Computers, Materials & Continua
DOI:10.32604/cmc.2022.021608
Article

Tech Science Press

Dynamic Audio-Visual Biometric Fusion for Person Recognition

Najlaa Hindi Alsaedi* and Emad Sami Jaha

Department of Computer Science, Faculty of Computer Science and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
*Corresponding Author: Najlaa Hindi Alsaedi. Email: Nalsaedi0024@stu.kau.edu.sa

Received: 08 July 2021; Accepted: 07 September 2021

Abstract: Biometric recognition refers to the process of recognizing a person's identity using physiological or behavioral modalities, such as face, voice, fingerprint, gait, etc. Such biometric modalities are mostly used in recognition tasks separately as in unimodal systems, or jointly with two or more as in multimodal systems. However, multimodal systems can usually enhance the recognition performance over unimodal systems by integrating the biometric data of multiple modalities at different fusion levels. Despite this enhancement, in real-life applications some factors degrade multimodal systems' performance, such as occlusion, face poses, and noise in voice data. In this paper, we propose two algorithms that effectively apply dynamic fusion at feature level based on the data quality of multimodal biometrics. The proposed algorithms attempt to minimize the negative influence of confusing and low-quality features by either exclusion or weight reduction to achieve better recognition performance. The proposed dynamic fusion was achieved using face and voice biometrics, where face features were extracted using principal component analysis (PCA), and Gabor filters separately, whilst voice features were extracted using Mel-Frequency Cepstral Coefficients (MFCCs). Here, the facial data quality assessment of face images is mainly based on the existence of occlusion, whereas the assessment of voice data quality is substantially based on the calculation of signal to noise ratio (SNR) as per the existence of noise. To evaluate the performance of the proposed algorithms, several experiments were conducted using two combinations of three different databases, AR database, and the extended Yale Face Database B for face images, in addition to VOiCES database for voice data. The obtained results show that both proposed dynamic fusion algorithms attain improved performance and offer more advantages in identification and verification over not only the standard unimodal algorithms but also the multimodal algorithms using standard fusion methods.

Keywords: Biometrics; dynamic fusion; feature fusion; identification; multimodal biometrics; occluded face recognition; quality-based recognition; verification; voice recognition

1 Introduction

Biometrics has long been known as a robust approach for person recognition that uses different physiological or behavioral traits, such as face, voice, fingerprint, iris, gait, and many others [1]. The greatest majority of existing real-life biometric systems are unimodal, which means they make use of a single biometric modality and need to be accurately enrolled in a database to train a discriminative algorithm, and then to be sufficiently acceptable and usable in the recaptured probe or test samples for achieving successful recognition. Consequently, such systems mostly suffer from different limitations against some challenges, such as noise in sensed data, intra-class variation, inter-class similarity, and non-universality [2]. Multimodal biometric systems have thus been alternatively used for person recognition, as they are expected to be more accurate and reliable than unimodal systems and may overcome some of the unimodal limitations, because of the presence of multiple independent pieces of evidence [3] to be integrated and fused at either of four fusion levels, namely, sensor level, feature level, matching score level, and decision level [4].

Fusion at feature level is based on a concatenation of different feature vectors extracted from different biometric modalities to create a new more powerful feature vector with a higher dimensionality that represents the individual more accurately [5,6]. Since the feature sets may contain richer information from biometric modalities than the mere match scores or final decisions, using fusion at the feature level is expected to achieve better recognition results [7]. However, in some systems feature level fusion is more difficult to implement than other fusion levels, because of the relationship between the features spaces of different biometric systems may not be known and concatenating two feature vectors might lead to the dimensionality problem. Also, since the multimodal system may not have access to the feature values of individual modalities due to their proprietary nature, in such cases, fusion at the matching score or decision levels are the only option [3]. The standard fusion technique is used in the majority of literature, where representations from multimodalities are simply and symmetrically concatenated at any fusion level. On the other hand, few studies apply dynamic fusion mostly at the score level or dynamic selection of the classifier or fusion algorithm based on the quality or context of data [8]. In several earlier research studies [8–10], dynamic biometric fusion or classifier selection could improve the recognition performance over their standard fusion counterparts [8–10].

In this research, information of face and voice modalities are used for dynamic fusion at feature level. Human face is the most natural, user-friendly, and non-intrusive biometric measure; it is extensively used in daily life for identification, authentication, and retrieval, in such a way does not disturb people being identified [11]. Recently, face recognition techniques have achieved high performance using some public databases [12]. However, the accuracy of face recognition may degrade for many likely reasons, such as illumination, occlusion, facial expressions, and poses [12]. The other modality used here is voice, which has recently become one of the most efficient measures used to provide protection to an individual's computerized and electronic belongings [13]. The idea of voice recognition is to capture the voice as well as linguistic behavior of the speaker [1]. Therefore, voice can be considered as a combination of physiological and behavioral biometric forms [3]. Voice biometric has high potentiality for growth; since no special or sophisticated sensor is required, where a PC that already contains a microphone will be sufficient [1]. Similar to face recognition, voice recognition also has drawbacks, which influence the recognition accuracy in real life usage, such as environmental noise, medical conditions (e.g., a common cold), emotional state, etc. [3].

Recently, the coronavirus disease 2019 (COVID-19) pandemic has increased the focus on hygienic and contactless person recognition methods [14]. However, people worldwide are legally allowed or even obliged to wear face masks, which in turn degrade or negatively affect face and voice recognition performance. Hence, the prevalence of masked faces everywhere has become a serious concern to consider and a real challenge for face biometric systems to confront for successfully achieving person identification or verification (authentication). For instance, several security-related face recognition issues have been escalated in many regions after dozens of crimes were committed by criminals taking advantage of the COVID-19 face-covering rules. Users of voice authentication and identification systems also have been somewhat impacted by some consequent circumstances of the COVID-19 epidemic, as wearing a face mask may affect speech production by presenting an obstacle to the usual transmission of speech sounds. The effect of face masks on voice recognition can be similar to those caused by acoustic filters, like the sound-absorbing fabrics used for sound insulation [15]. Motivated by the challenging context of recognizing low-quality face and voice probe data besides the increased and urgent needs for developing such robust capabilities, we propose two dynamic feature-level fusion algorithms for improved and adaptive multimodal biometric identification/verification. The main contributions of this research are:

- A new proposed dynamic feature fusion method, with dynamic feature vector size, based on biometric data quality analysis and assessment of probe samples.
- A new proposed dynamic feature fusion algorithm, with dynamic weighting for features, based on the biometric data quality analysis and assessment of probe samples.
- An investigation of the effects and performance of using the proposed dynamic feature level fusion of face and voice biometrics for person identification and verification tasks.
- Performance comparisons of the proposed dynamic fusion approaches with two standard unimodal approaches using voice or face, and with a multimodal approach using the standard fusion of face and voice biometrics.

The rest of this paper is organized as follows. Section 2 shows a literature review of occluded face recognition, noisy voice recognition, standard face and voice fusion, and dynamic and quality-based person recognition systems. Section 3 shows the proposed framework of dynamic biometric fusion and the biometric data quality assessment along with two proposed dynamic feature fusion algorithms. The methodology, including databases description, a brief background about the used feature extraction methods, preprocessing, feature fusion, and the classifier used in our experiments, is shown in Section 4. Section 5 shows the experimental results and analysis. This research is eventually concluded in Section 6.

2 Related Work

2.1 Occluded Face Recognition

A number of recent research studies are concerned with removing the occlusion area to effectively recognize occluded faces [16–20]. Jianxin et al. [16] proposed a block-oriented method for partially occluded face recognition. Their proposed algorithm was designed to segment the face image and extract Histogram of Oriented Gradient (HOG) and Local Binary Pattern (LBP) features from each block of the image to obtain HOG-LBP joint features. Then to attain identification using sparse representation reconstruction residual, where their algorithm enhanced the recognition performance and provided more robustness against occlusions compared with other traditional algorithms. In [17], two approaches were proposed for masked faces recognition. The first approach was designed using an attention-based module to focus on the region around the eyes, which improved the performance of masked face recognition over other attention modules in comparison. The second approach was a cropping-based investigating the optimal cropping for each case in masked face recognition, which also improved the recognition performance as per their experimental results. Another proposed method for partially occluded face recognition [18] was implemented by dividing each image into sub-images and detecting occluded regions using eigenfaces, then using Gabor filters to extract features from unoccluded sub-images, where only those unoccluded sub-images were used in matching. Wu et al. [19] proposed a partial occlusion facial attitude estimation algorithm based on HOG in the direction of the pyramid. Their approach divides a detected face horizontally into two sub-images, to predict the existence of any occlusions in these two sub-regions individually. After that, pyramid HOG features are extracted from unoccluded sub-images and used with a Support Vector Machine (SVM) classifier to recognize a person's identity. Moreover, Song et al. [20] developed a different approach for occluded face recognition based on Pairwise Differential Siamese Network (PDSN) such that, a mask dictionary is established using the differences between the top convoluted features of occluded and unoccluded face pairs, which indicated the correspondence between occluded facial areas and damaged feature elements. Their experimental results on synthetic and realistic occluded face datasets showed that their proposed approach achieved higher performance compared with some state-of-the-art results.

2.2 Noisy Voice Recognition

Voice recognition nowadays is increasingly explored and effectively utilized in numerous applications, such as security, access control, and forensics [21]. Despite the significant advances and increased accuracy in performance of this biometric modality, it still faces serious challenges, such as noise in probe data [22]. Consequently, several research studies explore the efficacy of voice recognition in noisy environments using Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction [13,21–24]. In [23], the researchers proposed an architecture using a convolutional neural network (CNN) and MFCC to identify the speaker in a noisy environment. The proposed methods are based on a hybrid feature extraction using CNN as a feature extractor combined with MFCC as a single feature set, then the speakers were classified using a deep neural network. They achieved an improved performance reaching up to 87% accuracy. Kawakami et al. [24] proposed a method for speaker identification in a noisy environment and investigated the effect of pitch synchronous phase information when combined with MFCC for speaker identification by combining Gaussian Mixture Model (GMM) based on MFCC with GMM based on phase information. This method improved remarkably the recognition accuracy from 31.8% to 61.6%. In [25], an architecture pipeline of near real-time speaker recognition was proposed, which exploited the advantages of hybrid feature extraction techniques by using Gabor filters, CNN, and statistical parameters. To minimize the influence of the environmental noise, their proposed system enforced the recursive least squares (RLS) as an adaptive filtering method for the noise cancellation. Their experimental results showed that the recognition accuracy was enhanced up to 9% when compared to the standard AlexNet architecture. In [26], a forensic speaker verification system was proposed investigating the efficiency of combining the features of MFCCs with MFCC extracted from the discrete wavelet transform (DWT). The performance of this system was evaluated for these features with and without warping, and the experimental results showed that the fusion feature warping DWT-MFCC and feature-warped MFCC approach achieved high verification performance under different environmental noise and reverberation conditions.

