Bimodal Biometric Person Authentication System Using Speech and Signature Features

Prof. M.N. Eshwarappa jenutc@rediffmail.com
Assistant professor Department of Telecommunication Engineering, Sri Siddhartha Institute of Technology, Tumkur-572101, Karnataka, India

Prof. (Dr.) Mrityunjaya V. Latte mvlatte@rediffmail.com
Principal and Professor Department of Electronics and Communication Engineering, JSS Academy of Technical Education, Bangalore-560060, Karnataka, India

Abstract

Biometrics offers greater security and convenience than traditional methods of person authentication. Multi biometrics has recently emerged as a means of more robust and efficient person authentication scheme. Exploiting information from multiple biometric features improves the performance and also robustness of person authentication. The objective of this paper is to develop a robust bimodal biometric person authentication system using speech and signature biometric features. Speaker based unimodal system is developed by extracting Mel Frequency Cepstral Coefficients (MFCC) and Wavelet Octave Coefficients of Residues (WOCOR) as feature vectors. The MFCCs and WOCORs from the training data are modeled using Vector Quantization (VQ) and Gaussian Mixture Modeling (GMM) techniques. Signature based unimodal system is developed by using Vertical Projection Profile (VPP), Horizontal Projection Profile (HPP) and Discrete Cosine Transform (DCT) as features. A bimodal biometric person authentication system is then built using these two unimodal systems. Experimental results show that the bimodal person authentication system provides higher performance compared with the unimodal systems. The bimodal system is finally evaluated for its robustness using the noisy data and also data collected from the real environments. The robustness of the bimodal system is more compared to the unimodal person authentication systems.

Keywords: Biometrics, Speaker recognition, Signature verification, Multimodal biometrics.

1. INTRODUCTION

Biometrics is the development of statistical and mathematical methods applicable to data analysis problems in the biological sciences. Introduction of this technology brings new security approaches to computer systems. Identification and verification are the two ways of using biometrics for person authentication. Biometrics refers to the use of physical or physiological, biological or behavioral characteristics to establish the identity of an individual. These

characteristics are unique to each individual and remain partially unaltered during the individual's life time [1]. Biometric security system becomes a powerful tool compared to electronics based security [2]. Any physiological and/or behavioral characteristic of human can be used as biometric feature, provided it possesses the following properties: universality, distinctiveness, permanence, collectability, circumvention, acceptability and performance [3]. The physiological biometrics related to the shape of the body. The oldest traits, that have been used for more than 100 years are fingerprints. Other examples are Face, Hand Geometry, Iris, DNA, Palm-prints and so on. Behavioral biometrics related to the behavior of a person. The first characteristic to be used, still widely used today, is the signature. Others are keystroke, Gait (way of walking), Handwriting and so on. Speech is the unique biometric feature that comes under both the categories [9]. Based on the application, selecting the right biometric is the crucial part. Unimodal biometric system, which operates using any single biometric characteristic, is affected by problems like noisy sensor data, non-universality and lack of individuality of the chosen biometric trait, absence of an invariant representation for the biometric trait. For instance, speech is a biometric feature whose characteristics will vary significantly if the person is affected by cold or in different emotional status. Some of these problems can be relived by using multimodal biometric system that consolidates evidence from multiple biometric sources. Multimodal or Multi-biometric systems utilize more than one physiological or behavioral biometrics for enrolment and identification. This work presents, such a multimodal biometric person recognition system and results obtained are compared to the unimodal biometric systems.

There are several multimodal biometric person authentication systems developed in the literature [3-7]. In 2004, A. K. Jain *et. al.*, proposed the frame work for multimodal biometric person authentication [3]. Even though some of the traits offering good performance in terms of reliability and accuracy, none of the biometrics is 100% accurate. With increasing global need for security, the demand for robust automatic person recognition systems is evident. For applications involving the flow of confidential information, the authentication accuracy of the system is always the prior concern. From this basic reason the use of multimodal biometrics are encouraged. Multi-biometrics is an integrated prototype system embedding different types of biometrics [35]. Multimodal biometric fusion and identity authentication technique help to achieve an increase in performance of identity authentication system [8]. Multimodal biometrics can reduce the probability of denial of access without sacrificing the False Acceptance Rate (FAR) performance by increasing by discrimination between the genuine and impostor classes. There are several multimodal biometric person authentication systems developed in the literature [4-8]. Applications of multi-biometrics are widely spread throughout the world. A wide variety of systems require reliable personal recognition schemes to either confirm or determine the identity of an individual requesting their services. The purpose of such schemes is to ensure that the rendered services are accessed only by a legitimate user, and not anyone else. Examples of such applications include secure access to buildings, computer systems, laptops, cellular phones and ATMs. In the absence of robust personal recognition schemes, these systems are vulnerable to the wiles of an impostor. Authentication systems built upon only one modality may not fulfill all the requirements, due to the limitations of unimodal systems. This has motivated the current interest in multimodal biometrics, in which several biometric traits are simultaneously used in order to make an identification decision. The objective of the present work is to develop a bimodal biometric system using speech and signature features to mitigate the effect of some of the limitations of unimodal biometric systems.

The present work mainly deals with the implementation of bimodal biometric system employing speech and signature as the biometric modalities. This includes feature extraction techniques and modeling techniques used in biometric system. The organization of the paper is as follows: Section 2 deals with bimodal databases used in bimodal person authentication system. Section 3 deals with unimodal biometric speech based person authentication system and section 4 deals with unimodal biometric signature based person authentication system. Bimodal biometric system by combining speaker and signature recognition systems is explained with different fusion techniques in section 5. Section 6 concludes the paper by summarizing the present work and adding few points regarding the future work.

