Comparative Study on Feature, Score and Decision Level Fusion Schemes for Robust Multibiometric Systems

Chia Chin Lip and Dzati Athiar Ramli

School of Electrical & Electronic Engineering, USM Engineering Campus, Universiti Sains Malaysia, 14300, Nibong Tebal, Pulau Pinang, Malaysia
Chinlip0102@hotmail.com, dzati@eng.usm.my

Abstract. Multibiometric system employs two or more behavioral or physical information from a person's traits for the verification and identification processes. Many researches have proved that multibiometric system can improve the performances of single biometric system. In this study, three types of fusion levels i.e feature level fusion, score level fusion and decision level fusion have been tested. Feature level fusion involves feature concatenation of the features from two modalities before the pattern matching process while score level fusion is executed by calculating the mean score from both biometrics scores produced after the pattern matching. Finally, for the decision level fusion, the logic AND and OR are performed on the final decision of the two modalities. In this study, speech signal is used as a biometric trait to the biometric verification system while lipreading image is used as a second modality to assist the performance of the single modal system. For speech signal, Mel Frequency Ceptral Coefficient (MFCC) is used as speech features while region of interest (ROI) of lipreading is used as visual features. Consequently, support vector machine (SVM) is executed as classifier. Performances of the systems for each fusion level is compared based on accuracy percentage and Receiver Operation Characteristic (ROC) curve by plotting Genuine Acceptance Rate (GAR) versus False Acceptance Rate (FAR. Experimental results show that score level fusion performance is the most outstanding method followed by feature level fusion and finally the decision level fusion. The accuracy percentages using 20 training data are observed as 99.9488%, 99.7534% and 99.6639% for the score level fusion, feature level fusion and decision level fusion, respectively.

Keywords: Multi-modal, speech signal, fusion level, verification, biometrics.

1 Introduction

Biometric system uses physiological or behavioral characteristics for verification and identification of individual. The applications have been used widely in various areas especially for smart access control, forensic identification, transaction verification system (ATM) and computer network security. There are two modes of biometric systems which are identification and verification. Identification is the process to

S. Sambath and E. Zhu (Eds.): Frontiers in Computer Education, AISC 133, pp. 941–948.
springerlink.com

identify the unknown user while verification involves the process of accepting or rejecting the claimed identity [1]. Typically, the structure of biometric verification process can be categorized into four main mechanisms i.e. data acquisition, feature extraction, pattern-matching and decision for accepting or rejecting the clients [2]. Data acquisition process starts with the enrollment of raw training data from the registered speakers who are legal users to access the systems for feature modeling. Then, the current user data collection is executed for the purpose of verification during real system implementation. Feature extraction is done in order to extract the relevant information from the raw data while pattern matching is the process of classifying the current user data into authentic or imposter by comparing with the developed model. By setting a particular threshold for soft decision method and employing Boolean logic for hard decision method, the decision whether to accept or reject the current user is then made in this mechanism.

Researches in multibiometric system reported that combining different biometric sources can improve the performance of the single biometric system [3].The level of fusion in multibiometric systems is categorized into two categories i.e. fusion before matching and fusion after matching. Fusion before matching includes fusion at the sensor and feature levels whereas fusion after matching includes fusion at the score and decision levels.

Fusion at a sensor module employs multiple sensors to capture biometric trait of an individual. For example, an infrared sensor and visible light sensor are used to capture of individual face information for face recognition system [4] while another experiment used optical and capacitive sensor to capture fingerprint image [5]. For the fusion at a feature module, the features of different biometrics are combined before executing the classifier. This study can be found in [6]. In this work the face and palmprint information are extracted using Gabor based image pre-processing and PCA techniques. In another experiment, side face and gait have been used as biometric traits [7]. The features of face is extracted using principle component analysis (PCA) from enhanced side face image (ESFI) while the gait features is obtained using gait energy image (GEI).

Consequently, example of the fusion at score level can be found in [8]. Score normalization techniques has been investigated in this study due to heterogeneous of matching score output by face, fingerprint and hand geometry modalities. Several normalization techniques such as min-max, z-score and tanh normalization have been explained in detail. Previous research by [3] proposed an adaptive weighting to fuse the audio and visual scores. Lastly, one example of fusion at decision level is presented by [9] which demonstrated multimodal speaker recognition system that combines audio, lip texture and lip motion. The fusion is performed by reliability weighted summation (RWS) decision rule. In another study by [10] proposed a new approach to combine decisions from face and fingerprint classifiers. In this study parametric and non parametric Linear Discriminant Analysis has been utilized.