2.3 Face and Voice Fusion

Several recent research efforts have presented different multimodal biometric schemes for person recognition using voice and face in fusion. In [27], a multimodal biometric system using face and voice was developed and different fusion techniques were tested and compared to measure their performance. The results showed that feature level fusion of face and voice data achieved the best recognition score, rank, and decision among the test fusion levels. In [4], they designed and developed

Android-based multimodal biometric authentication system, using LBP and MFCC features for face and voice, respectively. Their experiments were conducted using the Georgia Tech face (GT_DB) database and TIMIT database for face and voice data, respectively. Their proposed method achieved higher verification performance than the other compared methods in their experiments. Another multimodal identification approach for human authentication based on face and voice recognition was proposed in [28], where face feature was extracted using principal component analysis (PCA) and eigenfaces, while voice feature extraction was done using MFCC, linear prediction coefficients (LPC), and linear prediction Cepstral coefficients (LPCC). The classification was attained using GMM, SVM, and artificial neural network (ANN) for each modality separately, as they used and compared matching score level fusion and feature level fusion. Thus, matching score fusion achieved 0.62% equal error rate (EER), whereas feature fusion achieved 2.81% EER. In [29], a face-voice multimodal recognition approach was proposed and a number of experiments were conducted in three feature fusion mechanisms, concatenation of pre-normalized features, merging normalized features, and multiplication of features. The results showed that the merging fusion was the most effective mechanism.

2.4 Dynamic and Quality-Based Recognition

Several research studies improved the person recognition performance by applying dynamic score fusion or dynamic classifier selection. In [9], the researchers proposed a method that optimizes the accuracy and computation time by performing a dynamic selection of classifiers and fusion schemes based on the quality and pose of the input biometric data, while in [8] a framework was proposed for dynamic classifiers selection and fusion based on the data quality of gallery and probe images for face and fingerprint biometrics. The experimental results showed that the quality-based classifier selection maintained good performance even though the quality of image data is not optimal. In [30], a quality-based recognition method was offered, in which the quality information was used to switch between different system modules depending on the data source and to reject channels with low-quality data during the fusion. This method achieved an overall improvement of 25% in terms of EER. A quality-based multimodal biometric system was explored in [10], which adaptively combines the scores from individual classifiers. Moreover, it was reported to achieve 99.5% accuracy with 0.5% EER, outperforming the other compared state-of-the-art methods. In [31], the suggested method improved the performance by incorporating quality measures in multimodal biometric fusion to determine the reliability of the results given by fusion methods. Furthermore, another research work developed a mobile biometric recognition system to analyze face and voice information using a score-level fusion scheme driven by the quality of the biometric samples, as in [32], which yielded increased accuracy by 4.14% and 7.86% over the counterpart unimodal face and voice respectively. Sellahewa et al. [33] demonstrated the usefulness of quality-based adaptive normalization and adaptive score fusion for face recognition, to overcome the adverse effects of varying illumination conditions. Tab. 1 summarize some of relevant methods shown in this section.

Table 1: A summary review of some most relevant methods

Reference	Biometric	Classifier	Dataset	Evaluation metric	Result	Limitations
[20]	Face	CNN	MegaFace, AR, LFW	Accuracy	99.2%	Requirements of paired pictures are difficult to be satisfied in real-life applications
[23]	Voice in noisy environment	CNN	Their own collected dataset.	Accuracy	87%	High dimensionality of feature vectors, and high computation time.
[27]	Face and voice	KNN	AR, voice	Accuracy	100%	Face features with high dimensionality & did not consider low quality challenges.
[33]	Face	KNN	Extended Yale B	EER	7.9%	Both high quality and low-quality parts of image is used in identification

3 Proposed Framework of Dynamic Biometric Fusion

Based on the literature, it appears that dynamic fusion at the feature level has yet to be extensively explored; therefore, in this research, we propose a framework of dynamic biometric fusion along with two dynamic feature fusion algorithms based on the data quality of test modalities. This is to improve the person recognition performance and robustness against various types of face occlusion and environmental noise in voice data. We firstly analyze and assess the biometric data quality of the probe samples in order to appropriately apply the dynamic fusion. In this framework, other than the standard processes of feature extraction, modeling, and biometric template storage, the training phase is followed by an additional process in which dynamicity occurs on training data to reconstruct the templates corresponding to the dynamic size or weight, as decided by the data quality-based analysis and assessment in the test phase as shown in Fig. 1. As such, the first algorithm adopts a dynamic size for the feature vector by excluding the low-quality features, whilst the second algorithm is based on dynamic weighting for affected features. As illustrated in Fig. 1, both dynamic fusion algorithms handle the training data initially in the same manner as the standard fusion. However, the dynamic fusion takes place for both training and testing data only at testing time.

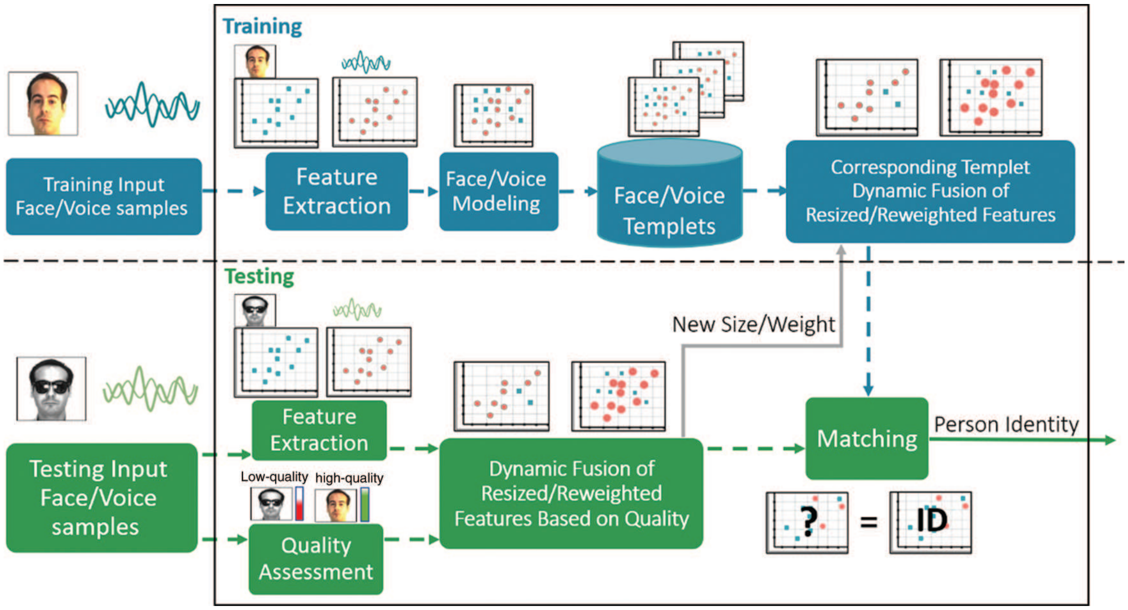

Training
Training Input Face/Voice samples
Feature Extraction
Face/Voice Modeling
Face/Voice Templets
Corresponding Templet Dynamic Fusion of Resized/Reweighted Features
Testing
New Size/Weight
Testing Input Face/Voice samples
Feature Extraction
Low-quality
high-quality
Quality Assessment
Dynamic Fusion of Resized/Reweighted Features Based on Quality
Matching
Person Identity
? = ID

Figure 1: Overview of the proposed dynamic biometric fusion framework

3.1 Biometric Data Quality Assessment

3.1.1 Face Data Quality Assessment

The quality of face image data varies due to different factors include occlusion, illumination, pose, aging, and expression [34,35]. Regardless of the type of face occlusion, the existence of occlusion in the face image can degrade the data/information quality, which consequently affects the quality of extracted features and its suitability for automated matching [15,36]. In the proposed framework, the quality assessment and analysis considered in the dynamic fusion framework are mainly based on the existence of face occlusion. Therefore, we consider four categories of face images, which are normal/unoccluded face image (does not contain any occlusion), face with sunglasses occlusion (upper part of face is occluded), face with scarf occlusion (lower part of face is occluded), and side occlusion (one side of face is vertically occluded). A normal face image is considered as a high-quality image; therefore, the whole extracted face feature vector will be used in the fusion standardly, whilst an image of a face with sunglasses, scarf, or side occlusion is considered as a low-quality image, where either dynamic resizing or reweighting will be applied in such cases.

In this research, SVM classifier is used to assess the image data quality by detecting the existence of occlusion in a probe image, since it is widely used and has proved its effectiveness for occlusion detection in several research studies [16,37,38]. SVM classifier is used here to classify images into one of the four aforementioned categories (i.e., normal/unoccluded face, face with sunglasses occlusion, face with scarf occlusion, and side occlusion). When a face image is classified as an image with sunglasses, scarf, or side occlusion, the occlusion area is then determined using the watershed algorithm, which is deemed as a fast, simple, and intuitive method for such detection processes [39]. It is an unsupervised algorithm designed for region segmentation that uses an intensity-based topographical representation of a grayscale image, where the bright pixels in the image represent higher altitudes or the 'hills' and the darker pixels correspond to the 'valleys' of the topography, which then is flooded from the bottom up, eventually the watersheds appear as lines dividing different regions in the image [40]. The resulting segmentation of an image consists of sets of connected pixels belonging to the same region, where the regions are non-overlapped, and sets of watershed pixels represent the border between regions. Fig. 2 shows an example of different categories of face occlusion detected and segmented using the watershed algorithm.