2. BIMODAL DATABASES FOR PERSON AUTHENTICATION

IITG Speech database (standard)
Number of speakers: 30(20 male, 10 female)
Sampling frequency: 8000Hz
Sentences considered for each speaker: 4
Number of utterances of each sentence for each speaker: 24
Training session: first 16 utterances
Testing session: remaining 8 utterances of each sentence of each speaker

IITG Signature database (standard)
Number of writers: 30 (20 male, 10 female)
Scanner: HP Scan jet 5300C
Resolution: 300dpi (digits per inch)
Data storage: 8-bit Gray scale image
Saved format: bmp (bits mapping)
Number of sample signatures of each writer: 24
Training session: First 16 signatures of all the writers
Testing session: remaining 8 signatures of all the writers

SSIT Speech database
Number of speakers: 30(20 male, 10 female)
Sampling frequency: 8000Hz
Sentences considered for each speaker: 4
Number of utterances of each sentence for each speaker: 24
Training session: first 16 utterances
Testing session: remaining 8 utterances of each sentence of each speaker

SSIT Signature database
Number of writers: 30 (20 male, 10 female)
Scanner: HP Scan jet 5300C
Resolution: 300dpi (digits per inch)
Data storage: 8-bit Gray scale image
Saved format: bmp (bits mapping)
Number of sample signatures of each writer: 24
Training session: First 16 signatures of all the writers
Testing session: remaining 8 signatures of all the writers

3. UNIMODAL SPEECH BASED PERSON AUTHENTICATION SYSTEM

As any other pattern recognition systems, a speech based person authentication system also consists of three components: (1) Feature extraction, which transforms the speech waveform into a set of parameters carrying salient speaker information; (2) Pattern generation, which generates from the feature parameters a pattern representing the individual speaker: and (3) Pattern matching and classification, which compares the similarity between the extracted features and a pre-stored pattern or a number of pre-stored patterns, giving the speaker identity accordingly. There are two stages in a speaker recognition system, training and testing. In the training stage, speaker models (or patterns) are generated from the speech samples with some feature extraction and modeling techniques. In testing stage, feature vectors are generated from the speech signal with the same extraction procedure as in training. Then a classification decision is made with some matching technique. Person authentication is a binary classification task [22]. The features from the testing signal are compared with the claimed speaker pattern and a decision is made to accept or reject the claim [10]. Depending on the mode of operation, speaker recognition can be classified as text dependent recognition and text independent recognition. The text dependent recognition requires the speaker to produce speech for the same text, both during training and testing: whereas the text independent recognition does not on a specific text being

spoken [11]. The present work follows text dependent speaker recognition approach.This work uses feature extraction techniques based on (1) Mel Frequency Cepstral Coefficients (MFCC) derived from Cepstral analysis of the speech signal and (2) Wavelet Octave Coefficients of Residues (WOCOR) derived from the Linear Prediction (LP) residual. The time frequency analysis of the LP residual signal is performed to obtain WOCOR [14]. WOCOR are generated by applying a pitch-synchronous wavelet transform to the residual signal. Experimental results show that the WOCOR parameters provide complementary information to the conventional MFCC features for speaker recognition [14]. The Vector Quantization (VQ) and Gaussian Mixture Modeling (GMM) are used for modeling the person information from these MFCC and WOCOR features [13-15]. State of the art system uses MFCC derived from speech as feature vectors and GMM as the modeling technique [13].

Feature Extraction from Speech Information

The speaker information is present both in the vocal tract and excitation parameters [12]. The vocal tract system can be corresponds to processing of speech in short (10-30ms) overlapped (5-15ms) windows. The vocal tract system is assumed to be stationary within the window and it can be modeled as all-pole-filter using LP analysis [21]. The most used form of speech signal for feature extraction is the Cepstrum. Different forms of Cepstral representation include Complex Cepstral Coefficients (CCC), Real Cepstral Coefficients (RCC), Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstral Coeeficeints (LPCC). Among these the mostly used one includes MFCC. In all the Cepstral analysis techniques the vocal tract information is obtained by taking *log* over spectrum of the speech signal. The LP residual signal, though not giving the true glottal pulse, is regarded as a good representative of the excitation source. The Haar transform and Wavelet transform are applied for the multi-resolution analysis of the residual signal and the derived the feature vectors termed as Wavelet Octave Coefficients of Residues (WOCOR). WOCOR are believed to be effectively capturing the speaker specific spectro-temporal characteristics of the LP residual signal.

Extraction of MFCC Feature Vectors

The state of the system builds a unimodal system by analyzing speech in blocks of 10-30 ms with shift of half the block size. The MFCC are used as feature vectors extracted from each of the blocks. The MFCCs from the training or enrolment data are modeled using Vector Quantization (VQ) and Gaussian Mixture Modeling (GMM) technique [12]. The MFCCs from the testing or verification data are compared with respective model to validate the identity claim of the speaker. The MFCCs represent mainly the vocal tract aspect of speaker information and hence take care of only physiological aspect of speech biometric feature. Another important physiological aspect contributing significantly to speaker characteristics is the excitation source [13]. A speech signal is obtained by the convolution of vocal parameters v(n) and excitation parameters x(n) given by equation (3.1). We can not separate these parameters in time domain. Hence we go for Cepstral domain. The Cepstral analysis used for separating the vocal tract parameter v(n) and excitation parameters x(n), from speech signal s(n).

$$s(n) = v(n) * x(n) \quad (3.1)$$

The Cepstral analysis gives the fundamental property of convolution used for separating the vocal tract parameters and excitation parameters [27]. The Cepstral Coefficients (C) of length M can be obtained by using equation (3.2).

