International Journal of Engineering & Technology, 14 (1) (2025) 1-6

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A novel multi-modal biometric authentication system with enhanced feature extraction using advanced mapped real transform

Lakshmi. S. Panicker [1] *, R. Gopikakumari [1]

[1] School of Engineering, CUSAT, Kochi, Kerala, India
**Corresponding author E-mail: lakshmispanicker@cusat.ac.in*

Abstract

In contexts where security is a top priority, biometric authentication, which is the act of identifying a person using physiological or behavioral modalities—acquires critical importance. Facial image and speech are the two most popular biometrics for personal verification. This paper introduces a new audio-visual biometric recognition system for authenticating humans, using the sequency-mapped real transform for feature extraction. The performance of this new system is compared with previously published results of other systems that only use audio, video, or both. The proposed system demonstrates better performance metrics by utilizing the complementary strengths of both modalities and by employing a new transform that cuts down on computational complexity.

Keywords: *Audio-Visual Biometrics; Feature Level Fusion; SMRT.*

1. Introduction

The act of confirming an individual's identity by using distinct biological characteristics, like fingerprints, facial features, iris patterns, voice, or behavioral traits such as typing patterns, is termed biometric authentication [1 - 3]. Biometric authentication systems are indispensable in various industries where security is critical, as traditional authentication techniques are more prone to hacking, phishing, and identity theft. Unimodal biometric systems that rely on a single modality for authentication are used in most real-world applications. However, such systems occasionally fail to recognize a person correctly because of a variety of problems, such as noisy data, intra-class variations, inter-class similarities, non-universality, spoofing, etc. Therefore, the desired characteristics of universality, distinctiveness, permanence, and collectability cannot be achieved by a unimodal biometric system. Multimodal systems, which rely on multiple biometrics for precise authentication, can readily overcome the shortcomings of unimodal systems. A review of the various systems and architectures for multimodal biometrics is presented in [4].

The most important stages in implementing a multimodal biometric system are feature extraction and fusion [5]. The relevant and compact features are to be extracted from multiple modalities, ensuring that the extracted features contain the information required to solve the underlying problem. Several feature extraction tools like DCT, Gabor filter, Mel-Frequency Cepstral Coefficients (MFCCs), and wavelets have been presented in the literature. The present research work aims to improve the performance of a multimodal biometric system with improved feature extraction tools. The sequency-mapped real transform is a transform that has shown excellent classification of audio and image signals [6], [7]. This research work highlights the relevance of sequency-mapped real transform as a feature extraction tool in multimodal applications and evaluates the resultant improvement in the performance of multimodal biometric systems.

While developing a multimodal system, one of the most challenging decisions is to zero in on a suitable fusion strategy for integrating information from different modalities. Fusion can be done at different levels that are categorized into sensor level, feature level, matching score level, and decision level [8]. Feature-level fusion is the process of concatenating the feature vectors extracted from different biometric modalities in order to create a new, more powerful feature vector with a higher dimensionality that represents the individual more accurately and provides the information required for the decision on identification.

The most commonly used biometric measure for authentication purposes is the facial image, as it provides the most natural, user-friendly, and non-invasive means to make a decision. Similarly, voice biometrics are considered the most efficient measures for authentication as they require no sophisticated sensors. Different architectures have been presented over the years to develop commercially successful speaker verification [9] and face verification systems [10], [11]. However, their performance could degrade dramatically under more challenging conditions. Hence, audio and visual modalities can be combined to develop a robust authentication system.

This paper presents a robust audio-visual biometric recognition system that uses the sequency-mapped real transform (SMRT) for feature extraction from audio and video modalities. The extracted feature vectors are combined using feature-level fusion, as it can model the correlated information between voice and face features. Unlike the previous work in the literature, the system is tested on video files of the training sample rather than a static facial image. Thus we have used the VidTIMIT database for testing, as it contains video and audio files

of the training sample thereby making it suitable for real-life implementation. The main contribution of this research work is the introduction of a new feature extraction tool that employs algorithms that avoid complex calculations, thus improving computational efficiency.
The remainder of the paper is organized as follows:
Section 2 presents the methodology used in the proposed system.
Performance metrics and results are presented in sections 3 and 4, respectively, followed by a conclusion in section 5.