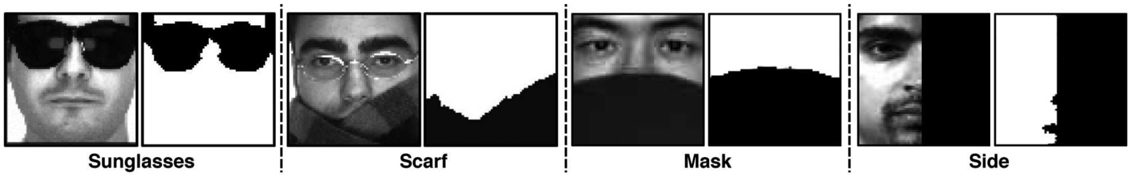


Figure 2: Pairwise examples of quality assessment result for low-quality occluded face data samples, in each pair, left: is the original image, and right: is the detected and segmented occlusion

3.1.2 Voice Data Quality Assessment

Speech signal quality is of fundamental importance for accurate speaker recognition [34]. The current situation of COVID-19 influences voice recognition systems, since covering mouth and nose by face masks often affects the production of speech by forming an obstacle to the usual transmission of speech sounds [15]. In this research, the data quality assessment and analysis of the probe voice sample were done by calculating signal to noise ratio (SNR) for each probe sample, since it is commonly used for such a purpose and its effectiveness has been verified in different research studies [41,42]. SNR can estimate speech quality by comparing the level of the desired signal to the level of background noise. Intuitively, higher values of SNR mean higher signal qualities, which lead to fewer errors in recognition. To calculate SNR, an energy-based voice activity detector was used to detect voiced and unvoiced sections as shown in [21], then the audio wave was separated into non-overlapping frames of 20 milliseconds, and the average energy was calculated for both voiced and non-voiced frames. Eventually, SNR can be computed as follows [21]:

$$SNR = 10 \times log\left(\frac{E_{voiced}}{E_{unvoiced}}\right) \tag{1}$$

where E_{voiced} and $E_{unvoiced}$ refer to the mean energies of the voiced and unvoiced sections, respectively. The result of SNR is used as voice quality assessment result. Our experiments confirm the result shown in [41], which shows performance reduction when SNR was less than 20 dB. Hence, the voice is considered low-quality when the calculated SNR is less than 20 dB.

3.2 Dynamic Multimodal Biometric Fusion Algorithms

3.2.1 Dynamic Size Fusion Algorithm

The first proposed algorithm is based on adopting a dynamic size of the feature vector nascent by the multimodal biometric fusion. Since the fusion is performed based on the quality of the probe image and audio, the feature vector size reduction was done by excluding unnecessary confusable information and minimizing the number of features of the low-quality modality/modalities. Hence, we can consider four cases of biometric feature fusion, which are:

First Case: When both face and voice samples are with high-quality information, the full extracted feature vectors of both face and voice information will be used in the fusion, so final feature fusion will be accomplished as same as the standard way.

Second Case: When the face sample is with high-quality information and the voice sample is noisy with low-quality information, the full extracted face feature vector will be used in fusion, while the extracted voice feature vector will be minimized to the most relevant high-quality feature information using PCA. As such, the final nascent feature vector size will be smaller than the feature vector of the standard fusion.

Third Case: When the face sample is with low-quality information and the voice sample is with high-quality information, the extracted face feature vector will be minimized by PCA to the most high-quality relevant feature, while the full extracted voice feature vector will be used in the fusion operation. Thus, the final resultant feature vector size will be smaller than the feature vector of the standard fusion.

Fourth Case: When both face and voice samples are with low-quality information, both extracted feature vectors will be minimized by PCA to the most high-quality relevant features. Thus, the final resultant feature vector size will be further smaller than the feature vector derived by the standard fusion.

Algorithm 1: Dynamic size fusion

```
INPUT: Test face and voice data
Detect occlusion in Test face image
Measure signal_to_noise_ratio
IF: occlusion_existence == true
    THEN: face_quality = low
    ELSE: face_quality = high
END IF
IF: signal_to_noise_ratio < threshold
    THEN: voice_quality = low
    ELSE: voice_quality = high
END IF
IF: face_quality = low
    THEN:
    Remove occlusion area from Test image
    Remove the same Test occlusion area from Train images
    Apply feature extraction to Test and Train face images
    Select only top-k high-quality unoccluded features
    Minimize size of Test face feature vector based on top-k
    Resize all Training face feature vectors accordingly
    ELSE:
    Use full-size for Test and all Training face feature vectors
END IF
IF: voice_quality = low
    THEN:
```

(Continued)

```
            Select top-k high-quality features
            Minimize size of Test voice feature vector based on top-k
            Resize all Training voice feature vectors accordingly
        ELSE:
            Use full-size for Test and all Training voice feature vectors
    END IF
    Combine face and voice feature vectors
    Match Test and Training feature vectors
    RETURN recognized person identity
```

3.2.2 Dynamic Weighting Fusion Algorithm

The second proposed algorithm is based on dynamic weighting for low-quality features, which adopts a fixed size of feature vectors after fusion, but a dynamic weight range of those features based on their feature information quality. When a face image sample is classified as a face with sunglasses, scarf, or side occlusion, the occlusion area is determined as explained in Section 3.1.1. After that, the extracted features from occlusion pixels will be minimized to range of weight smaller than the remaining features extracted from the pixels of unoccluded face region, where this is carried out by multiplying the pixels of occlusion by a reduction weight ranging between 0 and 1. Also, when a voice sample has an SNR less than 20 dB, which means the audio is either captured in a noisy environment or the speaker is speaking in a low voice, then the features will be minimized to range of weight less than the range of the face features to be fused with them. Thus, there are four cases of feature weighting in this algorithm, which are:

First Case: When both face and voice samples are with high-quality information, the full feature vectors will be normalized using min-max normalization and used as they are.

Second Case: When the face sample is with low-quality information and the voice sample is with high-quality information, the features derived from the occlusion pixels will be normalized using min-max normalization and multiplied by a suited dynamic reduction weight (0–1) determined based on empirical analysis to minimize their adverse effects on recognition performance, whilst the voice feature vector will be normalized using min-max normalization and used as it is.

Third Case: When face sample is with high-quality information and the voice sample is with low-quality information, the voice features will be normalized using min-max normalization then multiplied by a suited dynamic reduction weight (0–1) calculated based on the resulting SNR to minimize their adverse effects on recognition performance, whereas the face features will be normalized using min-max normalization and used as they are. The weight, in this case, is calculated as follows:

$$Weight = \frac{SNR}{20} \tag{2}$$

where SNR refers to the calculated signal-to-noise ratio of test audio data.

Fourth Case: When both face and voice are with low-quality information, they both normalized using min-max normalization, then the voice features will be multiplied by a suited dynamic reduction weight (0–1) computed relatively to the SNR using the Eq. (2). Furthermore, the face features extracted from occluded pixels will be multiplied by a suited dynamic reduction weight (0–1). While the face features inferred from unoccluded pixels will be used as they are.

Algorithm 2: Dynamic weighting fusion

```
INPUT: Test face and voice data
Detect occlusion in Test face image
Measure signal_to_noise_ratio
IF: occlusion_existence == true
THEN: face_quality = low
ELSE: face_quality = high
END IF
IF: signal_to_noise_ratio < threshold
    THEN: voice_quality = low
    ELSE: voice_quality = high
END IF
IF: face_quality = low
    THEN:
    Segment occlusion pixels and determine occlusion_features
    weighted_occlusion_features = occlusion_features × reduced_weight
    Replace occlusion_features with weighted_occlusion_features
    ELSE:
    Use same weighting for Test and all Training face feature vectors
END IF
IF: voice_quality = low
    THEN:
    weighted_voice_features = voice_features × reduced_weight
    Replace voice_features with weighted_voice_features
END IF
Combine face and voice feature vectors
Match Test and Training feature vectors
RETURN recognized person identity
```

4 Methodology

4.1 Databases Description

In this research, three different standard databases were used for conducting the experimental work, comprising AR [43] and Extended Yale Face database B [44,45] datasets for face image data, besides VOiCES database [46] for voice data. Note that the usage of these different face image databases enables evaluation and comparison of variation in recognition performance of the proposed dynamic fusion techniques from different aspects and on different realistic or synthetic occlusion forms.

4.1.1 Face Databases

AR database is one of the widely used databases to evaluate recognition performance on occluded faces using various algorithms [17,47,48]. It includes 26 frontal face images with different facial expressions, lighting conditions and occlusions for each of 126 distinct persons, 70 males, and 56 females. The images of most persons were taken in two sessions (13 images per session), separated by two weeks. Fig. 3 demonstrates some AR database face samples used for training and testing.

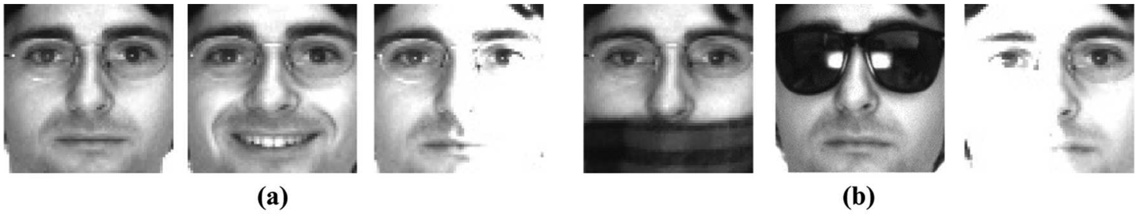


Figure 3: Face data samples of AR database: (a) high-quality (no occlusions) data samples used for training, and (b) high-/low-quality (unoccluded/occluded) data samples used for testing

Extended Yale Face database B has been also extensively used to evaluate the recognition performance of different algorithms under synthesized occlusions [17,47,48]. It consists of 16128 frontal images of 28 distinct persons with nine different poses and 64 illumination conditions. In our experiments, approximately 14 different images per person with different illuminations and poses were used for training, whereas a synthetic occlusion was added to a set of about 80% of face testing images to emulate different forms of realistic occlusions such as mask, scarf, and side occlusion. Figs. 4a and 4b demonstrates some Extended Yale face database B samples used for training and testing, respectively.