$$C = \text{real}\ (\text{IFFT}\ (\log |\text{FFT}\ (s(n))|)) \quad (3.2)$$

The nonlinear scale i.e., relation between the Mel frequency (f_{Mel}) and physical frequency (f_{Hz}) is used for extracting spectral information from the speech signal by using Cepstral analysis.

$$f_{Mel} = 2595 \log_{10} \left(1 + \frac{f_{Hz}}{700}\right) \quad (3.3)$$

Using equation (3.3) we construct a spectrum with critical bands which are overlapped triangular banks i.e., we map the linear spaced frequency spectrum (f_{Hz}) into nonlinearly spaced frequency spectrum (f_{Mel}). By this we can mimic the human auditory system and based on this concept MFCC feature vectors are derived. Windowing eliminates the Gibbs oscillations, which occur by truncating the speech signal. Using equation (3.4), Hamming window coefficients are generated, with which corresponding speech of frame is scaled.

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad (3.4)$$

But, due to Hamming windowing, samples present at the verge of window are weighted with lower values. In order to compensate this, we will try to overlap the frame by 50%. After windowing, we compute the log magnitude spectrum of each frame for finding the energy coefficients using equation (3.5).

$$Y(i) = \sum_{k=0}^{\frac{N}{2}} \log \mid S(k,m) \mid Hi\left(k\,\frac{2\pi}{N-1}\right) \quad (3.5)$$

where $H_i\left(k\,\frac{2\pi}{N}\right)$ is the i[th] Mel critical bank spectra and N is the number of points used to compute the discrete Fourier transform (DFT). The M number of Mel frequency coefficients computed by using discrete Cosine transforms (DCT), by using equation (3.6), which is nothing but the real IDFT of critical band filters log energy outputs.

$$C(n,m) = \left(\frac{2}{N}\right)\sum_{k=1}^{\frac{N}{2}-1} Y(k)\cos\left(k\,\frac{2\pi}{N}n\right) \quad (3.6)$$

Where n=1, 2, 3…….M.

The present work also takes care of channel mismatch by using the Cepstral Mean Subtraction (CMS) and the effect of different roll off from the different channels on Cepstral coefficients by Liftering procedure [21].

Extraction of WOCOR Feature Vectors

The Linear Predictive (LP) residual signal is adopted as a good representative of the vocal source excitation, in which the speaker specific information resides on both time and frequency domains. The resulting vocal source feature, WOCOR feature, can effectively extract the speaker-specific spectro-temporal characteristics of the LP residual signal. Particularly, with pitch-synchronous wavelet transform, the WOCOR feature set is capable of capturing the pitch-related low frequency properties. Only voiced speech is kept for subsequent processing. In the source-filter model, the excitation signal for unvoiced speech is approximated as a random noise [22, 26]. Voicing decision and pitch extraction are done by the robust algorithm for pitch tracking [32]. We believe that such a noise-like signal carries little speaker-specific information in the time-frequency domain [28]. For each voiced speech portion, a sequence of LP residual signals of 30ms long is obtained by inverse filtering the speech signal, i.e.,

$$e(n) = s(n) - \sum_{k=1}^{12} a_k s(n-k) \quad (3.7)$$

where the filter coefficients a_k are computed on Hamming windowed speech frames using the autocorrelation method [22]. The e(n)'s of neighboring frames are concatenated to get the residual signal, and their amplitude is normalized within [-1,1] to reduce intra-speaker variation. Once the pitch periods estimated, pitch pulses in the residual signal are located. For each pitch pulse, pitch-synchronous wavelet analysis is applied with a Hamming window of two pitch periods long. The windowed residual signal is denoted as $e_h(n)$. The wavelet transform of $e_h(n)$ is computed as

$$w(a,b) = \frac{1}{\sqrt{a}}\sum_{n} eh_{(n)}\psi^*\left(\frac{n-b}{a}\right) \quad (3.8)$$

Where a = {2^k|k=1, 2 …….K} and b = 1, 2 ……N, and N is the window length. $\Psi^*(n)$ is the conjugate of the fourth order Daubechies wavelet basis function $\Psi(n)$, a and b are the scaling parameter and the translation parameters, respectively [33].

The four octave groups of wavelet coefficients, i.e.,

$$W_k = \{w(2^k, b) | b = 1, 2……..N\}, \text{ where } k=1, 2, 3, 4. \quad (3.9)$$

Each octave group of coefficients is divided evenly into M subgroups, i.e.,

$$W_k^M(m) = \left\{ w(2^k, b) \mid b \in \left(\frac{(m-1)N}{M}, \frac{mN}{M} \right) \right\} \quad \text{m = 1, 2……..M} \tag{3.10}$$

The two-norm of each sub-group of coefficients is computed to be a feature parameter. As a result, the complete feature vector is composed as

$$WOCOR_M = \{ \| W_k^M(m) \| \quad \text{For m = 1, 2….M and k = 1, 2, 3, 4.} \} \tag{3.11}$$

where $\|.\|$ denotes the two norm operation.

For a given speech utterance, a sequence of $WOCOR_M$ feature vectors is obtained by pitch-synchronous analysis of the LP residual signal. Each feature vector consists of 4M components, which are expected to capture useful spectro-temporal characteristics of the residual signal.