2. Methodology

The flow diagram of the basic methodology employed is shown in Fig. 1. A description of the database, followed by a detailed discussion of feature extraction techniques and classifiers employed, is presented in this section.

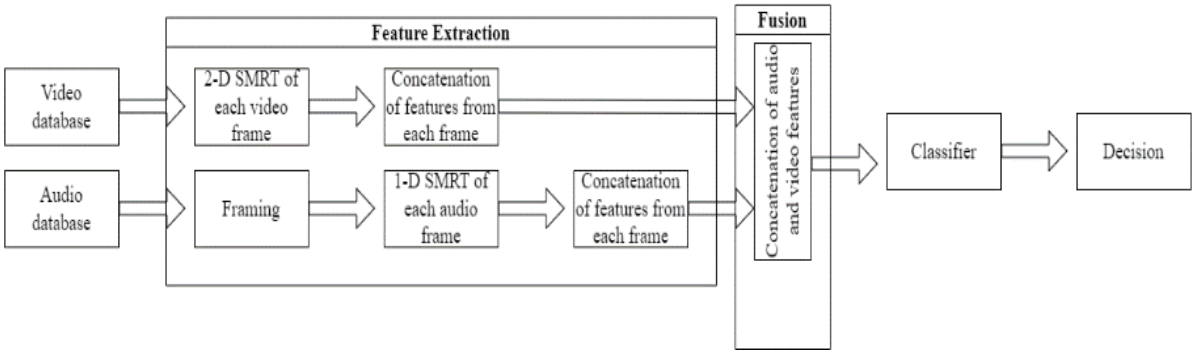

Fig. 1: Flow Diagram of Basic Methodology.

2.1. Database

The VidTIMIT database [12] is an audio-visual database comprised of audio-visual recordings of 43 people reciting sentences from the test section of the TIMIT corpus. It was recorded in three sessions. There are ten sentences per person, six of them belonging to session 1 and two each to sessions 2 and 3. Two sentences are common to all speakers, while the other eight sentences are generally different for each speaker, facilitating text-independent speaker recognition research. The video of each person is stored as a sequence of JPEG images with a resolution of 512×384 pixels.

2.2. Feature extraction

2.2.1. Sequency-mapped real transform

Transforms are the tools in signal processing that facilitate the analysis of a signal by mapping it into an alternative domain so as to extract hidden information from it. The Discrete Fourier Transform (DFT) is a powerful analytical tool, and a variety of approaches are utilized to minimize its computational complexity. The Fast Fourier Transform (FFT) is the most common algorithm for DFT implementation and it is extremely effective for 1-D signals. However, FFT is not efficient in the 2-D case, as the data is converted to a complex form, which results in increased computation time and storage requirements. Four real multiplications and two real additions are involved in a complex multiplication, and two memory locations are required to store one set of complex data. Finally, computation time would also be higher for multiplication than for addition. Therefore, processing speed can be increased by decreasing the number of complex multiplications.
A modified 2-D DFT representation [13] was developed in terms of real additions by expressing $N \times N$ DFT in terms of 2×2 DFT and utilizing the periodicity and symmetry properties of the twiddle factor. In this scheme, a signal is represented in terms of signal components that are mapped onto the twiddle factor axes and grouped according to the phase of the twiddle factor in terms of real arithmetic. This led to a new method of signal representation called Mapped Real Transform (MRT) [14] [15]. MRT coefficients $Y^{p}_{k_1,k_2}$ of a 2-D signal $x(n_1, n_2)$, $0 \leq n_1, n_2 \leq N-1$ is given by

$$Y^{p}_{k_1,k_2} = \sum_{\forall(n_1,n_2)\Rightarrow z=p} x(n_1, n_2) - \sum_{\forall(n_1,n_2)\Rightarrow z=p+M} x(n_1, n_2) \quad (1)$$