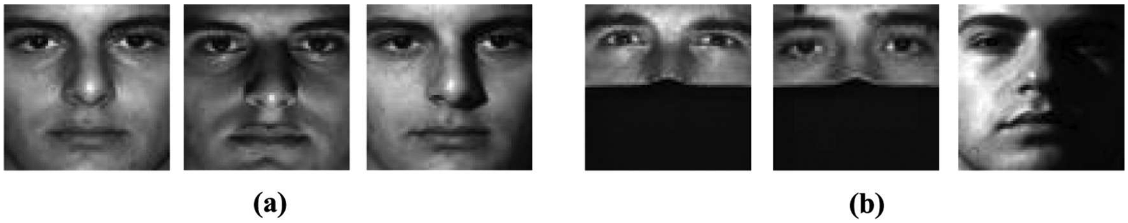


Figure 4: Face data samples of Extended Yale database B: (a) high-quality (no occlusions) data samples for training, and (b) high-/low-quality (unoccluded/synthetically occluded) data samples for testing

4.1.2 Voice Database

VOiCES database consists of 3,903 clear audio files for 300 different speakers. In our experiments, approximately 14 audio files per speaker were used for training. For testing, to emulate environmental noise we added synthetic non-stationary noise obtained from two different noise types (street, rain). Hence, testing was accomplished using a combination of clear and noisy audio files, where synthesized noises with different levels and durations at different times were added to most of the testing audio files. Fig. 5 shows clear and noisy voice wave samples, illustrating the difference of amplitude between the unvoiced sections of the clear and noisy voices.

It is noteworthy that in our experimental work we have conceptually and consistently assigned persons' face data with speakers' data, so each subject in our experiments has both unique person images and unique speaker audio files, assuming that they are the face and voice biometrics of the same person.

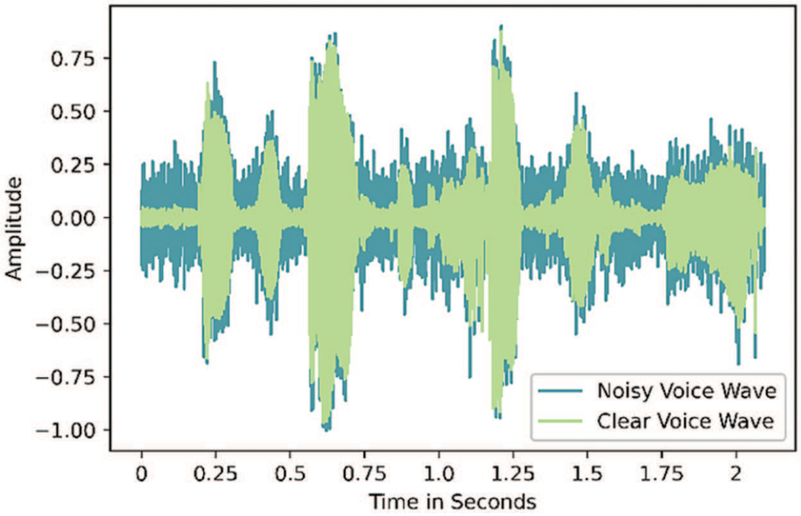


Figure 5: Clear and noisy voice waves: blue represents noisy voice and light green represents clear voice

4.2 Preprocessing

4.2.1 Face Data Preprocessing

Since, the used face images are from different databases with different properties, a suitable preprocessing is necessary for normalization. Consequently, we carry out a face detection using Haar cascade classifier which is object detection method proposed by Paul Viola and Michael Jones were used to localize the face region to be then cropped and used as the area of interest. For face detection using various multiple Haar features, each feature produces a numerical value by calculating the difference between the number of pixels under the white area and the number of pixels under the black area [49]. At some predefined threshold, the Haar features can classify face existence in a processed image as positive or negative [50]. At the end of Haar face detection process, we remove the other non-face or background parts of the image and keep a square image area containing only the face. Finally, the cropped face image is resized to 64 × 64 pixels. Fig. 6 shows a face sample before and after preprocessing.

4.2.2 Voice Data Preprocessing

Since the voice samples of VOiCES database have different durations and to prepare voice data for feature extraction, we implement cropping for each sample to maintain and use only the first voiced seven seconds. Moreover, external synthetic noises with different levels at different times and durations have been added only to test samples to investigate and compare the effect of environmental noise on the performance of the proposed dynamic fusion algorithms *vs.* their standard fusion counterpart.

4.3 Feature Extraction

4.3.1 Face Feature Extraction

After preprocessing, the face features are extracted using different feature extraction methods to verify the effectiveness of the proposed dynamic fusion techniques. A number of experiments are conducted using Gabor-based and PCA-based feature extraction from face data, for either constructing a unimodal feature vector or standard fusion feature vector (as baselines for performance comparison), or to be used in dynamic fusion by either minimizing the size of the feature vector or reweighting its features.

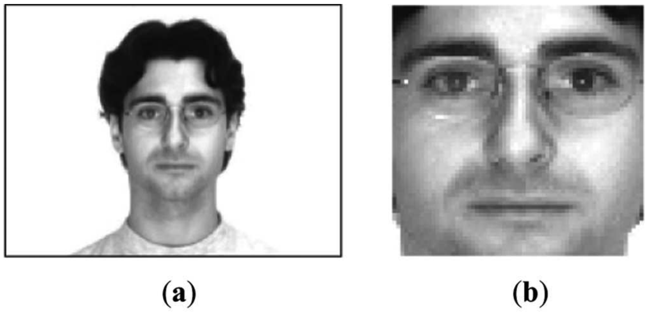


Figure 6: AR face sample before and after preprocessing: (a) original image, and (b) preprocessed image

- *Gabor-based features:*

Gabor features have attracted considerable attention and achieved enormous success in many face recognition purposes due to their capabilities for analyzing the visual appearance of an image and extracting discriminative feature vectors [51,52]. Gabor filters have been used and confirmed to be useful in several biometric applications, including face detection or recognition, iris recognition, and fingerprint recognition [53]. There are several research studies, as reviewed in [47], where Gabor based algorithms have achieved high accuracies in occluded face recognition. In this research, 2D Gabor filters were used as one of feature extraction methods, which can be mathematically represented as follows:

$$\Psi_F\left(u, v; f, \theta\right) = e^{\frac{-\pi^2}{f^2}\left(\gamma^2 (u\prime - f)^2 + \eta^2 v\prime^2\right)} \tag{3}$$

where

$$u' = u\cos\theta + v\sin\theta, \quad v' = -u\sin\theta + v\cos\theta \tag{4}$$

where u and v are the variable pair of the filter frequency, f refers to the central frequency of the filter, θ refers to the rotation angle of the Gaussian major axis and the plane wave, γ and η refers to the sharpness along the major and minor axes, respectively [51].

In this research, face features were extracted using 14 different Gabor filters were used, with two different scales (3 and 6), and seven different rotation angles θ (15°, 30°, 45°, 60°, 75°, 90°, and 120°), as demonstrated on the spatial domain in Fig. 7. An example of the resulting 14 face images after applying these Gabor filters on an image sample are shown in Fig. 8.

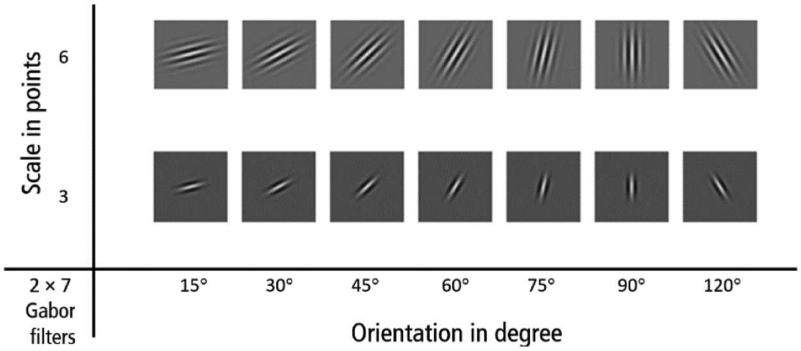


Figure 7: The 14 created and used Gabor filters

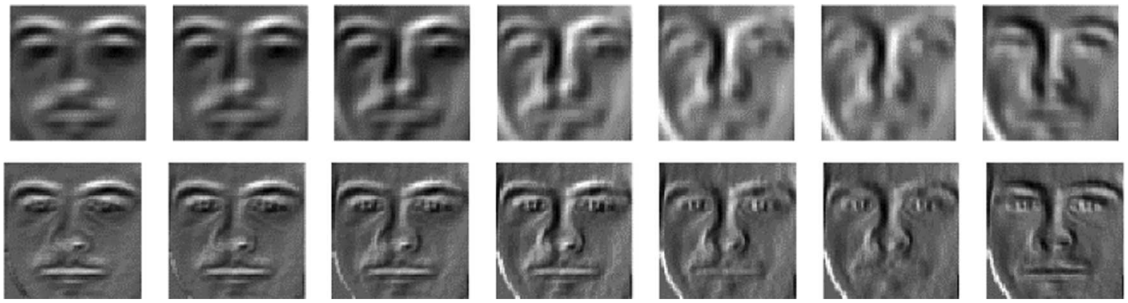


Figure 8: A sample of face image after filtering

Gabor filters-based feature extraction methods are usually computationally expensive due to the high dimensionality in calculation. Hence, the dimensionality of the whole extracted feature vector was reduced using an effective dimensionality reduction method proposed in [53]. This method was designed to address a single large image comprising all 14 concatenated filtered images. In this large image, for each *i* row-group consisting of *d* rows and each *j* group of columns consisting of *d* columns, the dimensionality is reduced by removing the d^{th} row from each *i* row-group and the d^{th} column from each *j* column-group. The used level of dimensionality reduction in our experiments was set to the dimension of ($d = 2$).

- *PCA-based features:*

The other feature extraction method for face images was PCA, which another widely used feature extraction and reduction method for face recognition [54]. It transforms the original data into a new less dimension set of data containing the most relevant information that may reveal characteristics of the data that were once hidden before the transformation [55]. This is done by constructing M feature vector from M training samples, each vector of size N, where $N =$ image height × image width. After that, an average image is constructed from the created M vectors using the following formula:

$$m = \frac{1}{M}\sum_{n=1}^{M}\Gamma_n \tag{5}$$

where m refers to the average image, and r refers to the image vector. After that, the average image is subtracted from each input image as follows:

$$x_i = \Gamma_i - m \tag{6}$$

The results are arranged on a matrix $X = [x_l, x_2, \ldots, x_n]$ of dimension $M \times N$, where each column in the matrix represents an image sample. Then a covariance matrix is obtained as:

$$C_X = XX^T \tag{7}$$

Next, eigenvectors and eigenvalues are calculated from the following formula:

$$C_X U = U\Lambda \tag{8}$$

where Λ is the diagonal matrix of eigenvalues of matrix C_X, and U is the associated eigenvector. For image vector x, there are N possible projections, defined as:

$$y_j = u_j^T x \tag{9}$$

where $j = 0, 1, \ldots, N$, and u_j refers to the eigenvectors of the covariance matrix C_X. The resulting y_j is the principal components that are also called the eigenfaces [56,57].

