Fusion Strategies for Multimodal Biometric System Using Face and Voice Cues

Poornima Byahatti\textsuperscript{1} and Madhura S Shettar\textsuperscript{2}

\textsuperscript{1}Department of Automation & Robotics, KLE Technological University, Hubballi, India.
\textsuperscript{2}School of Computer Science & Engineering, KLE Technological University, Hubballi, India.
\textsuperscript{1}poornima.byahatti@kletech.ac.in, \textsuperscript{2}madhura.shettar@kletech.ac.in

Abstract. Deployment of biometric systems in the applications of real world includes the most of unimodal biometric systems. Unimodal biometric system based on the information collected from single source. Sometimes single source of information may not identify the individual correctly because of some limitations such as Non-universality, Noisy data, Intra-class variation, Spoof attacks and Intra-class similarities. Various limitations of unimodal biometric systems are overridden by the multimodal biometric systems which involves multiple sources of information. Multimodal systems can be constructed by fusing information of multiple modalities. This fusion can take place at various steps of processing such as at image acquisition, extraction of features of the traits, matching of test vectors with trained vectors and during decision taking based on classification. This paper presents a system of multimodal biometrics using face and voice biometric traits by including four fusion methods. Fusion takes place at i) feature level using concatenation of face and voice features, ii) score level using method involving the maximum of mode of scores obtained from two matchers, iii) rank level using borda count & iv) decision level fusion using logical conjunction (AND). Fusing of Log Gabor & Local Binary Pattern (LBP) takes place at the facial feature extraction. The voice features are also fused using Mel Frequency Cepstral coefficients (MFCCs) and Linear Predictive Coefficient features (LPC). Computation of similarity between test feature vectors and training vectors is carried out using Euclidian distance during matching process. KNN Classifier is used during decision making. Performance evaluation of these techniques are also carried out using performance measures such as Accuracy, False Acceptance Rate (FAR), False Rejection Rate (FRR) and ROC curves.

Keywords: Log Gabor, LBP, MFCCs, LPC, Feature fusion, Score fusion, Rank fusion and Decision fusion.

1. Introduction
Implementation of recognition techniques are needed for protection of resources in nowadays as the World of technology evolves. Various ways are available to achieve authentication & authorization techniques. Most of the techniques are being overridden by biometric system because it recognizes the individual identity automatically based on behavioral & anatomical characteristics like signature, voice, face, iris, fingerprint etc. There are two main divisions of biometric systems depending upon the number of modalities used in the biometric system. Those are unimodal biometric systems (make use of single
trait such as face, voice, etc.) & multimodal biometric systems (make use of two or more traits such as (face & iris), (face & voice), etc.)

Some of the limitations embedded in unimodal biometric systems are Non-universality, Noisy data, Intra-class variation, Spoof attacks and Intra-class similarities. Multimodal biometric systems override the several disadvantages of unimodal biometric systems [1]. Unimodal biometric system is not suited for all applications because it makes use of single trait.

There are five ways in which we can achieve multi-biometric system. Those are: 1). Multi-algorithm biometric systems, which involve processing of same biometric information with more than one algorithm. 2). Multi-sensor biometric systems in which acquisition of images takes place using different sensors. 3). Multi-instance biometric systems make use of multiple instances of same biometric trait. 4). Multi-sample biometric systems in which multiple samples of biometric trait are collected from the single sensor. 5). Multi-modal biometric systems make use of the biometric data collected from multiple traits. Only single modality can be used to achieve authentication in first four ways in the above list but in fifth way, multiple modalities can be used to attain authentication. Fusing of the data of multiple traits like fingerprint, iris, voice, palmprint, face, etc. constitutes the multimodal biometric system.

There are two main different paths in which fusion of information can be achieved. Those are: fusion before matching and fusion after matching. Attainment of fusion before matching can be again done in two traditions as 1). Sensor fusion technique: In this technique many sensors collect the information from the same biometric trait followed by the integration of such information to construct new information. Features are extracted from the new data then they are used for further processing in recognition. 2). Feature fusion technique: Extraction of features from multiple traits is carried out in this technique at first. Then there is concatenation of features of multiple traits to new feature joint vector. Sometimes these joint vectors are high dimensional vectors so some reduction techniques are needed for selection of useful information for further processing.