Where

$$0 \leq k_1, k_2 \leq N-1,\ z = \left((n_1k_1 + n_2k_2)\right)_N$$

However, MRT is a highly redundant and expansive transform as it maps a 2-D signal of size $N \times N$ into M matrices each of size $N \times N$ where N is an even integer and $M = N/2$. Hence Unique MRT (UMRT) [16] was developed to remove redundant coefficients and arrange the N^2 unique coefficients in an $N \times N$ matrix. This arrangement of the unique coefficients resulted in the scattered placement of phase components associated with a specific frequency and posed issues in many applications. Visual representation of the unique MRT coefficients show unique patterns of sign changes. They were arranged in the order of row-wise and column- wise sequencies to derive the sequency-mapped real transform (SMRT) [17]. The (k_1, k_2, p) placement of SMRT coefficients for $N = 8$ is shown in Fig.2. Sequency packets are shown with dark lines in the increasing order of packet index. SMRT representation of a signal is an integer-to-integer transform that requires additions only and is therefore computationally efficient in comparison with other transforms.

0,0,0	0,1,0	0,1,1	0,1,2	0,1,3	0,2,0	0,2,2	0,4,0
1,0,0	1,1,0	3,1,0	5,1,0	7,1,0	1,2,0	3,2,0	1,4,0
1,0,1	1,1,1	3,1,1	5,1,1	7,1,1	1,2,1	3,2,1	1,4,1
1,0,2	1,1,2	3,1,2	5,1,2	7,1,2	1,2,2	3,2,2	1,4,2
1,0,3	1,1,3	3,1,3	5,1,3	7,1,3	1,2,3	3,2,3	1,4,3
2,0,0	2,1,0	2,1,1	2,1,2	2,1,3	2,2,0	6,2,0	2,4,0
2,0,2	6,1,0	6,1,1	6,1,2	6,1,3	2,2,2	6,2,2	2,4,2
4,0,0	4,1,0	4,1,1	4,1,2	4,1,3	4,2,0	4,2,2	4,4,0

Fig. 2: (k_1, k_2, p) Placement of 8×8 SMRT Coefficients.

A 2-D SMRT feature [18] can be obtained as absolute sum of coefficients corresponding to different p values of a particular (k_1, k_2) which can be obtained as

$$f_{k_1,k_2} = \frac{\sum_{i=1}^{N_b} \sum_p \left| Y_{k_1,k_2}^p \right|}{I \times I} \tag{2}$$

Where $I \times I$ is the size of image, $N \times N$ is the block size, $M = N/2$, N_b is the number of blocks. Total number of features for a given image block of size N is $3N - 2$. Here non overlapping blocks are considered.
Similar concepts can be extended to 1-D case also. 1-D MRT for a signal $x(n)$, $0 \le n \le N - 1$ can be obtained as

$$Y_k^p = \sum_{\forall n|((nk))_N = p} x(n) - \sum_{\forall n|((nk))_N = p+M} x(n) \tag{3}$$

Where $0 \le k \le N - 1$, $0 \le p \le M - 1$ and 1-D SMRT [19] can be obtained as a sequency ordered placement of unique MRT coefficients. (k, p) placement of SMRT coefficients for $N = 8$ is shown in Fig 3. Here sequency packets are separated by dark borders with $(0,0)$ as S_0 represents sequency packet 0, $[(1,0), (1,1), (1,2), (1,3)]$ as $[S_1, S_2, S_3, S_4]$ belong to the first sequency with the four-phase terms termed as packet1, $[(2,0), (2,2)]$ as $[S_5, S_6]$ belong to the second sequency with 2 phase terms forming the packet2 and $(4,0)$ as S_7 forms sequency packet 3. This can be extended to higher orders also.

0,0	1,0	1,1	1,2	1,3	2,0	2,2	4,0

Fig. 3: (k, p) Placement of SMRT Coefficients For $N = 8$.

2.2.2. Facial feature extraction

Each video frame in the VidTIMIT database features a frontal view of the person of interest in it. The Viola-Jones face detection algorithm detects the face and discards background information [20]. The detected face image is then resized to a standardized boxed image size of 32×32 pixels, and visual features are extracted using 2-D SMRT from each video frame with a block size of 8×8. Thus, each video frame results in $3N - 2 = 22$ features. The final feature vector can be obtained by concatenating the feature vectors from each frame.