4.3.2 Voice Feature Extraction

- *MFCC-based features:*

The acoustic features were extracted using MFCCs, which is perhaps the most commonly used feature extraction technique for speaker recognition [36]. MFCCs are based on short-term analysis, which carries the speech in the frequency domain with short segments and computes the MFCC feature vector from each segment [23]. In this research, MFCCs were used as they were found to be robust against noise and capable to detect speech characteristics even in low-frequency regions [23]. In MFCC, the digitized audio signal was blocked into small duration frames of 22 to 32 milliseconds [23,58]. Then a hamming window function was performed on each individual frame. This offers the bell-shaped weighting function with no zero at the edges of the window, to minimize the spectral distortion. This window can be defined as:

$$w_n = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right), \quad \text{where} \ \ 0 \le n \le N \tag{10}$$

where N is the number of samples in each frame [23,58]. The result of the windowing step can be defined as:

$$Y_1[n] = x[n] \cdot w[n] \tag{11}$$

where $x[n]$ is the n^{th} speech sample in the frame. After that, Fast Fourier transform (FFT) is performed on every frame to get the magnitude frequency. The FFT can be formulated on the set of N as follows:

$$Y_2[n] = \sum_{k=0}^{N-1} Y_1[k]e^{\frac{-2\pi jkn}{N}} \tag{12}$$

Here, $n = 0, 1, 2, \ldots, N{-}1$, and j refers to the imaginary unit, which is $\sqrt{}-1$. Then to compute the spectrum, the square of the magnitude for each frequency component is taken as follows [13]:

$$Y_3[n] = (real(Y_2[n])^2) + (imag(Y_2[n])^2) \tag{13}$$

To simulate non-linear of the human ear, we need to convert signals in frequency into Mel-frequency by Mel-scale, using the following formula [13]:

$$mel(f) = 2595 * log_{10}\left(\frac{1+f}{700}\right) \tag{14}$$

where f is the frequency in Hz. To simulate human perception, a Mel Filter Bank filters an input power spectrum through a bank of Mel-filters. The filter banks are a set of triangular windows spaced uniformly on the Mel-scale. The output is an array of filtered values, typically called Mel-spectrum, calculated as follows:

$$Y_4[n] = \sum_{i=0}^{\frac{N}{2}} Y_3[i] \times melweight[n][i], \quad \text{where } 0 < \text{n} < \text{k} \tag{15}$$

where k is number of filters. At the end of this step, only the useful features are preserved [23,58]. The resultant values from the Mel filter bank are reduced by calculating the natural logarithm of each value as follows:

$$Y_5[n] = ln(Y_4[n]), \quad \text{where } 0 \leq \text{n} < \text{k} \tag{16}$$

Eventually, Discrete Cosine Transform (DCT) is used to convert log Mel back into the time domain, where the result of this step is called the MFCC [21] which can be calculated as:

$$Y_6[n] = \sum_{k=1}^{K} Y_5[k] \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad n = 1, 2, 3, \ldots K \tag{17}$$

The process of MFCC extraction is summarized as shown in Fig. 9.

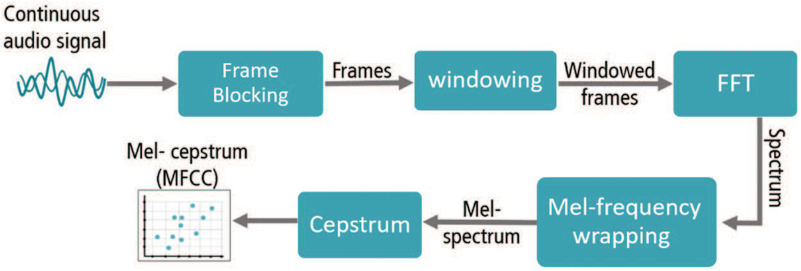


Figure 9: Block diagram for MFCC extraction process

4.3.3 Feature Fusion

After face and voice feature extraction, the fusion of multimodal feature vectors is applied based on the proposed dynamic fusion methods described in Section 3.2.1 and 3.2.2. The output of such fusions is a single feature vector of higher dimensionality obtained by the concatenation of face and voice feature vectors based on dynamic size/weight.

4.4 Classifier Training

To investigate the validity and potency of the dynamic fusion framework for improving the recognition performance, we train and use an SVM-based classifier for experimenting the proposed dynamic fusion approaches. The idea of SVM is based on structural risk minimization that tries to find an optimal hyperplane that maximizes the margin between classes [59,60]. The separation can be tuned by the C value (known as regularization parameter or penalty factor), which refers to the softness of the margin. A minimal value of C refers to a softer margin, where this yields some classification errors to maximize the separation margin, whereas the large value of C makes the SVM fitting better the training data with regard to the decision function's margin maximization [61].

In this research, an SVM classifier with soft margin and linear kernel is adopted and learned for person recognition. In each conducted experiment, a corresponding SVM classifier is trained and tested using four different types of feature vectors derived using PCA, Gabor and MFCC feature extractors as follows: **first**, feature vectors based on *unimodality* derived for each of face and voice data, as two face unimodal feature vectors are separately extracted and tested using PCA and Gabor feature extractors; **second,** a *standard fusion* consisted of the whole extracted features constructed using (PCA/Gabor-based) face and voice feature vectors; **third**, a *dynamic size fusion* comprising a minimized feature vector derived using the first proposed dynamic fusion method; and **fourth**, a *dynamic weight fusion* implicating the whole extracted features with reduced weights for low-quality features was performed using the second proposed fusion method of dynamic weighting for extracted features. To investigate and compare variation in recognition performance, the same experiments were separately conducted on each combination of the AR with VOiCES databases and Extended Yale B with VOiCES databases.

5 Experiments and Analysis

In this section, the conducted experiments and the achieved results are shown and analyzed, where all experiments were similarly conducted per face database using the three different types of feature fusion, 'standard fusion', '*dynamic size fusion*', and '*dynamic weight fusion*', as well as three unimodal face and voice algorithms, which were all initially extracted using either PCA or Gabor filtering-based approaches for face features, and MFCC for voice features. Here, the performance was evaluated for identification and verification. As such, the identification performance was evaluated using different standard evaluation measurements, including accuracy, precision, and F1 score, as reported in Tab. 2, which can be calculated using the following formulas [62]:

$$\text{Accuracy } = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \tag{18}$$

$$\text{Precision } = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{19}$$

$$\text{F1 score } = 2 \times \frac{\text{Precision } \times \text{ Recall}}{\text{Precision} + \text{Recall}}, \text{ where Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{20}$$

On the other hand, verification (authentication) performance was measured using appropriate standard evaluation metrics including receiver operator characteristic (ROC), where SVM was used for computing the likelihood as the conditional probability of each sample x [62] as follows:

$$ROC = \frac{P\left(x|positive\right)}{P\left(x|negative\right)} \tag{21}$$

Table 2: Identification performance comparison of dynamic fusion, unimodal and standard fusion

Metric	Feature	AR database				Extended Yale database			
		Unimodal	Standard fusion	Dynamic size	Dynamic weight	Unimodal	Standard fusion	Dynamic size	Dynamic weight
Accuracy	PCA	81.6	PCA & MFCC 87.3	PCA & MFCC 92.9	PCA & MFCC **95.7**	82	PCA & MFCC 82	PCA & MFCC 90	PCA & MFCC 94
	MFCC	49.3				49.3			
	Gabor	84.5	Gabor & MFCC 85.9	Gabor & MFCC 92.9	Gabor & MFCC 90.1	88	Gabor & MFCC 92	Gabor & MFCC **96**	Gabor & MFCC 96
Precision	PCA	82.9	PCA & MFCC 86	PCA & MFCC 90.8	PCA & MFCC **94.3**	87.8	PCA & MFCC 91.1	PCA & MFCC 97.3	PCA & MFCC 93.3
	MFCC	44.1				44.1			
	Gabor	82.3	Gabor & MFCC 82.3	Gabor & MFCC 89.9	Gabor & MFCC 89.2	96.2	Gabor & MFCC 91.1	Gabor & MFCC **97.6**	Gabor & MFCC 97.1
F1 score	PCA	82.2	PCA & MFCC 86.6	PCA & MFCC 91.9	PCA & MFCC **94.9**	84.8	PCA & MFCC 86.3	PCA & MFCC 93.5	PCA & MFCC 93.6
	MFCC	46.5				46.5			
	Gabor	81.6	Gabor & MFCC 84	Gabor & MFCC 91.3	Gabor & MFCC 89.6	91.9	Gabor & MFCC 91.5	Gabor & MFCC **96.7**	Gabor & MFCC 96.5

Moreover, area under the curve (AUC), and EER were deduced from ROC analysis for verification performance comparisons from different aspects, as reported in Tab. 3. EER refers to a common value of the false acceptance rate (FAR) and false rejection rate (FRR), where the FAR, FRR, and EER were calculated as [63]:

$$False\ AcceptanceRate = False\ Positive\ Rate = \frac{False\ Positive}{(False\ Positive + True\ Negative)} \tag{22}$$

$$False\ Rejection\ Rate = False\ Negative\ Rate = \frac{False\ Negative}{(False\ Negative + True\ Positive)} \tag{23}$$

$$Equal\ Error\ Rate = \frac{False\ Acceptance\ Rate + False\ Rejection\ Rate}{2} \tag{24}$$

Hence, the lower EER value indicates the higher performance of the system. In addition, AUC was calculated, as a measure that analyzes the verification performance with respect to each ROC curve by computing the area under it. Thus, the higher AUC value refers to the better performance, where the values close to 0.5 indicate low performance similar to random performance [64]. To prove the effectiveness of the proposed dynamic feature fusion framework, both identification and verification were conducted per database for the four different types of biometric feature vectors, explained in Section 4.4, then all tested methods were compared as summarized in Tabs. 2 and 3.

