Use of Vertical Face Profiles for Text Dependent Audio-Visual Biometric Person Authentication

Vinod Pathangay and B. Yegnanarayana
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai- 600 036, India
{vinod, yegna}@cs.iitm.ernet.in

Abstract

In this paper, a technique is proposed for text dependent audio-visual biometric person authentication using Dynamic Time Warping (DTW). A combination of features derived from video and audio is used as a representation of the utterance. The use of mid-face vertical intensity profiles as a representation of facial movements associated with the utterance is proposed. The time varying profiles are extracted from the video sequence after detecting the face using a motion guided template matching technique. The visual feature is combined with Linear Prediction Cepstral Coefficients (LPCC) extracted from the speech waveform to obtain a temporally synchronous joint audio-visual feature. The joint audio-visual feature sequences are matched using the DTW algorithm to obtain the distances between the test and the reference utterances. The performance of the system is evaluated for a database of 25 speakers, and the results are discussed.

1. Introduction

Biometric person authentication is the process of verifying the identity of a person based on his/her physiological or behavioral characteristics. As human speech contains the characteristics of the speaker, it is currently exploited in speaker recognition systems. Human speech is also bimodal, as it contains visual information in terms of facial movements along with the audio. In the audio-visual person authentication, the objective is to use the information present in the audio and visual components of speech production to authenticate a person. Techniques such as face recognition and speaker verification can be compromised by the use of face masks and pre-recorded utterances of a genuine person. A bimodal person authentication system using audio-visual information can offset some of the problems inherent in systems based on a single modality. This is supported by the fact that humans perceive speech in both audio and visual modes. Thus the use of bi-modal information can increase the robustness of the person authentication system.

In this study, the audio-visual person authentication problem is mapped as a bimodal pattern recognition task, where both modalities of audio and video are jointly processed at the feature level before making a decision. The issues involved are:

1. Face detection and tracking in video.
2. Feature extraction from video and audio.
3. Combination of modalities at feature level.
4. Pattern matching of test with reference features.

Extraction of facial features requires detection of the face in the image frame. Face detection using templates and motion information is discussed in Section 2. Section 3 discusses extraction of visual features. The Linear Prediction Cepstral Coefficients (LPCC) extracted from speech are combined with the visual features. Section 4 explains the methods of integration of audio and video modalities at the feature level. Pattern matching using DTW and performance evaluation of the proposed technique and discussed in Section 5. Section 6 concludes the paper by discussing the results of our experiments and directions for future work.

2. Face Detection and Tracking

In the audio-visual person authentication, the first step is face detection. After detection in the first image frame of the video, the face has to be tracked in the subsequent frames, to compensate for any displacement during the utterance of the sentence. In this study, the following assumptions are made about the face in the video:

1. There is a single upright frontal face.
2. The lighting conditions during enrollment and testing remain nearly the same.

Figure 1: Average face template

Figure 2: Different scales of upper face templates used for face detection

3. Size of the face varies between 75x75 to 125x125 pixels in the image frame.

4. The person does not wear spectacles.

With these assumptions, the face is detected by a template matching technique where the search space is pruned by using motion information.

2.1. Face Detection using Motion Guided Template Matching

In order to detect a face in a still image, a number of techniques have been reported [1]. Feedforward neural networks for classifying face and non-face inputs [2] and models based on skin colour have been reported [3, 4]. In our technique, the face is detected in a video sequence using multi-scale average face template. The average face template is shown in Figure 1. This was generated by averaging a set of 80 faces that were manually extracted from the FERET face dataset [5]. The upper part of the face template is used instead of full face templates reported in [6]. This is to allow variation in the lower face during speech and to make the matching less sensitive to cases where there is facial hair present.

The average face template is matched with overlapping windows taken from the image frame. Multiple scales of the upper face template as in Figure 2 are matched with overlapping windows taken from the image. The size of the window taken is same as that of the template. The absolute correlation coefficient is used as the extent of match between the face template T and the window W. The correlation coefficient ρ is given as

Figure 3: A sample image frame from a video sequence

$$\rho_{ij} = \frac{\sum_{k=0}^{K-1}\sum_{l=0}^{L-1}(W_{i+k,j+l} - \overline{W_{ij}})(T_{kl} - \overline{T})}{\sqrt{\sum_{k=0}^{K-1}\sum_{l=0}^{L-1}(W_{i+k,j+l} - \overline{W_{ij}})^2 \sum_{k=0}^{K-1}\sum_{l=0}^{L-1}(T_{kl} - \overline{T})^2}} \quad (1)$$

where (i, j) are the (row, column) indices of the image, (k, l) are the (row, column) indices of the window W and template T. $\overline{T}$ and $\overline{W}$ are the mean values of the pixels in the template and window, respectively. Different scales of the template as shown in Figure 2 are used for matching. The scale of the template that returns the highest value of ρ is taken as the best match scale. The location of the match is taken as those values of i, j where there is a maximum value of ρ.