Modeling Techniques

For speaker recognition, pattern generation is the process of generating speaker specific models with collected data in the training stage.The mostly used modeling techniques for modeling includes Vector Quantization (VQ) and Gaussian Mixture Modeling (GMM) [13-15]. The VQ modeling involves clustering the feature vectors into several clusters and representing each cluster by its centroid vector for all the feature comparisons. The GMM modeling involves clustering the feature vectors into several clusters and representing all these clusters using a weighted mixture of several Gaussians. The parameters that include mean, variance and weight associated with each Gaussian are stored as models for all future comparisons. For speaker recognition, the Gaussian Mixture Model (GMM) has been the most popular clustering technique. A GMM is similar to a VQ in that the mean of each Gaussian density can be regarded as a centroid among the codebook. However, unlike the VQ approach, which makes “hard” decision (only a single class is selected for feature vector) in pattern matching, the GMM makes a “soft” decision on mixture probability density function. This kind of soft decision is extremely useful for speech to cover the time variation.

Vector Quantization (VQ)

Once the MFCC feature vectors are computed for the entire frame of the speech signal for the individual speaker, we have to find the sequence of feature vectors of training speech signal which is the text dependent template model. The dynamic time warping finds the match between the template matching and, it is time consuming as the number of feature vectors increases. For this reason, it is common to reduce the number of training feature vectors by some modeling technique like clustering. The cluster centers are known as code vectors, and the set of code vectors is known as codebook. In this work, the Vector Quantization (VQ) method is used for pattern matching [14]. Vector quantization process is nothing but the idea of rounding towards the nearest integer i.e. Minimum Mean Square Error (MMSE). The two popular codebook generation algorithms namely k-means algorithm [15-16] and Linde-Buzo and Gray (LBG) algorithm [17] are used for generating speaker based vector quantization (VQ) codebooks for speaker verification.

Gaussian Mixture Modeling (GMM)

Generally, speaker models can be classified into two categories: the generative model and the discriminative model. Generative models attempt to capture all the underlying distribution, i.e., the class centroids and the variation around the centroids, of the training data. The most popular generative model in speaker recognition is the stochastic model, e.g., Gaussian Mixture Models (GMM), hidden Markov Model (HMM), etc. Discriminative models, on the other hand, not necessary model the whole distribution, but the most discriminative regions of the distribution. The template models, e.g., Vector quantization (VQ) codebooks, can also be regarded as a generative model, although it does not model the variations. Unlike the template models, the stochastic models aim at the distribution, i.e., the centroid (mean) and the scattering around the centroid (variance) as well, of feature vectors in a multi-dimensional space. The pattern matching can be formulated as measuring the probability density (or the likelihood) of an observation given the Gaussian. As for speaker recognition, the Gaussian Mixture Model (GMM) has been the most popular clustering technique. The likelihood of an input feature vectors given by a specific GMM is the weighted sum over the likelihoods of the M unimodal Gaussian densities [29], which is given by equation (3.12).

$$P(x_i \mid \lambda) = \sum_{j=1}^{M} w_j b\left(x_i \mid \lambda_j\right) \tag{3.12}$$

where $b(x_i|\lambda_j)$ is the likelihood of x_i given the j^{th} Gaussian mixture

$$b(x_i \mid \lambda_j) = \frac{1}{(2\pi)^{D/2} \mid \sum_j \mid} \exp\{-\frac{1}{2}\left(x_i - \mu_j\right)^T \sum_j{}^{-1} \left(x_i - \mu_j\right)\} \tag{3.13}$$

Where D is the vector dimension, μ_j and $\sum_j$ are the mean vectors and covariance matrices of the training vectors. The mixture weights w_j are constrained to be positive and sum to one. The parameters of a GMM, μ_j, $\sum_j$ and w_j can be estimated from the training feature vectors using the maximum likelihood criterion, via the iterative Expectation-Maximization (EM) algorithm [31]. A GMM can be regarded as providing an implicit segmentation of the sound units without labeling the sound classes. The sound ensemble is classified into acoustic classes, each of which represents some speaker-dependent vocal system configurations, and modeled by a couple of Gaussian mixtures.

Performance of Speaker Recognition System

In the recognition stage, feature vectors are generated from the input speech sample with same extraction procedure as in training. Pattern matching is the task of calculating the matching scores between the input feature vectors and the given models in recognition. The input features are compared with the claimed speaker pattern and a decision is made to accept or reject the claiming. The performance of a system operating in verification mode is specified in terms of two error rates. They are false acceptance rate (FAR) and false rejection rate (FRR). The FAR may be defined as the probability of an impostor being accepted as a genuine individual and FRR may be defined as the probability of a genuine individual being rejected as an imposter. In pattern matching, the training speech of each speaker is processed in blocks of 20ms and 10ms block shift to extract MFCC and WOCOR features. These features are modeled using VQ and GMM modeling techniques. In this way, speaker models are developed. We will have in total four models per speaker. These include VQ-MFCC, GMM-MFCC, VQ-WOCOR and GMM-WOCOR combinations. The testing speech is also processed in a similar way and matched with the speaker models using Euclidean distance in case of VQ and Likelihood ratio in case of GMM. Testing stage in the person authentication system includes matching and decision logic. During testing the test feature vectors are compared with the reference models. Hence matching gives a score which represents how well the feature vectors are close to the claimed model. Decision will be taken on the basis of matching score, which depends on the threshold value. The alternative is to employ verification through identification scheme. In this scheme the claimed identity model should give the best match. The test speech compared with the claimed identity model, if it gives best match, then it is accepted as genuine speaker, otherwise, rejected as imposter.

In order to check the performance of different algorithms, we use IITG standard speech database. The speaker verification system is implemented with different combination of feature extraction techniques and modeling techniques. As a result, the four unimodal biometric systems were developed individually and conducted experiments with 30 user's database. The performance of different unimodal systems, with and without noise (noise with SNR=15dB), and also there combination systems using some simple rules of combination like score level fusion are tabulated in Table1.