In this research, the performance of fusion approach by employing speech and lip data are experimented and three levels of fusions i.e. feature, score and decision are integrated separately in the architecture of the multibiometric system so as to distinguish between a genuine and imposter user. The main objective of this study is to obtain the most outstanding fusion approach for the multibiometric system architecture. For validation, the performances of the single systems of the audio and visual systems are first experimented and subsequently followed by the fusion

systems. Performances based on different numbers of training data for modeling process are also evaluated in this study.

2 Methodology

The database of both audio and visual data used in this project is obtained from Audio-Visual Digit Database from [11]. This database consists of the digitized speech signals of the recording voices of 37 speakers (16 female and 21 male) stored as monophonic 16 bit, 32 kHz and in WAV format. The recording process is performed in three different sessions. Each speaker performed 20 repetitions of digit zero to nine for each session. Therefore, 60 audio data for each speaker obtained from each session. It consists of 2220 data for digit zero. The visual data of 37 speakers is stored as a sequence of JPEG images with a resolution of 512 x 384 pixels. Each speaker consists of 60 sequences of face images (20 sequences from each session) hence in total of 2220 images from entire speakers. For the purpose of this study, both audio and visual data from the first session are used for modeling while the testing data are used from the data of the second and third sessions.

For feature extraction, two types of features have been extracted i.e. lip and speech signal information. For lip feature extraction, the information is extracted in term of region of interest (ROI) while for speech signal feature extraction; the method of mel frequency cepstral coefficient (MFCC) is employed. Both techniques can be found in [12] and [13], respectively. For the lipreading image, the sequence of lip images is extracted as the region of interest (ROI) while the speaker uttering the word zero to nine. Comparing to the static lip, lipreading image contains more information in term of behavioral and physical characteristics which gives extra advantages for the accurate verification system [14]. Further description on lipreading feature extraction can be found in [14].

Speech signal feature extraction is analyzed on a frame by frame basis with frame length 20 ms^{-1} and frame interval 10 ms^{-1}. The preprocessing procedure consists of preemphasis, framing and windowing. MFCC is processed in frequency domain by computing the entire framed and windowed signals using the Discrete Fourier transform (DFT). After obtaining the signal's spectrum, the filter bank processing is then employed in order to obtain the spectral features at defined frequency at its exit. Log energy computation which consists of computing the logarithm of the square magnitude of the filter bank outputs is then performed. The final step for MFCC processing is the computation of mel frequency cepstrum by performing the inverse DFT on the logarithm of the magnitude of the filter bank output.

SVM theory can be found in [15] and [16]. SVM is a classifier which can classify samples within two or more classes. In this study, N-class classifier is used which employ the method of winner take all classification or one against the others. The value of N is depends on the number of person in the database thus N binary SVM models are created. Each model in database is trained to discriminative one class from the remaining N-1 classes. SVM in its simplest form, linear and separable case can be defined as the optimal hyper plane that maximizes the distance of the separating hyper plane from the closest training data point called the support vectors. To summarize, let

W be normal to the decision region and the N training examples represented as the pair (x^i, y^i), $i = 1,2, \ldots L$.

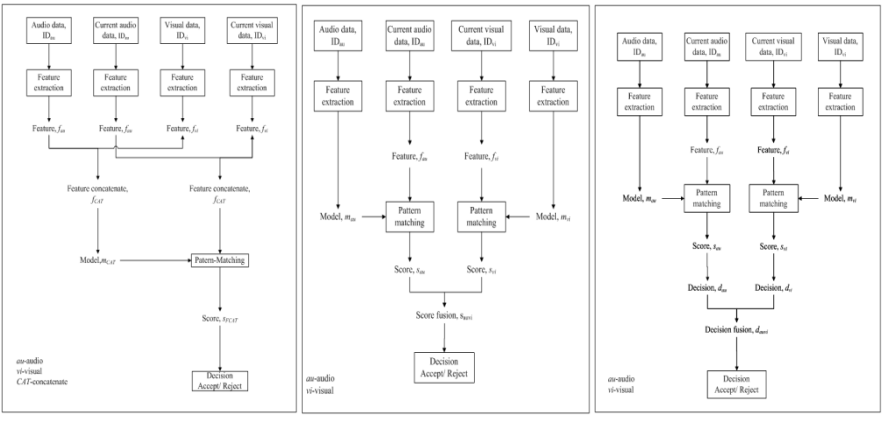
$$D = \{(x^1, y^1), \ldots, (x^L, y^L)\}, \quad x \in \Re^n, \ y \in \{-1,1\} \tag{1}$$

The points that lie on the hyper plane to separate data satisfy,

$$\langle w, x \rangle + b = 0 \tag{2}$$

where b is the distance of the hyper plane from the origin. The hyper plane should have the same distance for each class nearest point and the margin distance is twice. The non-linear mapping is imperative in the case of the linear boundary is inappropriate. In practice, the kernel function is introduced to transform the data points in the input space to the feature space. In this study, the polynomial function is used.