In overview, the experimental results show that the voice unimodal achieved the worst identification and verification performance, as 49.3% accuracy of identification and 0.308 EER of verification, while both unimodal face identification and verification achieved better results as Gabor-based achieved 84.5% accuracy with 0.0871 EER and 88% accuracy with 0.692 EER, when tested on AR and Extended Yale B, respectively. The PCA-based unimodal achieved about 82% accuracy of identification in both face databases, whereas for verification performance it received around 0.097 EER in both databases. The standard fusion scores show a slight performance improvement over the unimodal methods as the average accuracy improvement in identification was about 2.7% for both PCA-based and Gabor-based standard fusion methods. Furthermore, likewise, a slight performance improvement in verification is observed for the standard fusion methods over the unimodal methods, as illustrated in Figs. 13 and 14. Note that the proposed dynamic fusion methods were tested and compared to the other aforementioned baseline methods (i.e., unimodal and standard fusion methods).

5.1 Dynamic Size-Based Fusion of Face and Voice Biometrics

5.1.1 Dynamic Size-Based Fusion of Gabor and MFCC Traits

In this experiment, the proposed *dynamic size fusion* algorithm was applied using Gabor (face) and MFCC (voice) feature sets extracted for mostly low-quality and few high-quality probe data for both face and voice modalities. The experimental results show a remarkable enhancement in the performance of the *dynamic size fusion* method by about 7% in terms of accuracy, precision, and F1 score over the corresponding standard fusion method, when tested on AR and VOiCES databases, as can be deduced from Tab. 2. Accordingly, the *dynamic size fusion* method definitely outperforms the Gabor-based unimodal and MFCC-based unimodal in identification by 8.4% and 43.6%, respectively, as shown in Fig. 10.

Table 3: Face verification performance comparison of dynamic fusion, unimodal and standard fusion

Metric	Feature	AR database				Extended Yale database			
		Unimodal	Standard fusion	Dynamic size	Dynamic weight	Unimodal	Standard fusion	Dynamic size	Dynamic weight
EER	PCA	0.0971	PCA & MFCC 0.074	PCA & MFCC 0.041	PCA & MFCC **0.016**	0.0968	PCA & MFCC 0.078	PCA & MFCC 0.054	PCA & MFCC **0.017**
	MFCC	0.308	Gabor & MFCC 0.077	Gabor & MFCC **0.035**	Gabor & MFCC 0.050	0.308	Gabor & MFCC 0.064	Gabor & MFCC **0.017**	Gabor & MFCC 0.026
	Gabor	0.087				0.069			
AUC	PCA	0.953	PCA & MFCC 0.980	PCA & MFCC **0.994**	PCA & MFCC 0.990	0.976	PCA & MFCC 0.982	PCA & MFCC 0.989	PCA & MFCC **0.999**
	MFCC	0.781	Gabor & MFCC 0.977	Gabor & MFCC 0.986	Gabor & MFCC 0.986	0.781	Gabor & MFCC 0.991	Gabor & MFCC 0.998	Gabor & MFCC 0.997
	Gabor	0.957				0.968			

The overall identification and verification results of this method comes second in performance after dynamic weight fusion of PCA and MFCC features. Nevertheless, it achieves the best attained performance results on Extended Yale face B and VOiCES databases, which reaches up to 96% of identification accuracy. The other compared results emphasize the superiority of the first proposed *dynamic size fusion* algorithm, by which exceeding the accuracy of the standard fusion by 4%, the Gabor-based unimodal by 8%, and the MFCC-based unimodal by about 47%, as demonstrated in Fig. 11.

Correspondingly, the ROC performance of all examined Gabor-based approaches is compared in Figs. 13a and 14a, where the *dynamic size fusion* approach provides the best verification metric values of EER and AUC, as can be observed in Tab. 3, where the EER of the *dynamic size fusion* is 0.035 indicating less errors than the EER rate received as 0.077 by the standard fusion, using AR and VOiCES databases. Moreover, the EER of *dynamic size fusion* using the Extended Yale B database is the best and offers similar EER to the dynamic weight fusion of PCA and MFCC features, this enhancement due to the minimizing bad occlusion and noisy features. The reported results in Tab. 3, clarify the achieved improvement in verification using *dynamic size fusion* of Gabor and MFCC features.

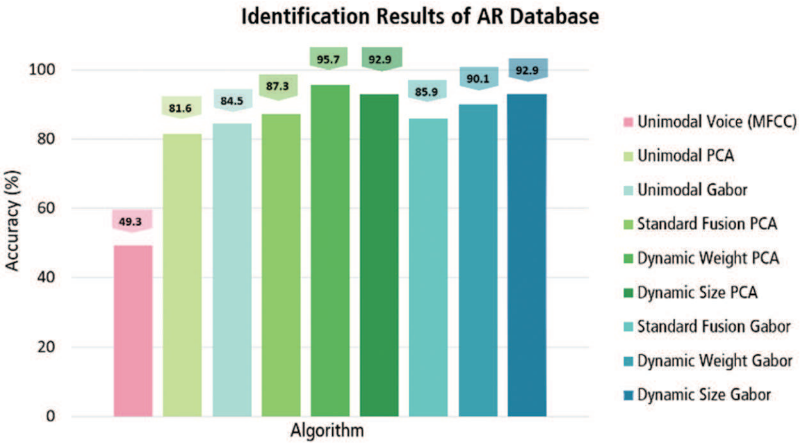


Figure 10: Face identification performance of standard and proposed methods on AR database

5.1.2 Dynamic Size-Based Fusion of PCA and MFCC Traits

In this experiment, the extracted PCA (face) and MFCC (voice) features were used to evaluate the first proposed method adopting *dynamic size fusion*. Here again the performance of *dynamic size fusion* of PCA and MFCC traits outperform the standard fusion of the same features by 5.6% on AR database, whilst the accuracy improvement on Extended Yale B reaches up to 8%, as shown in Figs. 10 and 11. These results signify consistent and similar performance enhancement to the results described for using the same *dynamic size fusion* method with Gabor features (Section 5.1.1).

Furthermore, the verification performance of this method achieves better results compared with the standard fusion and unimodal in terms of EER and AUC, as reported in Tab. 3. The *dynamic size fusion* achieves better EER and AUC than the standard fusion, and improve their scores by 0.033 and 0.014, respectively, on AR and VOiCES databases. On the other hand, the inferred EER when testing this method on Extended Yale B and VOiCES databases outperforms the standard fusion and unimodal by about 0.024 and 0.043, respectively. These results emphasize the superiority of *dynamic size fusion* over the standard fusion, and unimodal methods regardless of the database and the used feature extraction method, this is due to low-quality features exclusion or size minimization, by excluding whole occlusion features, and minimizing the size of noisy voice features. The achieved results shown in Figs. 13b and 14b. Fig. 12 shows few examples of correctly and incorrectly identified samples from different used databases and with different occlusion categories.

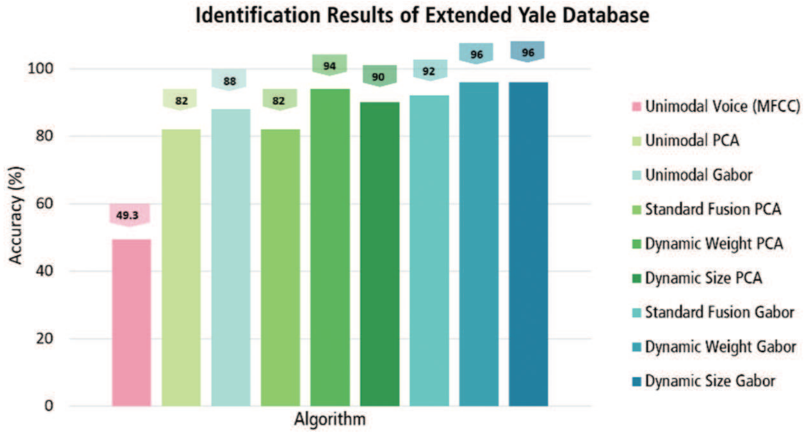


Figure 11: Face identification performance of standard and proposed methods on Extended Yale database

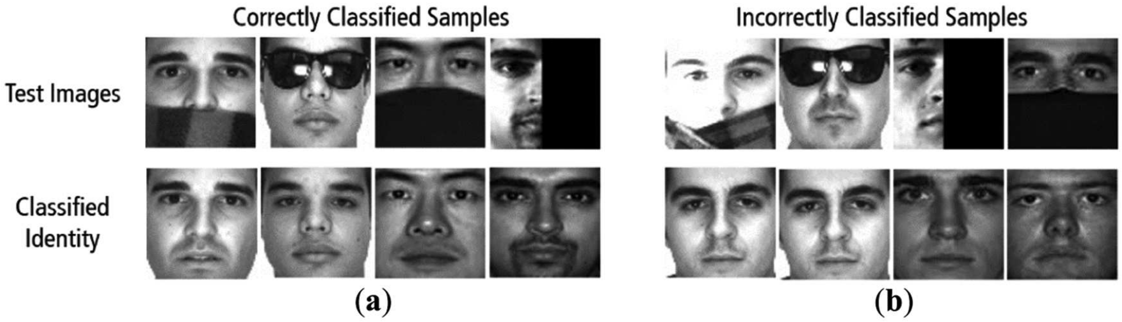


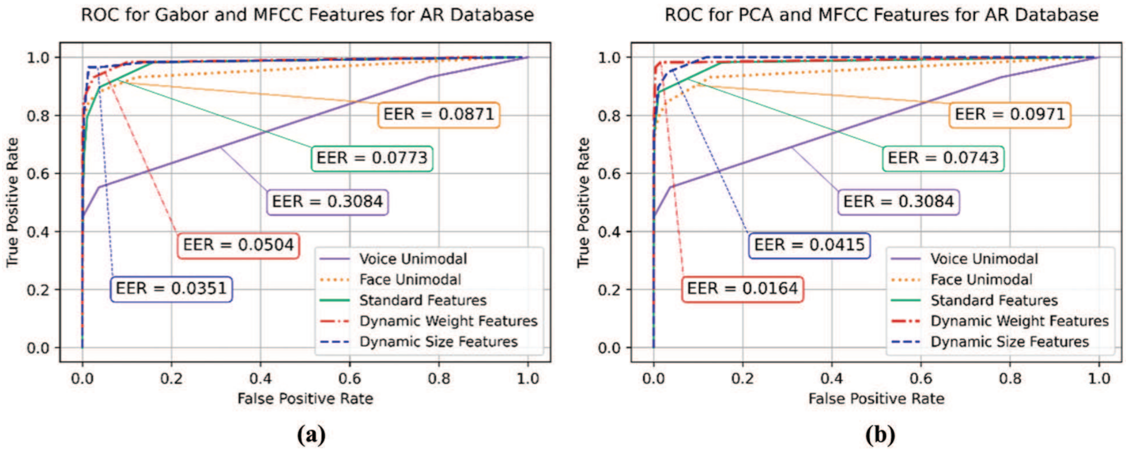
Figure 12: Examples of occluded face classification: (a) correct identification, (b) incorrect identification