Table1: Speaker system verification performance(IITG database)

Unimodal System	FAR (%)	FRR (%)	Average error (%)
MFCC-VQ (clean data)	0.001	0.003	0.002
MFCC-VQ (noise data)	0.3862	12.1667	6.2765
MFCC-GMM (clean data)	0	0	0.000
MFCC-GMM (noise data)	2.4623	20.133	11.2976
WOCOR-VQ (clean data)	0.11	1.232	0.671
WOCOR-VQ (noise data)	1.242	28.1267	14.6844
WOCOR-GMM (clean data)	0	0	0.000
WOCOR-GMM (noise data)	0.0123	3.222	1.6725

The experiments are also conducted for SSIT database, which is our own database created under practical environments for 30 users. The experimental setup was same as that of IITG database. The experimental results are shown in Table 2. The Table 2 resembles the Table 1, which shows that the proposed techniques yield good performance irrespective of any database. The result comparative evaluation process is done with available literatures and the values listed out within bracket in Table 2, the average error obtained for clean data. The experimental result shows the present system performance is better.

Table 2: Speaker system verification performance(SSIT database)

Unimodal System	FAR (%)	FRR (%)	Average error (%)
MFCC-VQ (clean data)	0	0	0.00 (0.1)
MFCC-VQ (noise data)	0.4885	14.1667	7.3276
MFCC-GMM (clean data)	0	0	0.0 (0.01)
MFCC-GMM (noise data)	3.4483	25.133	14.29065
WOCOR-VQ (clean data)	0.22	1.624	0.922 (1.06)
WOCOR-VQ (noise data)	2.0402	29.1667	15.60345
WOCOR-GMM (clean data)	0	0	0.0 (0.0)
WOCOR-GMM (noise data)	0.0144	3.333	1.7241

4. UNIMODAL SIGNATURE BASED PERSON AUTHENTICATION SYSTEM

Unimodal signature based person authentication system is more commonly termed as signature verification system. Signature verification is the task of verifying signatories by using their signatures [18]. Signature verification systems require contact with the writing instrument and an effort on the part of the user. The signature verification system finds use in government, legal and commercial applications. Signature is a behavioral biometric which is characterized by a behavioral trait. Signature, which is similar to handwriting, is learnt and acquired over a period of time rather than a physiological characteristic. Signature verification methods are divided into two types, offline signature verification and online signature verification. Online signature verification uses additional information collected dynamically at the time of signature acquisition along with the signature information and is also called as dynamic signature verification [19]. Offline signature verification uses only the scanned signature image for verification which is static and is also called static signature verification. The offline signature signal is two-dimensional nature and offline signature recognition becomes a pattern recognition problem. The techniques used in the literature for offline signature recognition are Support Vector Machines (SVM), Hidden Markov Models (HMM), Neural Networks, Graph Matching, GSC features (gradient, structure and concavity) and Dynamic Time Warping [19]. The present work employs offline signature verification system for person authentication.

Feature Extraction from Signature Information

Feature extraction plays a very important role in offline signature verification. In offline signature recognition there are two groups of features, static and pseudo dynamic features fall under one group, global and local features constitute the other group. In our work we implemented an offline signature identification system using Vertical Projection Profile (VPP), Horizontal Projection Profile (HPP) and Discrete Cosine Transform (DCT) features [23]. The VPP and HPP are static features of a signature and DCT is a global feature of a signature image. The size of VPP is equal to the number of columns in the signature image. VPP also a kind of histogram indicates the intensities around which the image pixels are concentrated. VPP gives the horizontal starting and ending points of the image. So, this can be used as a unique feature of a signatory. Since, the size of signature regions are not constant even for a single user, in this work we are taking average value of vertical projection profile as a feature. Horizontal Projection Profile (HPP) is an array contains sum of pixels of each row in a signature image. The size of HPP is equal to the number of rows in the signature image. HPP is also a kind of histogram. Just like histogram indicates the intensities around which the image pixels are concentrated. HPP gives vertical starting and ending points of the image. So, this can be used as a unique feature of a signatory.

Since, the size of signature regions are not constant even for a single user, in this work we are taking the average value of horizontal projection profile as a feature. The equations (4.1) and (4.2) give just average values of VPP and HPP of signature image.

$$vpp_{avg} = \frac{1}{N}\sum_{q=1}^{N}\sum_{p=1}^{M} A_{(p,q)} \tag{4.1}$$

$$hpp_{avg} = \frac{1}{M}\sum_{p=1}^{M}\sum_{q=1}^{N} A_{(p,q)} \tag{4.2}$$

where M is number of rows in an image, N is number of columns in an image, p and q are the row and column indices respectively, and A(p,q) is the intensity of the signature image at p^{th} row and q^{th} column. There are various transforms available to extract the feature of images. Among them, Karhunen-Loeve (KL), Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT) are the transforms. Since DCT is real arithmetic and having less computational complexity to other transforms, in our work we are using analysis to extract the features of our signatures. Equation (4.3) shows the two dimensional discrete cosine transform of the input image A. Where B_{pq} is the output DCT coefficient corresponding to p^{th} row and q^{th} column. M and N are total number of rows and columns of input image respectively.