In order to reduce the search space for template matching, motion information is used. This is under the assumption that the face of a speaker in a video may undergo displacement during an utterance, which is evident from Figure 4 where the difference image between two successive image frames is shown.

If the video $\mathbf{V}$ is represents a sequence of image frames $\mathbf{F_l}$ for an utterance of L frames,

$$\mathbf{V} = \{\mathbf{F_0}, \mathbf{F_1}, ..., \mathbf{F_l}, ..., \mathbf{F_{L-1}}\} \quad (2)$$

then the difference image $\mathbf{D_l}$ for the frame l is given as

$$\mathbf{D_l} = |\mathbf{F_l} - \mathbf{F_{l+1}}| \quad (3)$$

and the accumulated difference image $\mathbf{A}$ for the sequence is obtained as

$$\mathbf{A} = \frac{1}{L-1}\sum_{l=0}^{L-2}\mathbf{D_l} \quad (4)$$

Figure 5 shows the binarized motion map obtained by thresholding the accumulated difference image $\mathbf{A}$. It can be seen from Figures 4 and 5 that, in order to locate a face in a video sequence of an utterance, the template matching needs to be performed only in the regions where there is change. Therefore in each window the number of pixels that have undergone change is taken from the motion

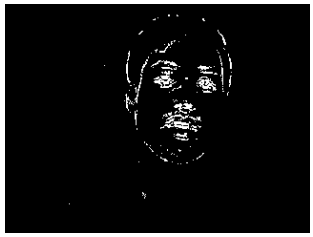

Figure 4: Difference image between two adjacent frame of the video sequence of an utterance

Figure 5: Thresholded Accumulated Differences for a video sequence of an utterance used as a binary motion map

Figure 6: Result of face detection in a cluttered background using upper face template (face enclosed in the white rectangle)

Figure 7: Result of face detection on person with facial hair

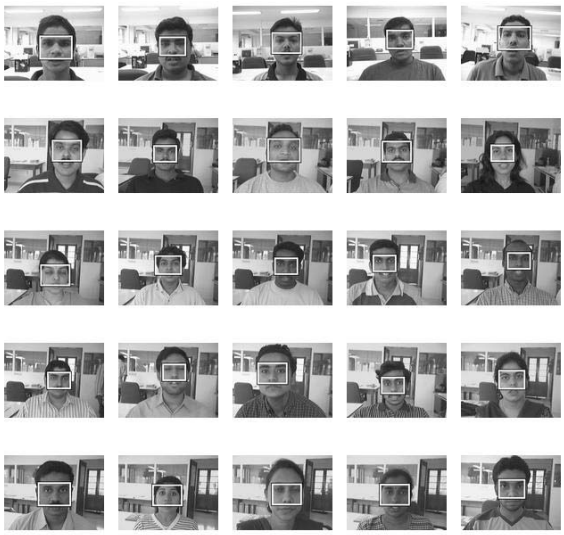

Figure 8: Results of the face detection technique applied on the database used

map. This pixel change count is used to decide if the ρ_{ij} has to be calculated or not. If the pixel change count is below a threshold, then ρ_{ij} is set to zero, else it is calculated as in (1). Figure 6 shows the result of this face detection technique for a test image sequence with a cluttered background as in Figure 3. The white box shows the location of the detected face. The technique also works with persons with facial hair as shown in Figure 7. Figure 8 shows the results of the face detection technique applied on the 25 speaker data set used in this study.