Accomplishment of fusion after matching may also be done in three traditions as 1). Score fusion technique: In this technique each individual system calculates its own scores that are nothing but matching similarity measures of input vectors with template vectors. Scores of individual matchers are fused in order to calculate final scores then; these scores will help in taking final decision. 2). Rank fusion technique: Identification and verification are the two types of recognition, where rank fusion technique commonly suits during the identification of the person instead of verification. In this fusion, final decision is being established by determining a new rank obtained from consolidation of multiple ranks associated with an identity. 3). Decision fusion technique: In this technique each biometric system classifies the persons separately and then decisions are fused in order to claim final decision based on various fusion methods available. This fusion method has less information for fusing as compared to other fusion methods.

In this paper, multimodal biometric system is employed in order to evaluate the performance of fusion strategies using face and voice biometric traits in terms of accuracy, FRR and FAR. Fusion of Log Gabor and LBP is carried out for facial feature extraction. MFCCs and LPC features are used during voice feature extraction. Fusion strategies employed in system are feature fusion using concatenation, score fusion using max-min, rank fusion using borda count and decision using AND methods. Classification is done by using KNN classifier with one nearest neighbour. This paper is organized as follows: section 2 summarizes literature survey, section 3 presents the proposed methodology, section 4 demonstrates the experimental results and section 5 illustrates the conclusion.

2. Literature Survey

The literature survey is carried out in order to study the different fusion approaches in multimodal biometric systems and to identify the challenges and issues in multimodal biometric systems. Several research papers have been referred and the outlines of these papers are described below:

Multimodal biometric system evaluates performance of decision fusion with logical operators AND & OR for iris and fingerprint biometric traits using K-nearest neighbour, Neural based classifier and Hidden markov model and proved that neural classifier with AND based fusion attains the highest accuracy of 90.30% in [2]. Multi-algorithmic biometric system is developed by fusing features of face
& iris and multimodal biometric system is employed by using score fusion method of weighted sum in [3]. Daugman’s Gabor filters are used for iris. For facial feature extraction Gabor filters and Local binary patterns are used. The value for EER of the developed system is 1.48% that is 85% improvement as compared to unimodal system. Multimodal system involving combination of iris, voice and face using score fusion method was developed. Simple sum rule is used for scores fusion after min-max normalization. Authors achieved accuracy in terms of GAR-92% in [4]. By making use of finger veins, new approach of multi-instance identification system is established based on score fusion method in [5]. EER values of proposed system using three different finger veins (index, middle, ring) in terms of sum weighted, min and max are 0.11415, 0.15714 and 0.08571 respectively. Rank fusion with improvement involving two combinations such as serial and parallel is adapted using in multi-instance, multi-algorithm and multimodal biometric systems in [6]. Values of RR for Multi-instance, multi-algorithm and Multi-modal are 83.6%, 92.63% and 93.4% respectively. Multimodal biometric system is presented using three biometric traits involving iris, voice and face in [7]. Delta Delta Mel Frequency Centrall Coefficients (DDMFC) features are used for voice recognition followed by extraction of face features using Scale Invariant Feature Transform (SIFT-based) features then LAB-based features are used in iris recognition. Decision module fusion based on majority voting is performed and attained 90% of recognition rate.

Comparative analysis of performance of various image fusion techniques is carried out and showed that feature fusion is the best method for fusing information than other techniques through results obtained during experimentation. Values of feature level fusion for accuracy, FAR and FRR are 99%, 0.029% and 1.0%. Values for score and decision level in terms of accuracy are 97% and 98% respectively in [8]. System evaluates the performance of unimodal and multimodal (score fusion) systems using face and voice biometric traits. Results varied based on the feature extracted from XM2VTS database and score fusion techniques used and proved multimodal with score fusion achieves the higher accuracy as compared to unimodal system in [9]. Score fusion strategy of weighted score for fusing the palmprint and speech signal is explored in [10]. Palm features are extracted by modified canonical method. Subband based Centrall coefficients are used for extraction of speech signal features. Clean & degraded databases are applied on system. Evaluation of system is done in terms of EER, FAR and FRR. Results showed that system attains higher accuracy and achieves the EER of 3.54% & 9.17% for clean and degraded dataset. Feature extraction level fusion is carried out based on the information from the two traits of fingerprint and iris in [11]. Decision about recognition is taken for multimodal system and for each unimodal biometric system. The values for FAR and GAR are 1% and 98% respectively. SPR fusion method based on the combination of simple sum and product rule by using face and voice biometric traits in [12]. The double sigmoid normalization technique is used for the normalization of scores before fusion. The sum rule, maximum rule, product rule and minimum rule is also performed in order to compare with the system. Experimental results declared that SPR fusion scheme attains higher accuracy of 89.12% on average. Multimodal system is developed for fusing face and speech using decision fusion techniques with both AND & OR in [13]. Combination of PCA and DCT are used for face identification and for speech text independent mode is used. Results proved that system achieves better results than unimodal systems. Identification of various issues and challenges in multimodal biometric systems is done through the literature survey such as sensors adjustment and accumulation of dust on sensors in score fusion, non-compatibility of features during feature fusion, normalization of scores due to heterogeneity during score fusion, selection of system design of serial and parallel during rank fusion and limited amount of information during decision fusion techniques.