For feature level fusion, features of speech and lip are combined as a concatenated feature after the process of feature extraction taking place. For the score level fusion, the scores produced after the pattern matching of both audio and visual systems are combined using the mathematical operators for examples sum rule, product rule, maximum and minimum. Finally, for the decision level fusion, the hard decision of accept or reject from both biometric systems is fed to a logic operation using i.e. AND method for the final decision. The architecture of the multibiometric systems using feature, score and decision level schemes are depicted in Fig. 1, Fig. 2 and Fig.3, respectively.



(a) (b) (c)

Fig. 1. Audio and visual verification systems using feature (a), score (b) and decision (c) level fusions

3 Results and Discussions

The performance evaluation is calculated based on the percentage of accuracy, genuine acceptance rate (GAR), false acceptance rate (FAR) and equal error rate (EER). Percentage of accuracy is defined as the capability of the classifier in

providing the correct output corresponding to the input data [17]. For the single modal systems, feature fusion system and score fusion system, the performances can be subsequently evaluated using GAR, FAR and EER by comparing the soft scores produced after the pattern matching with a specified threshold. GAR is the percentage of authorized persons is admitted by the system. Meanwhile, FAR which is also known as false non match rate is when actual impostor is accepted as genuine. EER is the equal value of FAR and FRR. In general, the lower EER value, the higher the performance of the system [18].

3.1 Audio and Visual Systems Using 3 Training Data

Comparison on accuracy percentages using feature, score and decision level fusion schemes for each speaker in the database is presented in Fig. 2 (left). Average of accuracy percentages for entire speakers are observed as 99.11435%, 99.5471% and 98.7381% for feature, score and decision fusion systems, respectively. Subsequently, performances of speech, lip, feature fusion and score fusion systems based on GAR and FAR percentages by using 3 training data are shown in Fig. 2 (right). At 10% of FAR, speech, feature fusion and score fusion systems obtain 97%, 99% and 100% of GAR, respectively. Score fusion achieves 100% of GAR at 0.3 % FAR. Based on EER percentages, 4.3206%, 3.9461% and 0.3425% are observed for speech, feature fusion and score fusion systems.

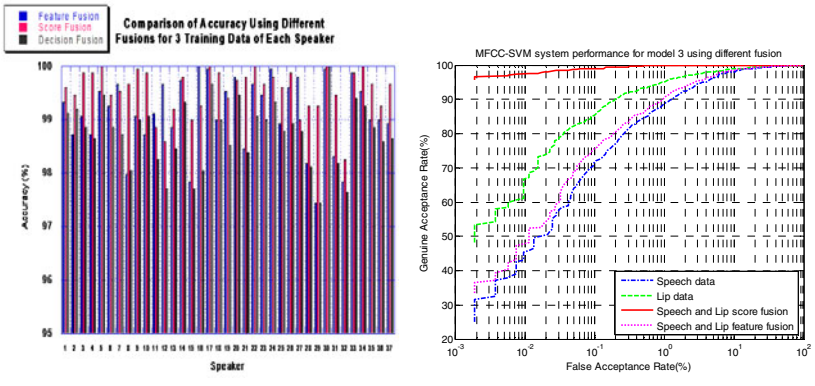


Fig. 2. System performance based on 3 training data

3.2 Audio and Visual Systems Using 6 Training Data

Fig. 3 (left) compares the performances of feature, score and decision level fusion schemes based on the accuracy percentages for each speaker in the database. 99.5452%, 99.8319% and 99.3517% of average of accuracy percentages for entire speakers are observed for the feature, score and decision fusion systems, respectively. Afterward, performances of speech, lip, feature fusion and score fusion systems based on GAR and FAR percentages by using 6 training data are shown in Fig. 3 (right). At

1% of FAR, speech, feature fusion and score fusion systems obtain 97%, 98% and 100% of GAR, respectively. Score fusion achieves 100% of GAR at 0.007 % FAR. Based on EER percentages, 1.8168%, 1.4471% and 0.0582% are observed for speech, feature fusion and score fusion systems.