2.2. Face Tracking

During utterance, it is observed that the position of the face does not remain constant. Therefore the face has to be tracked across the entire video sequence for accurate feature extraction. The face is tracked by performing template matching over a limited region around the original location of the face in the first image frame of the video. The face detected in the first image frame is used as a template. If (p, q) represent the (row, column) of the location of the face in the first image frame $\mathbf{F_0}$ of the video, the matching region in the second frame onward is taken as a sub-image $\mathbf{F_1}(i, j)$ where

$$p - S \leq i \leq p + T_h + S \tag{5}$$

$$q - S \leq j \leq p + T_w + S \tag{6}$$

where (i, j) are the (row, column) indices of the image frame, S is the offset around the original face location and T_h, T_w are the height and width of the face respectively. The

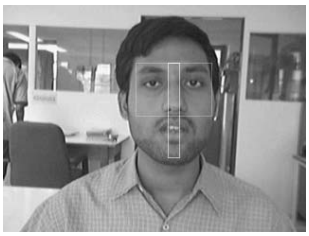

Figure 9: Extraction of vertical intensity profile from face as shown by the vertical box along the middle of the face. The profile is averaged at each row of the box.

offset S is set to a value taking into consideration the extent of head movement during an utterance. Thus the face is localized across all the image frames of the video sequence of the utterance.

3. Visual Feature Extraction

In order to represent the variations on the face that occur during a speech utterance, visual features are extracted from the facial region. The features for representing visual speech can be classified as morphological features and photometric features. Morphological features are parameters related to the shape of the lips such as lip height, lip width and contour shape parameters extracted using active contour models [7]. Photometric features are extracted from the gray level value of the image, such as eigenlips [8]. Active appearance models use statistical and morphological properties of the lip sub-image [9]. Optical flow vectors estimated between images frames have also been used as visual features [10]. But the flow vectors are computationally expensive, and are not accurate when used to represent the motion of non-rigid objects such as face.

In this paper, vertical face profile is proposed as a feature for representing visual speech for the audio-visual person authentication task. This feature has the advantage of being simple and easier to extract, than the other reported visual features. Figure 9 shows the extraction of mid-face vertical intensity profile. The face profile vector is taken as the row-averaged pixel values along the vertical mid-face region. The pixels are row-averaged in order to eliminate noisy spikes in the profile. The use of vertical face profiles is motivated by the fact that there is maximum change on the face only in the vertical direction during speech. This is due to the displacement of the lower jaws during a speech utterance. Figure 10 shows a 3-D plot of the face profiles extracted from a sequence of frames of an utterance. It can be seen that the dynamic variations on the face is captured by the face profile sequence. The variation is between row indices 50 and 90 in Figure 10. The temporal variation of this region is caused by the lip movements.

The profile extracted has static and dynamic parts. The

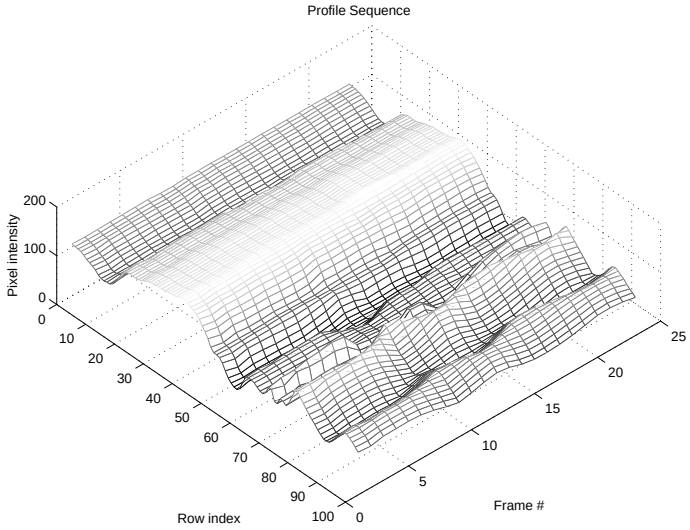

Figure 10: Profiles extracted over 25 frames of an utterance

static part corresponds to the upper region of the face where there is no variation during an utterance. The dynamic part corresponds to the lip region that registers some variation during an utterance. The temporal variation of each pixel of the face profile is shown in the variance plot shown in Figure 11. The correct length of the profile is the sum of the lengths of static and dynamic parts. When a sufficiently long fixed length profile is extracted from the face, the size is adjusted by using the size of static and dynamic parts. The end of the dynamic part can be obtained by thresholding the temporal variance. In Figure 11, the end of the dynamic part corresponds to the pixel index 80, therefore the corrected length of the profile is in this case is 80 pixels.

