Multimodal Fuzzy Fusion for Biometric Identity Management

GIRIJA CHETTY, DAT TRAN, and DHARMENDRA SHARMA
School of Information Sciences and Engineering
University of Canberra, Australia

Abstract: - Biometric identity management based only on the single biometric modality is not accurate or robust enough to be used in uncontrolled environments. This paper describes a fusion of face and voice biometric traits, based on fuzzy logic approach for speaker identity verification. In this approach, a scheme based on membership function and fuzzy integral is proposed to fuse information from the two modalities. Equal Error rate is used to evaluate the fusion scheme. Experimental results show the fusion scheme improves identity verification performance substantially and makes the system robust to environmental degradations such as acoustic noise and visual compression artefacts.

Key-Words: - Multimodal, Human computer interfaces, Face, Voice, Fusion, Fuzzy logic

1 Introduction

Recently there has been significant interest in use of biometric based identity management solutions as compared to pin and password based approaches [1]. Biometric identity management based on user-friendly biometric traits such as face and voice, find better user acceptance in applications involving multi-modal human computer interfaces in the areas such as banking, and E-commerce systems [2]. It is known that humans perceive in a multimodal manner, and people with impaired hearing use lip-reading to complement information gleaned from their perceived degraded audio signal. Previous work in this area is usually based on the use of audio, [3], or static facial images (face recognition) [4]. Multimodal interfaces based on face and voice biometric traits find better user acceptance due to their non-intrusiveness and availability of low-cost off the shelf components. However, most of human-computer interfaces pre-dominantly use the single mode voice only or face only biometric traits. But fusion of audio and visual modes, it is possible to improve the performance and robustness to environmental degradations. Moreover, the addition of the dynamic visual mode complements the audio mode, and increases the reliability for noisy conditions and increase the performance for clean conditions. Also, it would be increasingly difficult for an imposter to impersonate both audio and dynamical visual information simultaneously.

The aim of the current study was to implement a fuzzy logic based approach for fusion of both dynamic visual and audio features, and to achieve a more reliable and robust multimodal speaker identity verification in noisy audio and visual operating environments. We describe the proposed fuzzy logic scheme in the next section, followed by details of experimental set up for fuzzy fusion approach in section 3. The details of the experimental results are described in section 4, and the paper concludes with conclusions and further scope of the study in section 5.

2 Multimodal Fuzzy Fusion Scheme

In this study, the multimodal fusion is between face and voice features of a video. The goal of fusion is to fuse important information from each modality. Fig. 1 illustrates the main steps of our fusion scheme, and the scheme is outlined below:

Step 1. We first compute acoustic and visual features from the video frames. In this study we used MFCCs for acoustic features and Eigen faces for visual features.

Step 2. The acoustic and visual features were normalized prior to fusion. The normalized vector v of an original vector $\wedge$ is defined as

$$\upsilon = \frac{\Omega}{\sqrt{(\Omega^T \Omega}} \tag{1}$$

Step 3. The purpose of fuzzification is to map input vector v from each modality to values from 0 to 1, representing evidence that the object satisfies the class hypothesis *kC*. The generation of membership function is very important [5] [6]. In this paper, we propose a histogram-based method for generating the membership function.