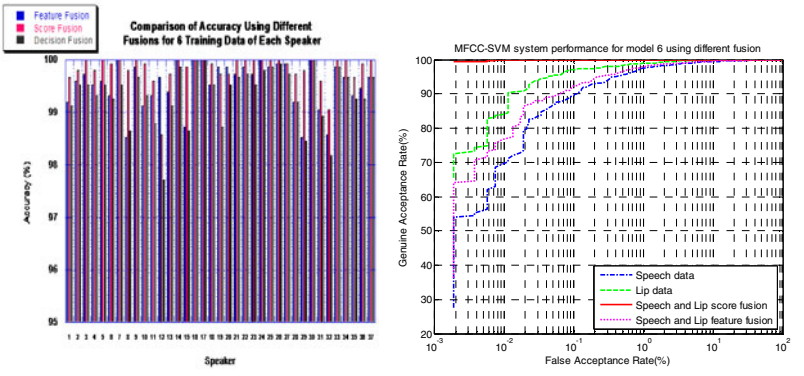


Fig. 3. System performance based on 6 training data

3.3 Audio and Visual Systems Using 10 Training Data

For the 10 training data, comparison on accuracy percentages using feature, score and decision level fusion schemes for each speaker in the database is presented in Fig. 4 (left). For feature, score and decision fusion systems, average of accuracy percentages for entire speakers are observed as 99.6840%, 99.8886% and 99.5157%, respectively. Consequently, Fig. 4 (right) shows the performances of speech, lip, feature fusion and score fusion systems based on GAR and FAR percentages. At 1% of FAR, speech, feature fusion and score fusion systems obtain 98%, 98.5% and 100% of GAR, respectively. Score fusion achieves 100% of GAR at 0.007 % FAR. Then, 1.5700%, 1.2660% and 0.0075% of EER percentages are observed for speech, feature fusion and score fusion systems.

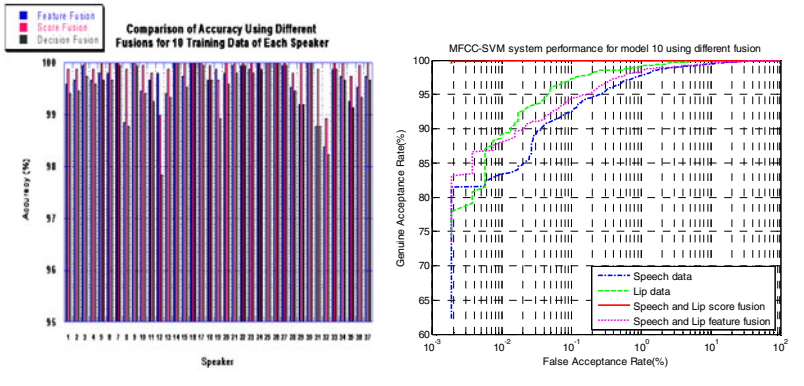


Fig. 4. System performance based on 10 training data

3.4 Audio and Visual Systems Using 20 Training Data

Finally, comparison on accuracy percentage using feature, score and decision level fusion schemes for each speaker in the database is presented in Fig. 5(left). From the observation, average of accuracy percentages for entire speakers are found as 99.7434% , 99.9488% and 99.6639% for feature, score and decision fusion systems, respectively. Subsequently, performances of speech, lip, feature fusion and score fusion systems based on GAR and FAR percentages by using 20 training data are shown in Fig. 5 (right). Speech, feature fusion and score fusion systems obtain 98.5%, 99% and 100% of GAR, respectively at 1% of FAR. Score fusion achieves 100% of GAR at 0.003 % FAR. Then, 1.1524%, 0.8268% and 0.0019% of EER percentages are observed for speech, feature fusion and score fusion systems.