2.2.3. Voice feature extraction

Audio features are extracted using 1-D SMRT with a frame size of 256 samples. Each frame of 256 samples is transformed into 256 SMRT coefficients, which are arranged in 9 packets as ($[S_0]$, $[S_1$ to $S_{128}]$, $[S_{129}$ to $S_{192}]$, $[S_{193}$ to $S_{224}]$, $[S_{225}$ to $S_{240}]$, $[S_{241}$ to $S_{248}]$, $[S_{249}$ to $S_{252}]$, $[S_{253}$ to $S_{254}]$ and $[S_{255}]$). The absolute sum of coefficients in each packet forms voice features, and each frame is converted into a nine element feature vector.
The audio and visual features extracted using the above technique result in feature vectors of extremely large size. The video sample in the database consists of an average of 100 frames per sample. The concatenation of the extracted feature vectors from each video frame would result in a feature vector with an average size of 2000 for facial features alone. Similarly, for the audio data, the average signal length is around 100K, which can result in 390 frames of size 256, contributing to a feature vector of length $9 \times 390 = 3510$. Finally, feature level fusion of video and audio features results in a huge feature vector of size roughly 5000. The large feature vector size would require more computation time. Moreover, the size of the multimodal feature vector varies, depending on the number of video frames and the length of the audio file.

2.2.4. Statistical analysis

The randomness of a signal can be measured using statistical analysis. The measure of randomness of a particular instance over all frames, termed ensemble average (EA), is used to derive reduced features, as shown in Fig. 4. Statistical parameters like average energy (e), mean (μ), standard deviation (σ), skewness(s), and kurtosis (k) [19] are computed on an EA basis which is useful in capturing dynamically changing features per frame.

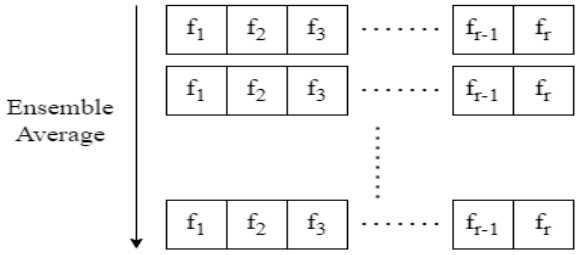

Fig. 4: Ensemble Average.

The detailed steps involved in video or audio feature extraction are shown in Fig. 5. SMRT is applied to all video and audio frames to extract base-level features. The number of transform coefficients extracted from each frame is represented by r which is 22 for the video frame and 9 for the audio frame. The number of frames varies depending on the size of the file and is represented by b. Once the base level features are extracted using SMRT reduced features are derived from them using five statistical parameters e, μ, σ, s and k computed on an EA basis.

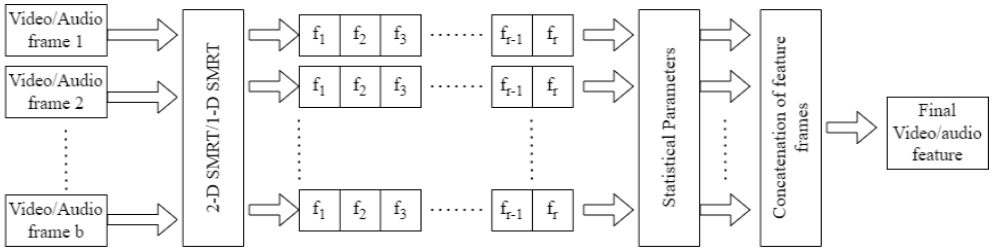

Fig. 5: Structure of Feature Extraction Model.

Thus the length of the reduced feature vector is given by $r \times 5$ which is $22 \times 5 = 110$ for the video file and $9 \times 5 = 45$ for the audio file. Hence, the size of the feature vector would remain constant irrespective of the number of frames.

2.3. Multimodal fusion

The choice of fusion strategy depends on the modalities being used and the application at hand. Since the application at hand is the design of audio-visual biometric system, it is necessary to take into account the correlation between audio and visual modalities, as the person's face deforms differently depending on what is being said and the underlying speaking style variations. Feature-level fusion effectively models correlated information between spoken utterance and the corresponding face appearance and is implemented by concatenating the feature vectors extracted from two modalities to create a new, more powerful feature vector that represents the individual more accurately. Here, the concatenation of audio and video feature vectors results in a combined feature vector length of 155.

2.4. Classifiers

Different classifiers are trained and tested to investigate the validity and potency of the proposed framework for improving the recognition performance.