The profile vector $\mathbf{p}$ extracted from the face in an image frame is represented as

$$\mathbf{p} = \{p_0, p_1, ..., p_n, ..., p_{N-1}\} \tag{7}$$

where p_n is the pixel value and N is the normalized length of the profile. The first K Fourier coefficients of $\mathbf{p}$ is represented as

$$\mathbf{p^{FC}} = \{f_0, f_1, ..., f_{K-1}\} \tag{8}$$

where f_k is the k^{th} Fourier Coefficient. In this study, the value of K was taken as 4 for $N = 50$. This is to eliminate the high frequency spikes in the profile. The Profile Fourier Coefficients (PFC) is taken as the feature vector derived from $\mathbf{p^{FC}}$ by taking the logarithm of each Fourier coefficient.

The begin and end image frames of the utterance are detected using the begin-end detection algorithm for speech proposed in [11]. This eliminates the leading and trailing silence regions in the utterance. Thus the visual speech is represented parametrically by a sequence of PFCs.

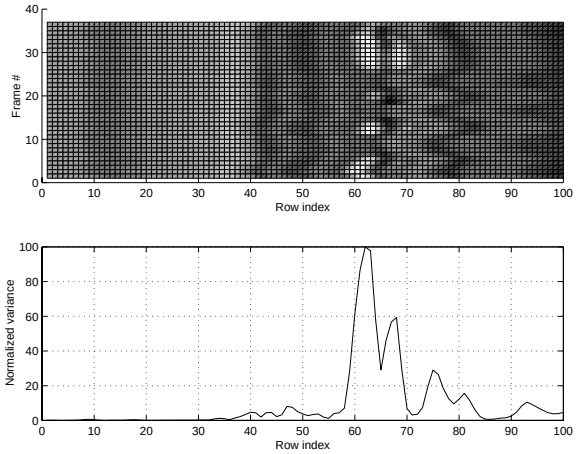

Figure 11: Profiles and their temporal variance for detecting static and dynamic regions

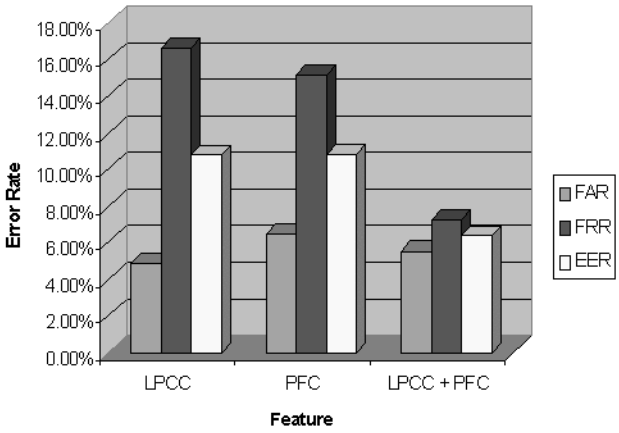

Figure 12: Performance of LPCC, PFC and their feature level combination in terms of FAR, FRR and EER

4. Audio-visual Integration

The audio stream can be integrated with video at two different levels, which are known as late fusion and early fusion [12]. In the late fusion, separate classifiers are used for audio and visual streams, and their decisions are combined. In the early fusion, the audio-visual modalities are combined at the feature level. In this paper, the early fusion strategy is explored, where the audio and video features are combined at the feature level. The motivation for using the early fusion strategy is that it captures the temporal synchronization of the audio and visual events.

The acoustic speech signal is parametrically represented as 19 dimension Linear Prediction Cepstral Coefficients (LPCC). A detailed description of LPCC extraction is given in [13]. The LPCCs are extracted for a frame size of 20ms and frame shift of 5ms. This gives a feature generation rate of 200 features per second.

The PFC extracted from each frame of the video is at a rate of 30 features per second. As the rate at which video features are generated is much lower than that of audio feature, the video features were repeated. Thus there are equal number of LPCCs and PFCs per second. A joint audio-visual feature is generated by concatenating the LPCC and the PFC vectors. Thus the temporal synchronization is maintained in the joint audio-visual representation.