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} A_{mn} \cos\left(\frac{\pi(2m+1)p}{2M}\right)\cos\left(\frac{\pi(2n+1)q}{2N}\right) \tag{4.3}$$

Where $\alpha_p = \left\langle \frac{1}{\sqrt{M}}$ for p=0 and $\alpha_p = \left\langle \sqrt{\frac{2}{M}}$ for 1 ≤ p ≤ (M-1)

$\alpha_q = \left\langle \frac{1}{\sqrt{N}}$ for q=0 and $\alpha_q = \left\langle \sqrt{\frac{2}{N}}$ for $1 \leq q \leq (N-1)$

The performance of the signature recognition system depends on the way in which the DCT coefficients are considered. As DCT is a transform which has high energy compaction property, most of the energy in the signature image is concentrated in very few coefficients. In threshold coding, the DCT coefficients in the transformed image have been sorted and a particular number of DCT coefficients have been taken as a feature vector representing the signature image. Instead of threshold coding, the zonal coding DCT coefficients are used for the better performance, which gives energy concentration at low spatial frequencies. In this work we are considering zonal coding of DCT coefficients of signature image.

Modeling Techniques

During training, calculate two-dimensional DCT of all the training images, and consider a specified number of coefficients according to zonal coding. Calculate the average value of all the average VPP values, average HPP values and all DCT coefficient vectors of all the signatures of a user. These become three kinds of feature models for signature recognition system. The simple time averaged VPP and HPP values cannot convey the person information present along its length. A modified system uses VPP and HPP vectors with Dynamic Time Warping (DTW) for the optimal cost. DTW is a pattern matching technique which aims at finding the minimum cost path between the two sequences having different lengths [23]. A very general approach to find distance between two time series of different sizes is to resample one of the sequence and comparing the sample by sample. The drawback of this method is that, there is a chance of comparing the samples that might not correspond well. This means that comparison of two signals correspond well when there is a matching between troughs and crests. DTW solves this method by considering the samples with optimum alignment. The DTW computation starts with the warping of the time indices of two sequences. The two sequences are compared with some distance measures like Euclidean distance at each and every point so as to obtain the Distance Matrix. These distances in the matrix are termed as local distances. Let the matrix be *d* and the sequences are A, B with lengths M, N respectively. Then *d* is calculated as:

$$d(i,j) = distance(A(i),B(j)) \tag{4.4}$$

where i vary from 1 to M and j varies from 1 to N.

The distance here considered is Euclidean distance. After computing this matrix, the minimum path is obtained from the matrix by considering some constraints. Apart from the direct Gray scale values in terms of VPP and HPP, some frequency information is obtained from DCT of the

image. So, in order to use this information, zonal coding of DCT coefficients of signature image is considered. The modified feature vectors obtained from the signature image A(i,j) of size MXN are given in equations (4.5) and (4.6). Apart from the VPP and HPP vectors, DCT features with zonal masking are used as it is.

$$vpp(j) = \sum_{i=1}^{M} A(i, j) \quad \text{Where j= 1, 2, 3……..……N} \tag{4.5}$$

$$hpp(i) = \sum_{j=1}^{N} A(i, j) \quad \text{Where i=1, 2, 3……..M} \tag{4.6}$$

The VPP, HPP and DCT features gives three models namely, VPP-HPP model, DCT model and VPP-HPP-DCT model for training the signatures. The third model should give better performance compared to the other two, since all three different information about signature is used while modeling. In this work, we propose the new method using VPP-HPP features with DTW method, for signature verification, instead of simple averaged values of VPP, HPP features.

Performance of Signature Recognition System

Signature verification is a pattern recognition problem [33]. After extraction of the features from a given image, distances are obtained from a testing image to all the users. Signature verification using VPP-HPP and DCT features involves the following steps:

a) Calculate the DTW distance values separately for VPP vectors and HPP vectors from all the users for all the training images to the testing image and obtain distances from each user using average distance method.

b) Obtain the two dimensional DCT and zonal coding of the coefficients for the testing image. Calculate the Euclidean distance of the signature feature vectors from the corresponding trained image DCT models. We get one distance for each model and for each user in the database.

c) Normalize each of the distance of a particular feature using one of the normalization methods and use sum rule for fusion of match scores obtained using each model.

d) Assign the test signature to the user who produces least distance in fused sum vector.

In order to check the performance of different algorithms, we use IITG standard signature database. Then, implemented signature verification system with different combination of feature extraction techniques and modeling techniques. The two unimodal biometric systems were developed individually and conducted experiments with 30 user's database. The results are tabulated for averaged values of VPP-HPP-DCT feature models and modified VPP-HPP-DCT feature models with DTW. Table 3 shows the performance of two different signature verification systems for with and without noise (salt and pepper noise=3%).

Table3: Signature system verification performance(IITG database)

Unimodal System	FAR (%)	FRR (%)	Average error (%)
VPP-HPP-DCT (clean data)	1.2232	36.23	18.7256
VPP-HPP-DCT (noise data)	2..342	70.161	36.251
Modified VPP-HPP-DCT (clean data)	0.061	3.212	1.6302
Modified VPP-HPP-DCT (noise data)	2.1332	66.426	34.2796

The experiments are also conducted for SSIT signature database. The experimental setup was same as that of IITG database. The results are shown in Table 4. The performance in Table 4 resembles the performance in Table 3. This means that the proposed technique gives good performance irrespective of any database. The result comparative evaluation process is done with available literatures and the values are listed out within the bracket in Table 4. The present system performed better with clean data.

Table4: Signature system verification performance(SSIT database)			
Unimodal System	**FAR (%)**	**FRR (%)**	**Average error (%)**
VPP-HPP-DCT (clean data)	1.2931	37.5	19.396 (20.013)
VPP-HPP-DCT (noise data)	2..4549	71.25	36.8534
Modified VPP-HPP-DCT (clean data)	0.1149	3.333	1.7241 (1.66)
Modified VPP-HPP-DCT (noise data)	2.3994	69.583	35.994

5. BIMODAL PERSON AUTHENTICATION SYSTEM

The main module in the bimodal person authentication system is the biometrics. One commonly used approach to the development of biometrics block is combining person information from different biometric features. There are different ways of combining biometric features like decision level fusion, score level fusion, feature level fusion etc. The present work employs score level fusion for the development of bimodal biometric person authentication system. Once we have the biometric block obtained by the fusion process, the person authentication performance will increase and also its robustness.