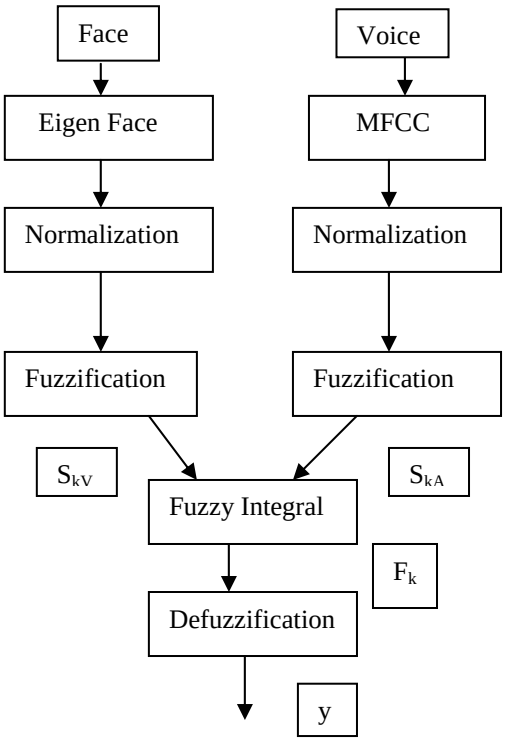

Fig. 1. Flowchart of the fuzzy fusion scheme

Let x be the distance between input object and its class, and $h(x)$ be the histogram of x, which provides information regarding the distribution of distance. Membership function $u(x)$ can be constructed as follows

$$u(x) = \int_{x}^{+\infty} h(x)dx \qquad (2)$$

From Eq. 2, we construct membership function $u(x)$ for each feature vector. Let $\varepsilon_k = \|v - v_k\|$, where v_k is the vector describing the k th class. The fuzzification result S_k is computed as

$$S_k = u(\varepsilon_k) \qquad (3)$$

Step 4. Fuzzy integral considers the objective evidence supplied by each source (called the h - function) and the expected worth of each source (via a fuzzy measure) [8]. Let x_i represent the visual features, and x_2 represent the audio features. The fuzzy density value $g^i = g\{x_i\}$ is determined via statistical measurements on errors rates of the single modality x_i . Thus the output of fuzzy integral Fk *can* be expressed as

$$F_k = \begin{cases} \max(\min(S_{kV}, g^1), S_{kA}) \\ \max(\min(S_{kA}, g^2), S_{kV}) \end{cases} \quad \begin{matrix} S_{kV} > S_{kA} \\ else \end{matrix} \qquad (4)$$

where S_{kV} *is* the fuzzification result of visual features, and S_{kA} *is* the fuzzification result of audio features.

Step 5. We classify the audio-visual features into a specific class if the fuzzy integral F_k had the output of fuzzy integral

$$y = \arg\max_k F_k \qquad (5)$$

3. Experimental Set Up

3.1 Experimental Data

The type of data used is the UCBN database [9], a free to air broadcast news database. The broadcast news is a continuous source of video sequences that can be easily obtained or recorded, and has optimal illumination, colour, and sound recording conditions. However, some of the attributes of broadcast news database such as near-frontal images, smaller facial regions, multiple faces and complex backgrounds require an efficient face detection and tracking scheme to be used. The database consists of 20-40 second video clips for anchor persons and newsreaders with frontal/near-frontal shots of 10 different faces (5 female and 5 male). Each video sequence is 25 frames per second MPEG2 encoded stream with a resolution of 720 × 576 pixels, with corresponding 16 bit, 48 kHz PCM audio.

Fig. 2. Sample Face Images from UCBN corpus

3.2 Audio and Visual Feature Extraction

The audio signal was first pre-emphasized to increase the acoustic power at higher frequencies using the filter H (z) =1/(1-0.97z -1). The pre-emphasized signal was divided into frames using a

Hamming window of length 20 ms, with overlap of 10 ms to give an audio frame rate, FA,, of 100 Hz. Mel-frequency cepstral coefficients (MFCC's) of dimension 8 were extracted from each frame.

The faces from the images were detected [7] and were converted to gray scale values. The images were histogram equalized and pixel mean subtracted. Then 20 Eigen-face coefficients were obtained from the face images by principal component analysis [8]. Since the audio and video was acquired at different frame rates, a matching frame rate was obtained with appropriate rate interpolation.

For audio-visual fusion, we first constructed membership function for each modality from acoustic MFCC and Eigen face features. We set $g^1 = 0.8$, and $g^2 = 0.9$ in Eq. 4 after statistically analyzing EER rates of two modalities individually. Verification performance in terms of EERs of each modality (using Eigen faces and MFCCs) and fusion (using the proposed scheme) are shown obtained and is shown in Table 1.