After carrying out the literature survey, it is observed that various fusion models are available based on the requirements. Numerous systems make use of different feature extraction methods, different fusion techniques at various levels [14]. Choosing among all for application is not an easy task so there is need of evaluation of those systems to know how reliable those are. Hence there is scope for developing system which evaluates the performance of fusion systems and eases way of selection.
3. Proposed Methodology

There is involvement of all processing modules such as image acquisition, extraction of features, creation of template, matching for similarity and concluding the decision in biometric multimodal systems as that of unimodal systems. Additionally, system involves the fusion of information at various levels. The fusion may take place at any of the following phases. 1). at the time of extraction of features of biometric traits. 2). at the time of comparison of input template with stored biometric templates. 3). at the time of concluding of decision based on classification.

Addressing of issues embedded in biometric systems like efficiency, applicability, robustness, accuracy & universality can be done with the implementation of fusion. Fig. 1 shows the block diagram for all levels of fusion. This system consists of two phases such as training & testing phase. In the proposed system, extraction of features from face biometric trait is carried out using both Log Gabor & Local Binary Pattern features. Voice feature’s extraction process makes use of Mel-Frequency Cepstral Coefficient & Linear Predictive Coefficient features from voice biometric trait. Respective features are stored during training according to fusion technique used. Euclidian distance is calculated for testing both the traits during comparison module. Finally, decision is made using K-Nearest Neighbour (KNN) classifier.

![Figure 1. Block diagram of fusion levels](image)

This proposed system involves the four fusion techniques and are described as:

During the process of feature level fusion at first, features are extracted from both face and voice biometric trait. Next, extracted features of both traits are joined together using concatenation fusion technique. Finally, resulting joint vector represents feature set for further processing. Principle component analysis is used in order to reduce the dimension of joint vector.

In this technique scores are calculated using Euclidian distance method for both face and voice trait’s features. These scores are fused using maximum score of mode method (max-min). Then final decision is taken based on the fused scores.

Rank fusion is almost similar to score level fusion with slight difference. It is well suited to identification process rather than verification. In proposed system Borda Count rank level fusion technique is used. Borda Count calculates sum of the ranks of individual matchers to obtain final rank.

Decision level technique is carried out using Logical Conjunction (AND) method. Individual matchers calculate scores for trait’s features then decisions are taken for both traits. The decisions output by individual matchers are fused using AND technique so when decisions of two matches are same then final decision is same about the matching.
3.1. Datasets
Image & voice sample acquisition is the first module of proposed system. The system is applied on two datasets. First system makes use synthetic database constructed by combining the AR database for face & voice database. AR database consists of 10 people where each image has size of 576x768 pixels. All frontal views of person are available. For AR database of faces some voice samples are downloaded and they are assigned to people with assumption that those are voice samples of faces of AR database. Fig. 2 shows the samples of face and voice of synthetic database.

Figure 2. Face and voice Samples of synthetic database

Second database applied on the project is own collected database of 20 persons. Both images and voice samples are collected from our surrounding people through camera and voice recording device respectively. Images are captured by camera with different poses. Voice samples involving utterance of 22 of each person are collected through device. Fig. 3 shows the face and voice samples of own collected database.

Figure 3. Face and voice samples of own database
3.2. Pre-processing of face and voice biometric traits

Pre-processing of face consists of three steps such as gray scale conversion which is the process of conversion of color RGB image to gray scale image, filtering process which is the process to highlight certain features or exclusion of some features of image. (Here, system makes use of Gaussian filtering) and enhancement of image which is the process of adjustment of digital images so that results are appropriate for display and further analysis of image.

Pre-processing of voice consists of following step of voice activity detection which is the process of separating speech from silent portion or other background noise. Here proposed system makes use of zero crossing rates for voice detection which is the number of times voice signal changes its sign.

3.3. Feature extraction

Features are the building blocks of biometrics so that features are extracted from traits in order to classify the individual through various feature extraction methods or techniques. In the proposed system following facial and voice feature extraction methods are used.