5.2 *Dynamic Weight-Based Fusion of Face and Voice Biometrics*

5.2.1 Dynamic Weight-Based Fusion of Gabor and MFCC Traits

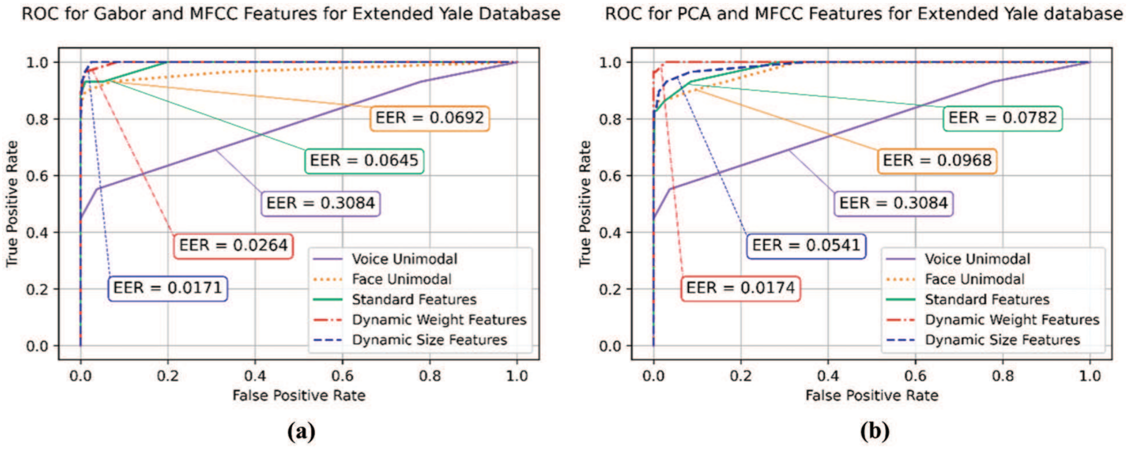
Here, the second proposed algorithm of *dynamic weight fusion* was experimented using Gabor (face) and MFCC (voice) features. As shown in Tab. 2 and Fig. 10, the results of this experiment on AR and VOiCES databases show a significant enhancement in identification performance over the standard fusion by 4.2% in terms of accuracy, 6.9% for precision, and 5.6% for F1 score. In addition, the identification result on the Extended Yale B and VOiCES databases offers the best attained results, reaching up to 96% of identification accuracy, as shown in Fig. 11. These results are similar to the results obtained by the first algorithm of dynamic size fusion, when examined on the same databases, with respect to accuracy. Although Gabor-based dynamic size fusion outperforms all other Gabor-based approaches, the overall identification performance of both dynamic fusion approaches is close on AR database and even much similar on Extended Yale B database in most comparable aspects.

Additionally, the verification results of this method using AR and VOiCES databases are improved, as shown in Fig. 13a, where this is presented as a reduction in EER value, such that the *dynamic weight fusion* provides the minimum EER value of 0.05, whilst the EER values of the standard fusion, the face unimodal, and voice unimodal in verification are worse larger values reported as 0.077, 0.087, and 0.308, respectively. The improved verification performance appears also as an increased AUC value, since the *dynamic weight fusion* achieves 0.986 AUC, whilst the AUC values achieved by the standard fusion and the face unimodal are 0.977 and 0.957, respectively. Similarly, the results of *dynamic weight fusion* for Extended Yale B show similar improvement for AUC and EER as illustrated in Tab. 3, and Fig. 14a.



(a) (b)

Figure 13: ROC verification performance of standard and proposed methods on AR database: (a) Gabor-based features, and (b) PCA-based features



(a) (b)

Figure 14: ROC verification performance of standard and proposed methods on Extended Yale B database: (a) Gabor-based features, and (b) PCA-based features

5.2.2 Dynamic Weight-Based Fusion of PCA and MFCC Traits

In this experiment, the second proposed *dynamic weight fusion* algorithm was tested for both identification and verification using PCA and MFCC features, see Tabs. 2 and 3. The performance evaluation on AR and VOiCES databases yields 95.7% accuracy as the best attained result, as shown in Fig. 10, by which outperforms the standard fusion and face unimodal by 8.4% and 14.1%, respectively.

As shown in Figs. 13b and 14b, the verification results of the same databases also show high performance by achieving the minimum-attained EER compared to all experimented approaches, which equals 0.016. Moreover, this method achieves 94% accuracy on the Extended Yale B and VOiCES databases, as shown in Fig. 11, which is also the best achieved result overall PCA-based methods using the Extended Yale B and VOiCES databases. This method consistently enhances identification with up to 12% over the standard fusion, which also represents the largest improvement and difference between standard and dynamic fusions. The ROC analysis in Fig. 14b shows comparable cures the methods based on PCA and MFCC feature set including the face/voice unimodal, the standard fusion, and both proposed dynamic fusion methods, when tested on Extended Yale B and VOiCES databases. The illustrated results show that the *dynamic weight fusion* achieves the best attained results, as shown in Fig. 13b on AR and VOiCES databases. The corresponding AUC results also confirm the superiority of the *dynamic weight fusion* by all means in verification performance. This can be observed as obvious increase in AUC values and decrease in EER values for both '*dynamic size fusion*' and '*dynamic weight fusion*'.

Finally, it can be noticed the overall performance of the *dynamic size fusion* is better when used with Gabor and MFCC features. This may be due to the large number of Gabor features compared with MFCC features, which might make Gabor face features more dominant than the other MFCC voice features in fusion and recognition as well. Finally, despite the achieved performance improvement when using dynamic fusion, the accurate selection of the weight and size values may have a great impact on the dynamic fusion performance.

6 Conclusion

In this work, we propose two data/information quality-based dynamic fusions at feature level, which are capable to improve person recognition performance of face and voice data based on dynamic multimodal biometric fusion. The first proposed fusion method adopts a dynamic size feature vector by excluding detected low-quality feature information, whilst the second proposed algorithm adopts dynamic weighting for detected low-quality feature information.

The experimental results show that both proposed dynamic fusion algorithms achieved high identification and verification performance under different realistic and synthetic low-quality data conditions adversely affecting face and voice biometrics. Moreover, multiple performance comparisons of the proposed dynamic fusion methods with other standard unimodal and multimodal fusion methods indicated remarkable improvements over the other standard counterparts. The overall obtained results show the dynamic size-based fusion performed better with Gabor and MFCC features, whereas the dynamic weight-based fusion achieved the best attained performance when using PCA and MFCC features.

The proposed dynamic biometric fusion framework provides promising results and likely potential solutions for several real life applications. The current nowadays situation of the COVID-19 pandemic is one good example of such applications, as it has posed many challenges to existing face and voice biometric recognition systems, which have been severely affected by global mask covering billions of people' noses and mouths and affecting their face looking and speech production. Therefore, there is an increasing need to develop accurate recognition systems and methods to be more suitable for use in such COVID-19 era.

Acknowledgement: The authors would like to thank King Abdulaziz University Scientific Endowment for funding the research reported in this paper.

Funding Statement: This research was supported and funded by KAU Scientific Endowment, King Abdulaziz University, Jeddah, Saudi Arabia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Deriche, "Trends and challenges in mono and multi biometrics," in *2008 First Workshops on Image Processing Theory, Tools and Applications*, Sousse, Tunisia, pp. 1–9, 2008.

[2] P. Sanjekar and J. Patil, "An overview of multimodal biometrics," *Signal & Image Processing*, vol. 4, no. 1, pp. 57, 2013.

[3] A. K. Jain, A. Ross and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[4] X. Zhang, D. Cheng, P. Jia, Y. Dai and X. Xu, "An efficient android-based multimodal biometric authentication system with face and voice," *IEEE Access*, vol. 8, pp. 102757–102772, 2020.

[5] M. Singh, R. Singh and A. Ross, "A comprehensive overview of biometric fusion," *Information Fusion*, vol. 52, no. 1, pp. 187–205, 2019.

[6] N. Larbi and N. Taleb, "A robust multi-biometric system with compact code for iris and face," *International Journal on Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 1–13, 2018.

[7] A. A. Ross and R. Govindarajan, "Feature level fusion of hand and face biometrics," in *Biometric Technology for Human Identification II*, Vol. 5779, pp. 196–204, 2005.

[8] H. S. Bhatt, S. Bharadwaj, M. Vatsa, R. Singh, A. Ross *et al.,* "A framework for quality-based biometric classifier selection," in *2011 Int. Joint Conf. on Biometrics (IJCB)*, Washington, USA, pp. 1–7, 2011.

[9] M. Vatsa, R. Singh, A. Noore and A. Ross, "On the dynamic selection of biometric fusion algorithms," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 470–479, 2010.

[10] K. Gupta, G. S. Walia and K. Sharma, "Quality based adaptive score fusion approach for multimodal biometric system," *Applied Intelligence*, vol. 50, no. 4, pp. 1086–1099, 2020.

[11] S. Park, H. Lee, J. H. Yoo, G. Kim and S. Kim, "Partially occluded facial image retrieval based on a similarity measurement," *Mathematical Problems in Engineering*, vol. 2015, no. 1, pp. 1–11, 2015.

[12] W. Wan and J. Chen, "Occlusion robust face recognition based on mask learning," in *2017 IEEE Int. Conf. on Image Processing (ICIP)*, Beijing, China, pp. 3795–3799, 2017.

[13] T. Mahboob, M. Khanum, M. S. H. Khiyal and R. Bibi, "Speaker identification using gmm with mfcc," *International Journal of Computer Science Issues (IJCSI)*, vol. 12, no. 2, pp. 126, 2015.

[14] N. Damer, F. Boutros, M. Süßmilch, M. Fang, F. Kirchbuchner *et al.,* "Masked face recognition: Human vs. machine," *arXiv,* vol. 2103.01924, 2021.

[15] M. Gomez-Barrero, P. Drozdowski, C. Rathgeb, J. Patino, M. Todisco *et al.,* "Biometrics in the era of COVID-19: Challenges and opportunities," *arXiv,* vol. 2102.09258, 2021.

