delta\mathbf{I}||\Delta\mathbf{T}|)}, \tag{2}$$

where α is a scalar value, $\rho_{(\mathbf{IT})}$ is the correlation coefficient between the image $\mathbf{I}$ and the template $\mathbf{T}$ defined as:

$$\rho_{(\mathbf{IT})} = \frac{\sum_{i=0}^{N-1}(\mathbf{I}_i - \mu_\mathbf{I})(\mathbf{T}_i - \mu_\mathbf{T})}{N\sigma_\mathbf{I}\sigma_\mathbf{T}} \tag{3}$$

and $\rho_{(|\Delta\mathbf{I}||\Delta\mathbf{T}|)}$ is the correlation coefficient between the absolute gradients $|\Delta\mathbf{I}|$ and $|\Delta\mathbf{T}|$.

Since two appearance classes $(\mathbf{T}_i, i \in \{c, o\})$ are used to model each lip corner, the similarity $s_i = s(\mathbf{I}, \mathbf{T}_i)$ is calculated for each appearance class. The similarity s for both classes can be calculated as the similarity between $\mathbf{I}$ and the mean value of all templates i. e.

$$s_a = s(\mathbf{I}, \frac{1}{2}\sum_{i\in\{c,o\}} \mathbf{T}_i), \tag{4}$$

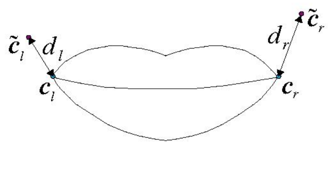

Fig. 3. The actual and estimated positions of lip corners used as error metric

Alternatively, similarity measure can be calculated individually for each template. Then the mean result is calculated according to Eq. 5.

$$s_c = \frac{1}{2}\sum_{i\in\{c,o\}} s(\mathbf{I}, \mathbf{T}_i) \tag{5}$$

The similarity s_c results in a better localization accuracy than s_a since it preserves the edge information. The computational complexity of s_c is about double than that of s_a. However, the computational disadvantage is compensated by restricting the search to the points with high gradient in the ROI. Hence the position in the image $\mathbf{I}$ giving the highest s_c is reported as the estimated position of the lip corner.

2.4. Error Metric

To evaluate the performance of the lip corner detection system, a relative error measure based on the distances between the expected and the estimated lip corner positions is used.

Let $\mathbf{c}_l, \mathbf{c}_r$ be the actual coordinates of the left and right lip corners in the image, and $\tilde{\mathbf{c}}_l, \tilde{\mathbf{c}}_r$ be the estimated coordinates of the corresponding lip corners (Fig. 3). The maximum of the distances $d_l = \|\mathbf{c}_l - \tilde{\mathbf{c}}_l\|$ and $d_r = \|\mathbf{c}_r - \tilde{\mathbf{c}}_r\|$, normalized by dividing it by the mouth width, to make it independent of scale of the face in the image and image size, is used as an error metric as depicted in Eq. 6.

$$E = \frac{\max(d_l, d_r)}{\|\mathbf{c}_l - \mathbf{c}_r\|} \tag{6}$$

In general, a feature is considered to be correctly located if $E \leq 0.1$ [7].

3. LIP FEATURE EXTRACTION AND CLASSIFICATION

After locating the two lip corners in the face image, suitable features are extracted from the mouth region in order to classify the image into the target viseme classes. Existing approaches fall into three categories: features based on image transform, shape model, and joint shape-appearance (for a literature survey, please refer to [8]). In a comparative study [9], image transform based features have been shown to outperform lip-contour and shape-model based features.

Therefore, we compare an image transform based feature [10] using 2D-DCT to a joint shape-appearance based feature using the Bézier curve model of the lips [11] for the purpose of viseme classification.

3.1. Image transform based features

The lip corner points are used to estimate the mouth center in each image. The image is scaled to a fixed size by using the distance between the eyes as reference. Then a ROI of size 64×64, centered around the mouth center is obtained. The 2D-DCT of the ROI is taken, the resulting DCT coefficients are zig-zag scanned, and the first 64 elements are used as feature vector.