- Support Vector Machine (SVM)

SVM is a classification tool that uses machine learning theory to make the most of predictive accuracy while automatically avoiding overfitting the data [21]. The idea of SVM is based on structural risk minimization, that tries to find an optimal hyperplane which maximizes the margin between classes. An SVM classifier with a soft margin and linear kernel is trained and tested.

- k-Nearest Neighbour

The idea of k-Nearest Neighbour method is to identify k samples in the training set whose independent variables x are similar to u, and to use these k samples to classify this new sample into a class, v.

- Random forests (RF)

One of the excellent ensemble machine learning techniques widely used in classification is Random forests (RF) [22]. The main idea of RF is to build many classification trees based on some randomly selected features from randomly selected samples with bagging strategy and then to use the trees to vote for a given input vector to get a class label.

3. Performance metrics

The performance of the proposed biometric system is evaluated using three standard evaluation metrics including accuracy, false acceptance rate (FAR), false rejection rate (FRR) and Equal Error Rate (EER). Accuracy can be evaluated as

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

TP is True Positive, TN is True Negative, FP is False Positive, FN is False Negative
FAR is the measure of the likelihood that the system will incorrectly accept an access attempt by an imposter.

$$\text{FAR} = \frac{\text{FP}}{\text{FP+TN}}$$

FRR is the measure of the likelihood that the system will incorrectly reject an access attempt by an authorized user

$$\text{FRR} = \frac{\text{FN}}{\text{FN+TP}}$$

EER is defined as the point where the value of FAR equals the value of FRR in the Receiver Operating Curve (ROC)

$$\text{EER} = \frac{\text{FAR+FRR}}{2}$$

Lower EER value indicates the higher performance of the system

4. Results

The main goal of this research is to improve the performance of biometric systems by combining information from audio and visual modalities with newly suggested SMRT features. Most of the existing systems are unimodal and rely only on a single modality, like video-only or audio-only systems. Video-only systems recognize the person from face images, and audio-only systems perform speaker verification from voice. Many research works have been carried out in this direction; however, their performance can be further enhanced by combining both audio and visual modalities. Thus, a robust audio-visual biometric system using SMRT features with feature-level fusion is proposed in this work. A comparison of the performance of audio-only, video-only, and audio-visual systems implemented using SMRT features is evaluated using different classifiers, including SVM, K-NN, and RF, and the results are tabulated in Table 1. The novelty of this work lies in the use of the sequency-mapped real transform for feature extraction which demonstrates how it can extract features from multiple modalities more efficiently with less computation.

Most of the research work on audio-visual biometric systems has been carried out on virtual databases that combine two different datasets for each modality. The main drawback of this type of database is that it cannot replicate correlations between the modalities. Moreover, most of the previous research work has been carried out on a face-image dataset rather than on a video dataset for the visual modality. Hence, the proposed system is tested on the VidTIMIT database, which offers video and audio files of the person of interest, which is useful to extract the correlated information from two modalities.

Table 1: Performance Evaluation

Methodology	Classifier	Accuracy	FAR	FRR	EER
Video-Only	SVM	94.2	4.3	8.7	6.5
	KNN	98.8	0.8	1.7	1.3
	RF	97.6	1.7	3.5	2.6
Voice-Only	SVM	96.8	2.4	4.8	3.6
	KNN	97.7	1.7	3.3	2.6
	RF	97.1	2.2	4.3	3.3
Audio-Visual	SVM	97.7	1.7	3.5	2.6
	KNN	99.4	0.4	0.9	0.6
	RF	98.26	1.3	2.6	1.9

It can be concluded from Table 1 that the highest accuracy of 98.8% and the lowest EER of 1.3 for a video-only system are attained with KNN classifiers. The performance of a voice-only system shows that although the performance of KNN and RF classifiers is comparable in terms of accuracy, the minimum EER is offered by the KNN classifier. Hence, it can be concluded that the KNN classifier offers better results for voice-only systems. Finally, the performance of the audio-visual system with different classifiers shows that the highest accuracy of 99.4% and the lowest EER of 0.6 are achieved with the KNN classifier. Thus, the performance comparison of audio-only, video-only, and audio-visual systems shows that the audio-visual system implemented with the KNN classifier offers the highest accuracy of 99.4% and the lowest EER of 0.6.