Facial feature extraction is carried out by fusing Local Binary Pattern and Log Gabor. LBP is an effortless and well-organized method which involves texture operator which identifies the image pixel by taking the threshold of all neighbours of that pixel & the result is considered as the binary number. The Equation 1 shows face description with LBP. Histogram of LBP can be represented as

$$Hi = \sum_{x,y} I\{ f(x,y) = i \}$$

(1)

Where \(i = 0,\ldots, n-1\) and \(n\) is the number of pixels of image output by the LBP operator.

Traditional selection for getting the localized frequency data is done through applying Gabor filters. But it has disadvantage of limited bandwidth so Log Gabor came into existence by Field from which it is possible to create filter with arbitrary bandwidth. Response of frequency for one dimensional Log Gabor filter is given in equation 2:

$$G(f) = \exp\left(\frac{-\left(\frac{\log(f/f_0)}{\sigma}\right)^2}{2\left(\frac{\log(f/f_0)}{\sigma}\right)^2}\right)$$

(2)

Where \(f_0\) and \(\sigma\) are the parameters of the filter.

Voice feature extraction is done by using linear predictor coefficients and mel-frequency Cepstral coefficients. Variance of prediction error is being returned by the LPC function of feature extraction based on the Levinson & Durbin algorithm. This method converts the analog voice signal into digital value in which a value is predicted by linear function of the past values of signal. The most common representation of linear prediction is given in equation 3.

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i)$$

(3)

Where \(\hat{x}(n)\) - predicted value, \(x(n-i)\) - previous observed values, and \(a_i\) the coefficients of prediction. The error generated by this estimate is \(e(n) = x(n) - \hat{x}(n)\) where \(x(n)\) - true value.

Mel-frequency Cepstral coefficients: MFCCs are nothing but collective coefficients that make up MFC. A type of Cepstral representation of the audio clip derives the MFCC. MFCC is a representation of the real Cepstral of a windowed short-time signal derived from the Fast Fourier Transform.

3.4. Template matching

After feature extraction, those features are stored in database during training phase this is nothing but knowledge base construction. In testing phase features are matched using Euclidian distance in our proposed system. There is use of K-nearest neighbour classifier for matching and for the recognition. KNN classifier is easy to interpret the output and it is instance based learning. This classifier is the
simplest among all classifiers. One nearest neighbour classifier is the most suitable classifier and its function is to assign point \( x \) to closest neighbour class in feature space. That is illustrated in equation 4:

\[
C_n^{h_n}(x) = Y_{(i)}
\]  

(4)

So in proposed system one nearest neighbour classifier is used for template matching. The proposed system better classifies using KNN classifier than others.

4. Experimental Results

The experimentation is carried out on two databases. First one is the synthetic database involving ten persons where six samples (face & voice) of each person is considered in which four are used for training and other two are used for testing. Second one is the own collected database with pose variation involving twenty persons. Here, also five samples of face & voice of individual is taken for experimentation in that three are used for training and two samples are used for testing. Performance evaluation of fusion techniques is carried out through performance measures such as FAR, FRR & RR. Where RR showed in equation 5: percentage of test sets that are correctly classified among the total number of recognition attempts.

\[
RR = \frac{\text{no. of correctly identified users}}{\text{total no. of recognition attempts}} * 100
\]  

(5)

FAR showed in equation 6: sometimes incorrect acceptance of an attempt by invalid user may also be possible. It is required to calculate rate of false acceptances ratio that is of wrongly identified users to total number of recognition attempts.

\[
FAR = \frac{\text{no. of wrongly identified users}}{\text{total no. of recognition attempts}} * 100
\]  

(6)

And FRR showed in equation 7: sometimes system may wrongly reject the access attempt of the genuine user hence it is required to calculate the rate of false rejections. Fraction of rejected genuine users to total number of recognition attempts.

\[
FRR = \frac{\text{no. of wrongly rejected users}}{\text{total no. of recognition attempts}} * 100
\]  

(7)

The results are tabulated in following tables from TABLE 1 to TABLE 2. Experimentation is carried out on all four levels using 10 people from synthetic DB and 20 people from own DB

| No of users | Fusion level | No of trained samples | No of tested samples | No of correctly identified samples | No of wrongly identified samples | RR in % | FRR in % | FAR in % |
|------------|--------------|-----------------------|---------------------|-----------------------------------|-------------------------------|--------|--------|--------|
| 10         | Feature      | 40                    | 20                  | 20                                | 00                            | 100    | 00     | 00     |
| 10         | Score        | 40                    | 20                  | 16                                | 04                            | 80     | 20     | 1.05   |
| 10         | Rank         | 40                    | 20                  | 20                                | 00                            | 100    | 00     | 00     |
| 10         | Decision     | 40                    | 20                  | 18                                | 02                            | 90     | 05     | 00     |

The graph of performance analysis on synthetic DB is shown in Fig. 4. In graph FFT means feature fusion technique, SFT means score fusion technique, RFT means rank fusion technique and DFT means decision fusion technique.
The graph of performance analysis on own DB is shown in Fig. 5. Abbreviations for FFT, SFT, RFT and DFT are considered as in previous graph.