3.2. Intensity profile features based on quadratic Bézier curve lip model

The two endpoints $\mathbf{p}_0$ and $\mathbf{p}_2$ of the quadratic Bézier curves were fixed to the two lip corner points. The control point $\mathbf{p}_1$ was shifted in 64 equal steps along the distances $-0.7d_2$ to $0.7d_2$ in the vertical direction, where d_2 is the mouth width (see Fig. 2). For each shift step, the intensity values were integrated along the path of the quadratic Bézier curve starting at a distance of $0.2d_2$ away from each corner. The intensity profile, obtained in this way, serves as a feature vector. By using a one-dimensional DCT, the length of this feature vector is further reduced.

3.3. Feature Classification

The extracted lip features are classified into two target viseme classes: open mouth ω_o and closed mouth ω_c. This is sufficient to detect presence of lip motion in a scene. Two types of classifier are used, namely K-nearest-neighbor (KNN) classifier and Bayesian classifier using Gaussian mixture model (GMM) as density estimator. The model parameters of the GMM are estimated from expectation maximization (EM) algorithm.

4. RESULTS AND DISCUSSION

The BioID face database [7], consisting of 1521 gray level images of 25 persons with 20 manually placed feature points on each image has been used for evaluation purpose. We manually classified each image to belong to either of the two classes. The entire dataset is partitioned equally to three disjoint sets, 3-fold cross validation is used by taking each of these partitions turn by turn as the training set and the other two as the test set.

We selected the intervals $d_2 = [0.85d_1, 1.35d_1]$, and $d_3 = [0, 1.5d_1]$ to determine the ROI for the lip corners (section 2.2), since for the BioID database all lip corners were always within this ROI ($\gamma = 0$).

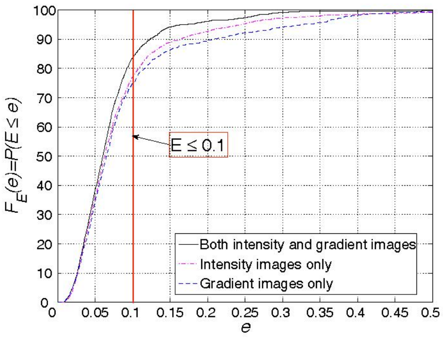

Fig. 4. Effect of using both intensity and gradient images as a similarity measure (Eq. 2) on localization error (Eq. 6). The weighting factor α in Eq. 2 was set to $0.0, 0.5$, and 1.0.

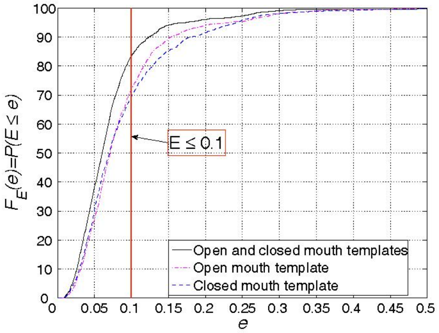

Fig. 5. Effect of using multiple templates for lip corner detection on localization error (Eq. 6)

First, the effect of using the intensity and the gradient images in calculating correlation (Eq. 2) is studied. The gradient correlation localizes features well, but is sensitive to changes in size and rotation. The intensity correlation is more robust against these variations, but is more sensitive to illumination changes. Let e be the value the error E can assume, and $F_E(e)$ be the cumulative distribution function of E. Using a maximum allowed localization error of $E \leq 0.1$ for a correctly located lip corner, results in Fig. 4 show a localization accuracy of 76% using gradient image only, and 78% using intensity image only. The combination of the two increases the detection accuracy to 84%.

The localization error produced by using open and closed mouth templates respectively, and then using both open and closed mouth templates simultaneously, is plotted in Fig. 5. The use of multiple templates increases the localization accuracy from about 71% to 84%. Hence there is a substantial gain of 13% when using two templates instead of one, making the additional computational cost well justified.