4 Experimental Results

We used a training set from 5 sets and used 2 sets for testing from the UCBN corpus. Ten subjects, five male and five female were used for training and testing in the text independent mode. To test the influence of acoustic and visual degradations we added white Gaussian noise of to degrade the signal to 12 dB signal to noise ratio (Original SNR was 30 dB), and JPEG compression of quality factor QF = 20 (Original QF of image is 60).
Training was done on clean set with no white Gaussian noise and no JPEG compression artificially added. For testing four different experiments were conducted.

Experiment 1: Audio only, Visual only and AV fusion with no acoustic noise and no compression artefacts.
Experiment 2: Audio only, Visual only, and AV fusion with acoustic noise (12 dB SNR) and no compression artefacts.
Experiment 3: Audio only, Visual only and AV fusion with no acoustic noise, and JPEG compression (QF= 20).
Experiment 4: Audio only, Visual only, and AV fusion with acoustic noise (12 dB SNR) and JPEG compression (QF=20).

Table 1 shows the equal error rates for Experiments 1 to 4. As can be seen in Table 1, there is a significant improvement in overall performance with fuzzy fusion approach as compared to audio only or video only case. For experiment 1 with no acoustic noise and visual artefacts, audio only mode performs best with 8.35%, and audio-visual fusion mode has EER of 8.05%, and does not lead to remarkable increase in performance.
With test conditions including acoustic noise and visual artefacts, the performance deteriorates for single mode audio and video only cases, with worst case EER of 18.45 % for video only case. However, the system is less sensitive for audio-visual mode, where the performance drops to just 8.46%, a marginal drop compared to audio only case. Hence, the proposed fuzzy fusion approach makes system less sensitive to acoustic and visual degradations and depicts operating efficiencies for more realistic scenarios as compared to the perfect databases developed in the controlled environments.

Table 1: Verification performance of fuzzy fusion

Experiment Type	Audio Only (EER)	Video Only (EER)	Audio Visual Fusion (EER)
Expt. 1	8.35 %	8.54 %	8.05 %
Expt. 2	13.96%	12.25%	7.98 %
Expt. 3	15.5 %	14.11%	8.01 %
Expt. 4	16.26 %	18.45 %	8.46 %

5 Conclusions

We presented a new fusion scheme to combine voice and face images for the purpose of speaker identity verification. Our scheme is based on using fuzzy integral to fuse the objective evidence supplied by each modality. We also designed a histogram-based method to generate membership. Experimental results show that the scheme is easy to implement, and improves verification performance substantially. Further work includes designing more effective schemes to deal with different types of challenging scenarios for next generation human computer interface applications.

6 References

[1] *Identity management solutions*, http://images.telos.com/files/external/Xacta_IDM_DS.pdf

[2] A.Corradini, M.Mehta, N.O. Bensen, J.C. Martin, and S. Abrilian, *Multimodal Input Fusion in Human Computer Interaction*, http://www.nis.sdu.dk/publications/2003/NATO-ASI_Armenia.pdf

[3] M.-C. Su, Y.-H. Lee, C.-H. Wu, and Y.-X. Zhao, *Low-cost Human Computer Interfaces for Disabled*, in Proceedings IASTED conference on Biomedical Engineering, 2003.

[4] L. Bretzner, I. Laptev, T. Lindeberg, S. Lenman, Y. Sundblad, *A prototype system for computer vision based human computer interaction* Technical report ISRN KTH/NA/P-01/09-SE, April 2001.

[5] Medasani, S., Kim, J., Krishnapuram, R.: *An overview of Membership Function Generation Techniques for Patter Recognition*. International Journal of Approximate Reasoning 19 (1998) 391–417.

[6] Keller, J. M., Osborn, J: *Training the Fuzzy Integral*. International Journal of Approximate Reasoning 15 (1996) 1–24.

[7] Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P. J.: Face Recognition: A Literature Survey. ACM Computing Survey 35 (2003) 399–458.

[8] Turk, M., Pentland, A.: Face Recognition Using Eigen face. IEEE Conf. on Computer Vision and Pattern Recognition (1991) 586–591