[16] Z. Jianxin and W. Junyong, "Local occluded face recognition based on HOG-LBP and sparse representation," in *2020 IEEE Int. Conf. on Artificial Intelligence and Computer Applications (ICAICA)*, Dalian, China, pp. 808–813, 2020.

[17] Y. Li, K. Guo, Y. Lu and L. Liu, "Cropping and attention based approach for masked face recognition," *Applied Intelligence*, vol. 51, no. 5, pp. 3012–3025, 2021.

[18] M. Sharma, S. Prakash and P. Gupta, "An efficient partial occluded face recognition system," *Neurocomputing*, vol. 116, no. 12, pp. 231–241, 2013.

[19] J. Wu, Z. Shang, K. Wang, J. Zhai, Y. Wang *et al.,* "Partially occluded head posture estimation for 2D images using pyramid HoG features," in *2019 IEEE Int. Conf. on Multimedia & Expo Workshops (ICMEW)*, Shanghai, China, pp. 507–512, 2019.

[20] L. Song, D. Gong, Z. Li, C. Liu and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, South Korea, pp. 773–782, 2019.

[21] A. Harriero, D. Ramos, J. Gonzalez-Rodriguez and J. Fierrez, "Analysis of the utility of classical and novel speech quality measures for speaker verification," in *Int. Conf. on Biometrics*, Alghero, Italy, pp. 434–442, 2009.

[22] Y. Elmir, O. Ghazaoui and F. Boukenni, "Multimodal biometrics system's resistance to noise," in *First National Conf. on Computer Science and Information and Communication Technologies (CTIC)*, Adrar, Algeria, pp. 25–28, 2012.

[23] A. Ashar, M. S. Bhatti and U. Mushtaq, "Speaker identification using a hybrid CNN-MFCC approach," in *2020 Int. Conf. on Emerging Trends in Smart Technologies (ICETST)*, Shah Faisal Town, Pakistan, pp. 1–4, 2020.

[24] Y. Kawakami, L. Wang and S. Nakagawa, "Speaker identification using pseudo pitch synchronized phase information in noisy environments," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conf.*, Kaohsiung, Taiwan, pp. 1–4, 2013.

[25] P. Dhakal, P. Damacharla, A. Y. Javaid and V. Devabhaktuni, "A near real-time automatic speaker recognition architecture for voice-based user interface," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 504–520, 2019.

[26] A. K. H. Al-Ali, D. Dean, B. Senadji, V. Chandran and G. R. Naik, "Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions," *IEEE Access*, vol. 5, pp. 15400–15413, 2017.

[27] P. Byahatti and M. S. Shettar, "Fusion strategies for multimodal biometric system using face and voice cues," in *IOP Conf. Series: Materials Science and Engineering*, Vol. 925, IOP Publishing, pp. 012031 2020.

[28] A. Abozaid, A. Haggag, H. Kasban and M. Eltokhy, "Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion," *Multimedia Tools and Applications*, vol. 78, no. 12, pp. 16345–16361, 2019.

[29] C. Dalila, B. Saddek and N. A. Amine, "Feature level fusion of face and voice biometrics systems using artificial neural network for personal recognition," *Informatica*, vol. 44, no. 1, pp. 85–96, 2020.

[30] F. Alonso-Fernandez, J. Fierrez, D. Ramos and J. Gonzalez-Rodriguez, "Quality-based conditional processing in multi-biometrics: Application to sensor interoperability," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 6, pp. 1168–1179, 2010.

[31] N. Poh and J. Kittler, "A unified framework for biometric expert fusion incorporating quality measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 3–18, 2012.

[32] M. Gofman, S. Mitra, K. Cheng and N. Smith, "Quality-based score-level fusion for secure and robust multimodal biometrics-based authentication on consumer mobile devices," in *Int. Conf. on Software Engineering Advances (ICSEA)*, Barcelona, Spain, pp. 274–276, 2015.

[33] H. Sellahewa and S. A. Jassim, "Image-quality-based adaptive face recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 805–813, 2010.

[34] M. Welvaert and Y. Rosseel, "On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data," *PLoS One*, vol. 8, no. 11, pp. e77089, 2013.

[35] Z. Lijun, S. Xiaohu, Y. Fei, D. Pingling, Z. Xiangdong *et al.,* "Multi-branch face quality assessment for face recognition," in *2019 IEEE 19th Int. Conf. on Communication Technology (ICCT)*, China, IEEE, pp. 1659–1664, 2019.

[36] S. Bharadwaj, M. Vatsa and R. Singh, "Biometric quality: A review of fingerprint, iris, and face," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, pp. 1–28, 2014.

[37] R. Min, A. Hadid and J. L. Dugelay, "Efficient detection of occlusion prior to robust face recognition," *Scientific World Journal*, vol. 2014, no. 3, pp. 1–10, 2014.

[38] G. N. Priya and R. W. Banu, "Occlusion invariant face recognition using mean based weight matrix and support vector machine," *Sadhana*, vol. 39, no. 2, pp. 303–315, 2014.

[39] H. Ng, S. Ong, K. Foong, P. S. Goh and W. Nowinski, "Medical image segmentation using k-means clustering and improved watershed algorithm," in *2006 IEEE Southwest Symp. on Image Analysis and Interpretation*, Colorado, USA, IEEE, pp. 61–65, 2006.

[40] A. S. Kornilov and I. V. Safonov, "An overview of watershed algorithm implementations in open source libraries," *Journal of Imaging*, vol. 4, no. 10, pp. 123, 2018.

[41] A. S. Abdulaziz and V. Kpuska, "The short-time silence of speech signal as signal-to-noise ratio estimator," *International Journal of Engineering Research and Applications (IJERA)*, vol. 8, pp. 99–103, 2016.

[42] J. Li, X. Zhang, C. Jia, J. Xu, L. Zhang *et al.,* "Universal adversarial perturbations generative network for speaker recognition," in *2020 IEEE Int. Conf. on Multimedia and Expo (ICME)*, London, United Kingdom, IEEE, pp. 1–6, 2020.

[43] A. M. Martinez, "The AR face database," *CVC Technical Report*, vol. 24, pp. 8, 1998.

[44] A. S. Georghiades, P. N. Belhumeur and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[45] K. C. Lee, J. Ho and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[46] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco *et al.,* "Voices obscured in complex environmental settings (voices) corpus," *arXiv,* vol. 1804.05053, 2018.

[47] D. Zeng, R. Veldhuis and L. Spreeuwers, "A survey of face recognition techniques under occlusion," *arXiv,* vol. 2006.11366, 2020.

[48] W. Zheng, C. Gou and F. Y. Wang, "A novel approach inspired by optic nerve characteristics for few-shot occluded face recognition," *Neurocomputing*, vol. 376, no. 4, pp. 25–41, 2020.

[49] A. Priadana and M. Habibi, "Face detection using haar cascades to filter selfie face image on instagram," in *2019 Int. Conf. of Artificial Intelligence and Information Technology (ICAIIT)*, Yogyakarta, Indonesia, pp. 6–9, 2019.

[50] A. Obukhov, "Haar classifiers for object detection with cuda," in *GPU Computing Gems Emerald Edition*, 1st ed., Burlington, USA: Elsevier, pp. 517–544, 2011.

[51] E. S. Jaha, "Augmenting Gabor-based face recognition with global soft biometrics," in *2019 7th Int. Symp. on Digital Forensics and Security (ISDFS)*, Kerala, India, pp. 1–5, 2019.

[52] M. T. Rahman and M. A. Bhuiyan, "Face recognition using gabor filters," in *2008 11th Int. Conf. on Computer and Information Technology*, Khulna, Bangladesh, IEEE, pp. 510–515, 2008.

[53] E. S. Jaha, "Efficient Gabor-based recognition for handwritten Arabic-Indic digits," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, pp. 112–120, 2019.

[54] B. A. Draper, K. Baek, M. S. Bartlett and J. R. Beveridge, "Recognizing faces with PCA and ICA," *Computer Vision and Image Understanding*, vol. 91, no. 1–2, pp. 115–137, 2003.

[55] R. Lionnie and M. Alaydrus, "Biometric identification system based on principal component analysis," in *2016 12th Int. Conf. on Mathematics, Statistics, and Their Applications (ICMSA)*, Banda Aceh, Indonesia, pp. 59–63, 2016.

[56] H. M. Maw, K. Z. Lin and M. T. Mon, "Evaluation of face recognition techniques for facial expression analysis," in *Int. Conf. on Intelligent Computing, Communication & Convergence (ICCC-2015)*, Odisha, India, 2015.

[57] H. M. Ebied, "Feature extraction using PCA and Kernel-PCA for face recognition," in *2012 8th Int. Conf. on Informatics and Systems (INFOS)*, Giza, Egypt, pp. MM-72–MM-77, 2012.

[58] A. Zulfiqar, A. Muhammad and M. E. AM, “A speaker identification system using MFCC features with VQ technique,” *2009 Third Int. Symp. on Intelligent Information Technology Application*, vol. 3, pp. 115–118, 2009.

[59] M. Pal and 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.

[60] Z. Rustam and S. Kharis, “Comparison of support vector machine recursive feature elimination and kernel function as feature selection using support vector machine for lung cancer classification,” *Journal of Physics: Conf. Series*, vol. 1442, no. 1, pp. 012027, 2020.

[61] M. Mohammadi, T. A. Rashid, S. H. T. Karim, A. H. M. Aldalwie, Q. T. Tho *et al.,* “A comprehensive survey and taxonomy of the SVM-based intrusion detection systems,” *Journal of Network and Computer Applications*, vol. 178, no. 4, pp. 102983, 2021.

[62] M. Sokolova, N. Japkowicz and S. Szpakowicz, “Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation,” in *Australasian Joint Conf. on Artificial Intelligence*, Hobart, Australia, pp. 1015–1021, 2006.

[63] R. Vyas, T. Kanumuri and G. Sheoran, “Iris recognition using 2-D Gabor filter and XOR-SUM code,” in *2016 1st India Int. Conf. on Information Processing (IICIP)*, Delhi, pp. 1–5, 2016.

[64] R. Jahangir, Y. W. Teh, N. A. Memon, G. Mujtaba, M. Zareei *et al.,* “Text-independent speaker identification through feature fusion and deep neural network,” *IEEE Access*, vol. 8, pp. 32187–32202, 2020.