In the score level fusion, scores obtained at the output of the classifier are fused using some rules. The simple rules of fusion are Sum rule, Product rule, Min rule, Max rule and Median rule. The Sum rule and Product rule assume the statistical independence of scores from the different representations [24-25]. In the present case, the entire work is carried out using Sum rule. The outputs of the individual matchers need not be on the same numerical scale. Due to these reasons, score normalization is essential to transform the scores of the individual matchers into a common domain prior to combining them. Score normalization is a critical part in the design of a combination scheme for matching score level fusion. Min-Max and Z-score normalization are the most popular techniques used for normalization [25]. Unimodal biometric person authentication systems are initially developed by using speech and signature biometrics features. The scores from the unimodal systems are normalized and combined. The combined score are treated as the output of bimodal biometric person authentication system. Therefore the combined score is evaluated to obtain the performance of bimodal person authentication system.

Performance of Signature Recognition System

Table 5 shows the performance of the bimodal biometric person authentication systems using speech and signature information. These include (i) MFCC features with VQ model and GMM model for speech with VPP-HPP and DCT features for signatures (ii) WOCOR features with VQ model and GMM model for speech with VPP-HPP and DCT features for signature (iii) MFCC features with VQ model and GMM model for speech with modified VPP-HPP and DCT features for signature (iv) WOCOR features with VQ model and GMM model for speech and modified VPP-HPP and DCT features for signature, are tabulated separately. We have conducted experiments on bimodal biometric system with and without noise. The random noise (SNR=15dB) added to the speech files under testing in the speaker recognition case. Similarly in the signature recognition case, we added salt and pepper noise (3%) to the signature files under testing. The IITG standard database and SSIT database are used for checking the performance of bimodal system. As it can be observed, the performance of bimodal system is better in all the cases. This demonstrates the usefulness of using bimodal and hence multimodal biometric features for person authentication.

Table5: Bimodal system verification performance(IITG database)			
Bimodal System	**FAR (%)**	**FRR (%)**	**Average error (%)**
MFCC-VQ for speech and VPP-HPP-DCT for signature (clean data)	0	0	0.00
MFCC-VQ for speech and VPP-HPP-DCT for signature (noise data)	1	70	35.5
MFCC-GMM for speech and VPP-HPP-DCT for signature (clean data)	0	0	0.00
MFCC-GMM for speech and VPP-HPP-DCT for signature (noise data)	2	70	36
WOCOR-VQ for speech and VPP-HPP-DCT with DTW for signature (clean data)	0	0	0.00
WOCOR-VQ for speech and VPP-HPP-DCT with DTW for signature (noise data)	1.86	57.75	30.72
WOCOR-GMM for speech and VPP-HPP-DCT with DTW for signature(clean data)	0	0	0.00
WOCOR-GMM for speech and VPP-HPP-DCT with DTW for signature(noise data)	0.86	54.75	27.8

Table 6: Bimodal system verification performance(SSIT database)			
Bimodal System	**FAR (%)**	**FRR (%)**	**Average error (%)**
MFCC-VQ for speech and VPP-HPP-DCT for signature (clean data)	0	0	0.00
MFCC-VQ for speech and VPP-HPP-DCT for signature (noise data)	2.42	70	36.2
MFCC-GMM for speech and VPP-HPP-DCT for signature (clean data)	0	0	0.00
MFCC-GMM for speech and VPP-HPP-DCT for signature (noise data)	2.4	69.58	36
WOCOR-VQ for speech and VPP-HPP-DCT with DTW for signature (clean data)	0	0	0.00
WOCOR-VQ for speech and VPP-HPP-DCT with DTW for signature (noise data)	2.86	58.75	30.72
WOCOR-GMM for speech and VPP-HPP-DCT with DTW for signature(clean data)	0	0	0.00
WOCOR-GMM for speech and VPP-HPP-DCT with DTW for signature(noise data)	1.86	53.75	27.8

6. CONSLUSION & FUTURE WORK

In this work, we have implemented a bimodal biometric person authentication using Speech and Signature biometric traits. The better performance can be achieved with different features and with different modeling techniques. The MFCC features with VQ model or GMM model and the WOCOR features with GMM model are best system for speaker verification. For the signature verification, the VPP-HPP with DTW method based system gives better performance. Thus, the experimental results proved that, the bimodal biometric person authentication system with respect to more number of users, and more number of biometrics. The future work needs to be done with respect to more number of users, and more number of biometrics. The future work may also be including with different sessions for speech data collection and signature data collection in practical environments. The new speaker recognition methods may be developed to extract feature vectors by combining two features like WOCOR and MFCC, different windowing techniques like triangular or rectangular or hamming used for framing in a linear frequency scale. The new signature verification system may be developed with the modifications were made to the basic DTW algorithm to account for stability of various components of a signature. Finally, the modified bimodal system performed significantly better than the basic system.