Table 2: Performance Comparison with Previous Published Methods

Method	Features		EER		
	Face	Voice	Video-Only	Voice-Only	Audio-Visual
[23]	2-D LDA	LPCC	2.1	2.7	1.2
[24]	DCT	MFCCs	6.2	8.13	5.7
[25]	DCT, GRD, CTR	MFCCs	3.2	4.2	0.73
[26]	MDLA	WLPCC	2.9	9.2	0.45
[27]	PCA, LDA, Gabor filter	MFCCs, LPCs. LPCCs	1.95	2.24	0.64
Proposed System	2-D SMRT	1-D SMRT	1.3	2.6	0.6

Table 2 indicates the obtained results in comparison with some of the previously published works.

In [23], facial and speech features are extracted using 2-D Linear Discriminate Analysis (2-D LDA) and Linear Prediction Cepstral Coefficients (LPCCs), respectively, and their EER is tabulated.

DCT and Mel Frequency Cepstral Coefficients (MFCCs) were used as feature extractors in [24].

Facial features extracted in [25] consist of three types of features: DCT features, grid-based lip motion (GRD) features, and contour-based lip motion (CTR) features; speech features are extracted using MFCCS.

Multiscale morphological erosion and dilation operations are used for extracting the facial features, visual and speech respectively, in [26]. The morphological dynamic link architecture (MDLA) method for face recognition uses multiscale morphological dilation and erosion under the elastic graph matching framework. Speech features are represented by the Weighted Linear Prediction Cepstral Coefficient (WLPCC).

A combination of different feature vectors is used in [27]. The feature vector for voice is MFCCs, LPCs, and LPCCs features, while the feature vector for face differentiation is excerpted using PCA, LDA, and the Gabor filter.

A comparison with other audio-visual systems shows that the proposed system offers better performance in terms of lower EER. The results reveal the ability of sequency-mapped real transform as an efficient feature extraction tool for multimodal applications, based on the better performance metrics obtained in comparison to the previously published results. This work can be extended to include more modalities as a future research project.

5. Conclusion

A novel method of feature extraction that is suitable to extract features from multiple modalities is proposed in this work. Performance comparisons of the proposed feature set with already existing feature extraction techniques show excellent results. The sequency-mapped real transform enables feature extraction with less computational complexity and therefore proves itself eminently suitable for bulk data handling scenarios such as real-time multimodal applications.

References

[1] N. a. M. M.-W. a. C. J.-T. a. N. L. a. M.-W. M. a. J.-T. C. Li, "DNN-Driven Mixture of PLDA for Robust Speaker Verification," IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 25, no. 6, p. 1371–1383, June 2017. https://doi.org/10.1109/TASLP.2017.2692304.

[2] C. S. X. Z. Y. Y. H. T. S. Fumin Shen, "Face image classification by pooling raw features," Pattern Recognition, vol. 54, pp. 94-103, 2016. https://doi.org/10.1016/j.patcog.2016.01.010.

[3] A. M. K. Javad Khodadoust, "Fingerprint indexing based on minutiae pairs and convex core point," Pattern Recognition, vol. 67, pp. 110-126, 2017. https://doi.org/10.1016/j.patcog.2017.01.022.

[4] L. N. Alessandra Lumini, "Overview of the combination of biometric matchers," Information Fusion,, vol. 33, pp. 71-85, 2017. https://doi.org/10.1016/j.inffus.2016.05.003.

[5] F. M. H. B. Jean-Philippe Thiran, Multimodal Signal Processing Theory and Applications for Human–Computer Interaction, Academic Press, 2010.

[6] K. M. R. G. B. Manju, "Prostate Disease Diagnosis from CT Images Using GA Optimized SMRT Based Texture Features,," in Procedia Computer Science, 2015.

[7] T. T. a. R. G. P. P. Mini, "Feature Vector Selection of Fusion of MFCC and SMRT Coefficients for SVM Classifier Based Speech Recognition System," in 2018 8th International Symposium on Embedded Computing and System Design (ISED), Cochin, India, 2018.

[8] R. S. A. R. Maneet Singh, "A comprehensive overview of biometric fusion," Information Fusion, vol. 52, pp. 187-205, 2019. https://doi.org/10.1016/j.inffus.2018.12.003.