In order to detect whether a person is speaking or not, it is

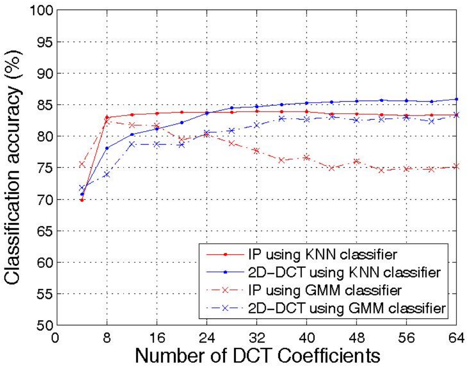

Fig. 6. Classification accuracy using intensity profile (IP) and 2D-DCT as a feature vector. Both KNN and GMM classifier are used to classify each feature vectors.

sufficient to classify the mouth images into two main viseme classes, namely open mouth class (ω_o), and closed mouth class (ω_c). The classifier is trained on the training set and then the classification accuracy is reported as the percentage of correctly classified images in the test set. The comparison results are shown in Fig. 6.

The results show that, when the feature vector contains more than 24 coefficients, the image transform based features using 2D-DCT result in better classification accuracy as compared to the joint-shape-appearance based features using integrated intensity profile (IP), irrespective of the classifier used.

5. CONCLUSION

This paper presented an approach for lip motion analysis to be used in conjunction with an authentication system based on face recognition. Multiple speaker dependent templates are used for locating the two lip corners. The localization accuracy is about 84% for the BioID face database. Then, features are extracted from the mouth area. The results show that the 2D-DCT based features have better overall performance than that of intensity profile (IP) along the quadratic Bézier curve model of the lips, when the length of the feature vector greater than 24 elements.

The extracted lip features are classified using the GMM and the KNN classifiers. When the 2D-DCT based feature vector is used, the classification accuracy achieved by the KNN classifier is about 86% on the BioID face database. Under the same conditions, the GMM classifier attains an accuracy of 83%, which is slightly lower than that of the KNN classifier. However, the computational cost of the KNN classifier during the classification stage is higher than that of the GMM classifier, especially when the training set is large. For the real-time operation, a compromise has to be made between the better accuracy and higher computational cost of the KNN classifier by limiting the size of the training set.

6. REFERENCES

[1] N. K. Ratha, A. Senior, and R. M. Bolle, “Automated Biometrics,” *Proceedings of ICAPR*, 2001.

[2] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, “Discriminative lip-motion features for biometric speaker identification,” in *ICIP*, 2004, pp. 2023–2026.

[3] L. L. Mok, W. H. Lau, S. H. Leung, S. L. Wang, and H. Yan, “Person authentication using ASM based lip shape and intensity information,” in *ICIP*, 2004, pp. 561–564.

[4] Jean-Luc Dugelay, Jean-Claude Junqua, C. Kotropoulos, Roland Kuhn, Florent Perronnin, and I. Pitas, “Recent advances in biometric person authentication,” in *ICASSP*, Orlando, USA, May 2002.

[5] S. Lucey, S. Sridharan, and V. Chandran, “Improved facial feature detection for AVSP via unsupervised clustering and discriminant analysis,” *EURASIP Journal on Applied Signal Processing*, pp. 264–275, 2003.

[6] S. Leung, S. Wand, and W. Lau, “Lip image segmentation using fuzzy clustering incorporating an elliptic shape function,” *IEEE transactions on image processing*, vol. 13, pp. 51–61, 2004.

[7] Oliver Jesorsky, Klaus J. Kirchberg, and Robert W. Frischholz, “Robust Face Detection Using the Hausdorff Distance,” in *Audio- and Video-Based Person Authentication*, Josef Bigun and Fabrizio Smeraldi, Eds. 2001, vol. 2091, pp. 90–95, Springer Verlag.

[8] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, *Audio-Visual Automatic Speech Recognition: An Overview. In: Issues in Visual and Audio-Visual Speech Processing*, chapter 10, MIT Press (In Press), 2004.

[9] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, “A comparison of model and transform-based visual features for audio-visual LVCSR,” *Proc. Int. Conf. Multimedia Expo., Tokyo, Japan.*, 2001.

[10] G. Potamianos, H. P. Graf, and E. Cosatto, “An image transform approach for HMM based automatic lipreading,” in *IEEE International Conference on Image Processing*, Chicago, 1998, vol. 1, pp. 173–177.

[11] Islam Shdaifat, *Design of a visual front end for audiovisual speech recognition*, Books on Demand GmbH, Norderstedt, 2005.