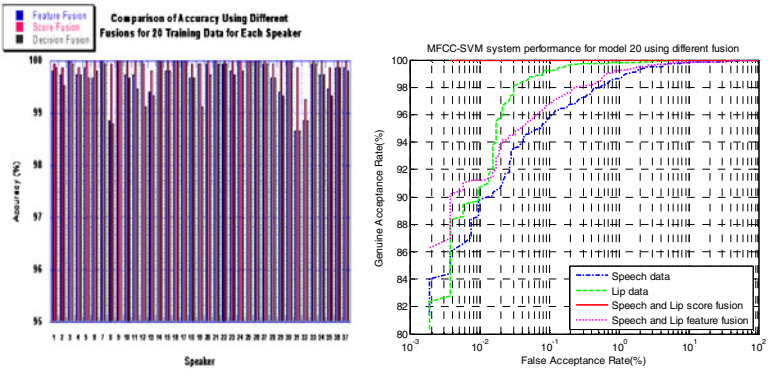


Fig. 5. System performance based on 10 training data

4 Conclusion

Three types of fusion levels i.e. feature, score and decision have been evaluated in order to investigate the outstanding level of fusion to be integrated in multibiometric system. In this study, lip trait has been used as a second modality to improve the performance of speech based biometric system. Experimental results shows that score level fusion gives the best performance compared to feature level fusion and decision level fusion while the performance of feature level fusion is better than decision level fusion. Besides that, the number of training data is also contributed to the performance of the system. Performance increases with the increasing of the number of training data. However, the increasing number of data causes high time consuming to the system.

Acknowledgments. This research is supported by the following research grants: Research University (RU) Grant, Universiti Sains Malaysia, 100/PELECT/814098 and Incentive Grant, Universiti Sains Malaysia.

References

1. Jain, A.K., Ross, A., Prabhakar, S.: An Introduction to Biometric Recognition. IEEE Transactions on Circuits and Systems for Video Technology 14(1), 4–20 (2004)
2. Reynolds, D.A.: An Overview of Automatic Speaker Recognition Technology. IEEE Transactions on Acoustic, Speech and Signal Processing 4, 4072–4075 (2002)
3. Ramli, D.A., Samad, S.A., Hussain, A.: A Multibiometric Speaker Authentication System with SVM Audio Reliability Indicator. IAENG International Journal of Computer Science 36(4), 313–321 (2008)
4. Kong, S.G., Heo, J., Abidi, B.R., Paik, J., Abidi, M.A.: Recent Advances in Visual and Infrared Face Recognition–A Review. Computer Vision an Image Understanding 97(1), 103–135 (2005)
5. Marcialis, G.L., Roli, F.: Fingerprint Verification by Fusion of Optical and Capacitive Sensors. Pattern Recognition Letters 25, 1315–1322 (2004)
6. Yong, F.A., Xiao, Y.J., Hau, S.W.: Face and Palmprint Feature Level Fusion for Single Sample Biometrics Recognition. Neurocomputing 70, 1582–1586 (2007)
7. Zhou, X., Bhanu, B.: Feature Fusion of Side Face and Gait for Video based Human Identification. Pattern Recognition 41, 778–795 (2008)
8. Jain, A., Nandakumar, K., Ross, A.: Score Normalization in Multimodal Biometric Systems. Pattern Recognition 38, 2270–2285 (2005)
9. Cetingul, H.E., Erzin, E., Yemez, Y., Tekalp, A.M.: Multimodal Speaker/Speech Recognition using Lip Motion, Lip Texture and Audio. Signal Processing 86, 3549–3558 (2006)
10. Patra, A., Das, S.: Enhancing Decision Combination of Face and Fingerprint by Exploitation of Individual Classifier Space, an Approach to Multi-Modal Biometry. Pattern Recognition 41, 2298–2308 (2008)
11. Sanderson, C., Paliwal, K.K.: Noise Compensation in a Multi-Modal Verification System. In: Proceeding of International Conference on Acoustic, Speech and Signal Processing, pp. 157–160 (2001)
12. Becchetti, C., Ricotti, L.R.: Speech Recognition: Theory and C++ Implementation. John Wiley & Son Ltd., England (1999)
13. Furui, S.: Cepstral Analysis Technique for Automatic Speaker Verification. IEEE Transactions on Acoustic, Speech Signal processing 29(2), 254–272 (1981)
14. Samad, S.A., Ramli, D., Hussain, A.: Person Identification Using Lip Motion Sequence. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAI), vol. 4692, pp. 839–846. Springer, Heidelberg (2007)
15. Gunn, S.R.: 2005.Support Vector Machine for Classification and Regression. Technical report. Faculty of Engineering, Science and Mathematics, University of Southampton (2005)
16. Vapnik, V.N.: The nature of statistical learning theory. Springer, Heidelberg (1995)
17. Hauck, W.W., Koch, W., Abernethy, D., Williams, R.L.: Making Sense of Trueness, Precision, Accuracy, and Uncertainty. Pharmacopeial Forum 34(3), 838–842 (2008)
18. Fawcett, T.: An Introduction to ROC Analysis. Pattern Recognition Letters 27, 861–874 (2006)