5. Pattern Matching using DTW

During enrollment, the audio and video features are extracted and stored as reference templates. During verification, the test features are matched with the reference templates. In order to match the test with the reference patterns, the Dynamic Time Warping (DTW) algorithm is used [14]. Let $Y_0, Y_1, Y_2, ..., Y_j, ..., Y_{J-1}$ represent the reference template for a speaker, and $X_0, X_1, X_2, ..., X_i..., X_{I-1}$ denote the test utterance. In general $I \neq J$, due to differences in speaking rate. DTW gives a warping path that associates features in X to those in Y. Thus matching score is computed as

$$D = \sum d(X_i, Y_{j(i)}) \tag{9}$$

where the template indices are determined using the DTW algorithm and $d(.)$ represents the distance between the feature vectors of the template sequence Y, which matches with the i^{th} vector of the input sequence X. A more detailed account on the use of DTW algorithm for text-dependent speaker verification is given in [11]. During verification, the test feature sequence is compared with all the reference templates, and the matches are arranged in order of increasing distances. A correct match is assumed when the test speaker ID is ranked within the three closest distances.

5.1. Experimental Results

The database used to evaluate the proposed technique consists of 25 speakers, with 10 sentences in 2 sessions per speaker. The first session was used for enrollment, and the second session was used for verification. During verification, a combination of 3 sentences was used, and the speaker is accepted only if the matching results come within the top 3 ranks for at least 2 of the 3 sentences. An exhaustive $^{10}C_3$ combinations were tested for each speaker of the database. The False Acceptance Rate (FAR) and the False Rejection Rate (FRR) were calculated, and the Equal Error Rate (EER) is taken as the average of the FAR and FRR.

Figure 12 shows a plot of the EER for PFC, LPCC taken

individually and the joint audio-visual features. It can be seen that there is an improvement in the performance of the combined audio-visual features.

6. Conclusion and Future Work

In this paper, the performance of the proposed face profiles as a feature for representing visual speech is evaluated for a text-dependent audio-visual person authentication task using DTW algorithm. The improved performance of joint audio-visual features over audio and video taken individually shows that use of bimodal information gives a better performance than the use of single modality. Hence the use of PFCs is justified.

Further improvements can be made in the visual features extraction by using a robust algorithm for face localization that is invariant to face pose. This will reduce the restrictions on the speaker. Use of other acoustic features such as pitch and intonation can further improve the performance of the system. The proposed PFC can also be used for a text-independent authentication using pattern modeling techniques based on Artificial Neural Networks and Hidden Markov Models.

References

[1] Ming-Hsuan Yang, David J. Kriegman, and Narendra Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 34–58, January 2002.

[2] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade, "Neural network - based face detection," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco, CA*, pp. 203–207, 1996.

[3] C. Garcia and G. Tziritas, "Face detection using quantized skin color region merging and wavelet packet analysis," *IEEE Transactions on Multimedia Vol.1, No. 3*, pp. 264–277, September 1999.

[4] R. L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 696–706, May 2002.

[5] P. Jonathon Phillips, Harry Wechsler, Jeffrey S. Huang, and Patrick J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms.," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.

[6] G. T. Poulton, N. A. Oakes, D. G. Geers, R. Y. Qiao, M. D. S. Seneviratne, N. E. Frampton, Y. Choi, and J. I. Agbinaya, "The CSIRO PC-check system," *Proceedings of International Conference on Audio and Video based Biometric Person Authentication, Washington, D.C.*, pp. 160–165, June 1999.

[7] K. L. Sum, S. H. Leung, Alan W. C. Liew, and K. W. Tse, "A new optimization procedure for extracting the point-based lip contour using active shape model," *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, 2001.

[8] C. Bregler and Y. Konig, "EIGENLIPS for robust speech recognition," *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, pp. II–669–II–672, 1994.

[9] Iain Matthews, Timothy F Cootes, J. Andrew Bangham, Stephen Coxand, and Richard Harvey, "Extraction of visual feature for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 198–212, February 2002.

[10] Y. Yacoob and L.S. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 636–642, 1996.

[11] Jinu Mariam Zachariah, *Text-Dependent Speaker Verification Using Segmental, Suprasegmental and Source Features*, M.S. Thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai, submitted 2001.

[12] Meera M. Blattner and Ephraim P. Glinert, "Multimodal integration," *IEEE Multimedia*, pp. 14–24, 1996.

[13] Jayant M. Naik, "Speaker verification: A tutorial," *IEEE Communications Magazine*, pp. 42–48, Jan. 1990.

[14] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.