As shown with results system works better on standard databases and achieves higher accuracy. Among all levels feature level provides accuracy of 100% and proved that fusion before matching achieves higher accuracy as compared to other fusion methods which take place after matching

5. Conclusion
Biometric system authorization is designed depend on two biometric traits using face & voice. Proposed system tests 50 users of own collected database involving pose variation and 10 users of AR database involving various expressions. Log Gabor and LBP features are extracted from face. MFCCs and LPC features are used during voice feature extraction. Proposed system implements four fusion levels. Feature fusion fuse features of face and voice using concatenation. Score level strategy implemented using maximum score of mode technique. Borda count technique that takes the sum of scores of
individual matchers is used for rank level fusion. For decision level logical conjunction (AND) is used. K-NN classifier is used for decision making. Experimental results prove that accuracy for feature level fusion is more as compared other fusion levels because fusion at early stage attains better recognition rate. Hence the proposed system is providing the results with better recognition rate. All fusion levels provide better recognition results but score level performs less as compared to other three as demonstrated in experimentation.

In future, it is planned to achieve higher accuracy rate with real time and other databases with more complexity samples.

Acknowledgement
I take this opportunity to express my gratitude to my colleague Ms. Madhura Shettar, my family, A & R Department and KLE Technological University for the valuable support.

References
[1] Madhura S, Hatture SM, 2017 A Model for Unconstrained Face Recognition System International Journal of Engineering Research & Technology 5 1-4
[2] Suneet NG, Vig R, Gupta S 2016 Analysis of Decision Level Fusion in Multimodal Biometrics using Iris and Fingerprint 3rd International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences 409-416
[3] Kirti VA, Kanade SG, Jadhav DV, Ajmera PK 2015 A Multi-modal and Multi-algorithmic Biometric System Combining Iris and Face International Conference on Information Processing Vishwakarma Institute of Technology, 496-501
[4] Sheetal C, Nath R 2015 A New Multimodal Biometric Recognition System Integrating Iris, Face and Voice International Journal of Advanced Research in Computer Science and Software Engineering, 5 145-150
[5] Fateme S, Nasri M 2015 A Multibiometric Finger Vein Verification System Based on Score Level Fusion Strategy Second International Congress on Technology, Communication and Knowledge, Mashhad Branch, Islamic Azad University, Mashhad, Iran, 501-507.
[6] Renu S, Das S, Joshi P 2015 Rank Level Fusion in Multibiometric Systems IEEE Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics 1-6
[7] Barbu T, Ciobanu A, Luca M 2015 Multimodal Biometric authentication based on Voice, Face and Iris. In 2015 E-Health and Bioengineering Conference (EHB) IASI, Romania 1-4
[8] Soruba Sree SR, Radha N 2014 A Survey on Fusion Techniques for Multimodal Biometric Identification International Journal of Innovative Research in Computer and Communication Engineering 2 7493-7497
[9] Fawaz A, Zahrani M, Alghamdi T 2013 Score-Level Fusion in Biometric Verification International Symposium on Biometrics and Security Technologies 193-197.
[10] Mahesh PK, Shanmukha Swamy MN 2012 Designing an Efficient Multimodal Biometric System using Palmprint and Speech Signal ACEEE Int. j. on Signal & Image Processing 3 76-80.
[11] David MD, Monica B 2012 A Data Fusion Technique Designed for Multimodal Biometric Systems 10th IEEE International Symposium on Electronics and Telecommunications, 155-158.
[12] Fouda YM 2012 Fusion of Face and Voice: An improvement International Journal of Computer Science and Network Security 12 37-43
[13] Samir A, Yahia B, Messaoud M and Chahir Y 2011 A Multi-Modal Recognition System Using Face and Speech, IJCSI International Journal of Computer Science Issues 8 230-236.
[14] Poornima B, Hatture SM 2017 A Fusion Model for Multimodal Biometric System, International Journal of Engineering Research & Technology (IJERT) 5.