7. REFERENCES

1. A. Jain L.Hang and S. Pankanti. *"Can multi-biometrics improve performance,"* Proceedings of Auto ID, 59-64, 1999.
2. L. Gorman, "*Comparing passwords, tokens, and biometrics for user authentication,"* IEEE Proceedings, vol 91, no12, Dec 2003.
3. A.K. Jain, A. Ross and Prabhaker, "An *introduction to Biometric Recognition,*" IEEE Transaction on Circuits and Systems for Video Technology, vol 14, no.1, 4-20, Jan 2004.
4. R. Bruneeli and D.Falavigna, *"Person identification using multiple cues,"* IEEE Transaction, PAMI, vol 12, no10, 955-966, Oct.1995.
5. A. K. Jain and L. Hong, "*Integrating faces and fingerprints for person identification,"* IEEE Transaction, Pattern Analysis and Machine Intelligence (PAMI), vol.20, no12, 1295-1307, Dec 1998.
6. V. Ghatis, A.G. Bors and I.Pitas, "*Multimodal decision level fusion for person authentication,"* IEEE Transaction and Systems, Man and Cybernetics, vol 29, no6, 674-680, Nov. 1999.
7. R.W. Frischolz and U Dieckman, "*Biod: a multimodal biometric identification system,"* IEEE Computer, vol.33, 64-68, Feb 2000.
8. B. Duc et. Al., "*Fusion of audio and video information for multimodal person authentication,"* Pattern Recognition letters, vol.18, 835-843, 1997.
9. B.S. Atal, "*Automatic recognition of speakers from their voices,"* IEEE Proceedings, vol. 64, no. 4, 460-75, Apr 1976.
10. A.E. Rosenberg, "*Automatic Speaker verification: A review,"* IEEE Proceedings, vol 64, no.4, 475-487, Apr. 1976.
11. H. Gish, and M. Schmidt, "*Text-independent speaker identification,"* IEEE Signal Process, Magazine, vol 18, 18-32, Oct. 2002.
12. A. Eriksson and P.Wretling, "How flexible is the Human Voice? A case study of Mimicry*,"* Proceedings of European Conference on Speech Technology, Rhodes,1043-1046, 1997.
13. D.A. Reynolds, "*Speaker identification and verification using Gaussian mixture speaker models," Speech Communication.,* vol 17, no1-2, 91-108, 1995.
14. S.R.M. Prasanna, C.S. Gupta, and B. Yegnanarayana, "*Extraction of speaker specific excitation information from linear prediction residual of speech,"* Speech Communication., vol 48, 1243-1261, Oct. 2006.
15. L. hanzo, F.C.A. Somerville and J.P. Woodard," *Voice compression and Communications,"* John B. Anderson, Wiley IEEE Pres series, 2001.
16. Y. Linde, A.Buzo and R.M. Gray, "*An algorithm for Vector Quantizer Design,"* IEEE Transaction on Communications, vol, COM_28, no.1, 84-96,Jan 1980.
17. R. Gray, *"Vector quantization,"* IEEE Acoustic Speech Signal Process, Magazine, vol 1, 4-29, Apr.1984.
18. V.S. Nalwa, *"Automatic on-line Signature verification,"* IEEE Proceedings, vol 85, no.2, 213-239, Feb.1997.
19. W. Hou, X. Ye and K. Wang, *"A survey of offline signature verification,"* IEEE Proceedings, International Conference on Intelligent Mechatronics and Automation, 536-541, Aug 2004.
20. Chaur-Heh Hsieh, *"DCT based code book design for vector quantization of images,"* IEEE Transactions, Circuits and Systems for Video Technology, vol.2, no.4, 401-409, Dec1992.
21. Makhoul J., *"Linear Prediction: a Tutorial review",* IEEE Proceedings, 561-580, Oct 1975.
22. L. Rabiner and B.H. Jung, *"Fundamentals of Speech Recognition",* Pearson Education, 326-396(1993).
23. T.M. Math and R. Manmatha, *"Word image matching using dynamic time warping",* IEEE Proceedings, computer Vision and Pattern Recognition, vol.2, 521-527, June 2003.
24. A. Jain, K. Nanda Kumar and A.Ross, *"Score normalization in multimodal biometric systems",* Pattern Recognition Journal- Elsevier, vol.38, 2270-2285, Jan. 2005.
25. F.Alsaade, *"Score-Level fusion for multibiometrics",* PhD Thesis, University of Hertfordshire, Jan.2008.
26. B.S.Atal *"Effectiveness of Linear prediction characteristics of the speech wave for Automatic Speaker Identification and Verification",* J. Acoust, Soc. Amer., 55(6); 1304-1312,1974.

27. S.Furu, *"Cepstral Analysis Technique for Automatic Speaker Verification",* IEEE Transaction, Acoustic and Speech Signal Processing, ASSp-29(2): 254-272, 1981.
28. G.Strang and T.Nguyen, *"Wavelets and Filter Banks",* Wellesley-Cambridge Press, 1996.
29. A. Sanker and C.H.Lee, *"A Maximum-Likelihood approach to stochastic matching for robust speech recognition",* IEEE Transaction, Speech-Audio Processing, 4(3): 190-202, 1996.
30. L.R. Rabiner, *"A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition",* IEEE Proceedings, 77/27, 257-286,1989.
31. L.E. Baum and T. Petie, *"Statistical inference for probabilistic functions of finite state Markov chains",* Ann. Mat. Stat., 37; 1554-1563, 1966.
32. D.Talkin, *"A Robust Algorithm for Pitch Tracking (RAPT)",* Speech Coding and Synthesis, W.B. Kleja and K.K. Paliwal, Eds., New York, Elsevier 1995.
33. I. Daubechies, *"Ten Lectures on Wavelets",* Philadelphia, PA: Siam, vol.6, 36-106,1992.
34. ZHENG Nengheng, *"Speaker Recognition using Complementary Information from Vocal Source and Vocal Tract",* PhD Thesis, The Chinese University of Hong Kong, Nov. 2005.
35. A.Ross and A.k.Jain, *"Multimodal Biometrics: an Overview",* Proceedings of 12th European Signal Conference (EUSIPCO), 1221-1224, Sept.2004.