[9] T. a. K. M. a. K. M. S. H. a. B. R. Mahboob, "Speaker identification using gmm with mfcc," International Journal of Computer Science Issues (IJCSI), vol. 12, 2015.

[10] S. L. H. Y. J.-H. K. G. K. S. Park, "Partially Occluded Facial Image Retrieval Based on a Similarity Measurement," Mathematical Problems in Engineering, 2015. https://doi.org/10.1155/2015/217568.

[11] W. W. a. J. Chen, "Occlusion robust face recognition based on mask learning," in 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017.

[12] C. Sanderson, "Biometric Person Recognition: Face, Speech, and Fusion," in VDM Verlag, 2008.

[13] R. Gopikakumari, Investigations on the development of an ANN model and visual manipulation approach for 2D DFT computation in image processing [PhD Thesis, Cochin University of Science and Technology, Kochi, India]. http://dyuthi.cusat.ac.in/purl/, 1998.

[14] R. C. &. G. R. Roy, "A new transform for 2-D signal representation (MRT) and some of its properties.," in 2004 International Conference on Signal Processing and Communications, SPCOM, 2, 363–367, 2004.

[15] R. C. Roy, Development of a New Transform : Mrt [PhD Thesis, Cochin University of Science and Technology, Kochi, India], 2009.

[16] V. Bhadran, Development & Implementation of Visual Approach and Parallel Distributed Architecture for 2-D DFT & UMRT computation [PhD Thesis, Cochin University of Science and Technology, Kochi, India], 2009.

[17] V. L. &. G. R. Jaya, " Sequency-based mapped real transform: properties and applications.," Signal, Image and Video Processing, vol. 11, no. 8, p. 1551–1558., 2017. https://doi.org/10.1007/s11760-017-1119-2.

[18] B. a. J. V. a. M. K. a. G. R. Manju, "8 × 8 SMRT Based Texture Descriptors," Lecture Notes on Software Engineering, vol. 3, pp. 295-298, 2015. https://doi.org/10.7763/LNSE.2015.V3.207.

[19] T. T. R. G. P.P. Mini, "EEG based direct speech BCI system using a fusion of SMRT and MFCC/LPCC features with ANN classifier," Biomedical Signal Processing and Control, vol. 68, p. 102625, 2021. https://doi.org/10.1016/j.bspc.2021.102625.

[20] P. J. M. Viola, "Robust Real-Time Face Detection," International Journal of Computer Vision, vol. 57, p. 137–154, 2004. https://doi.org/10.1023/B:VISI.0000013087.49260.fb.

[21] M. P. a. G. M. Foody, "Feature Selection for Classification of Hyperspectral Data by SVM," IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 5, pp. 2297-2307, 2010. https://doi.org/10.1109/TGRS.2009.2039484.

[22] L. Breiman, "Random Forests," Machine Learning , vol. 45, p. 5–32, 2001. https://doi.org/10.1023/A:1010933404324.

[23] R. A. K. G. Raghavendra R, "Multimodal person verification system using face and speech," in Procedia Computer Science, 2010. https://doi.org/10.1016/j.procs.2010.11.023.

[24] D. a. H. K. J. a. N. S. S. SHAH, "ROBUST MULTIMODAL PERSON RECOGNITION USING LOW-COMPLEXITY AUDIO-VISUAL FEATURE FUSION APPROACHES," International Journal of Semantic Computing, vol. 4, pp. 155-179, 2010. https://doi.org/10.1142/S1793351X10000985.

[25] M. W. Girija Chetty, "Robust face-voice based speaker identity verification using multilevel fusion," Image and Vision Computing, vol. 26, pp. Pages 1249-1260, 2008. https://doi.org/10.1016/j.imavis.2008.02.009.

[26] B. Y. S. Palanivel, "Multimodal person authentication using speech, face and visual speech," Computer Vision and Image Understanding, vol. 109, pp. 44-55, 2008. https://doi.org/10.1016/j.cviu.2006.11.013.

[27] H.Kasban, "A Robust Multimodal Bio metric Authentication Scheme with Voice and Face Recognition," Arab Journal of Nuclear Sciences and Applications, vol. 50, no. 3, pp. 120-130, 2017.