Vol. 06, No. 3 (2024) 1047-1056, doi: 10.24874/PES06.03.018

Proceedings on Engineering Sciences

www.pesjournal.net

A CONTACTLESS SPEAKER IDENTIFICATION APPROACH USING FEATURE-LEVEL FUSION OF SPEECH AND FACE CUES WITH DCNN

Khushboo Jha[1]
Aruna Jain
Sumit Srivastava

Received 01.08.2023.
Received in revised form 21.10.2023.
Accepted 26.12.2023.
UDC – 004.032.26

Keywords:

Deep Convolutional Neural Network, Data Fusion Methods, Dimensionality Reduction, Feature Extraction Techniques, Feature Level Fusion, Multimodal Biometric System

ABSTRACT

This paper evaluates the effectiveness of feature-level fusion through the concatenatio method, of two independent and emerging modalities, speech and face. The majo benefit of face modality (physiological) is that the data acquisition does not requir much user cooperation or awareness, as seen in airports or public places in mass Speech (physiological and behavioural) based recognition, for disabled and illiterat people, is the most convenient and reliable user identification technique due to the eas with which a contactless speech-receiving device can be accessed. Furthermore, should be noted that adverse conditions, such as low illumination for facial recognitio and a noisy environment for speech recognition during data acquisition, are nc interdependent and function autonomously. Consequently, the acoustic and distinctiv facial features are the paramount (fused) features in achieving higher use identification accuracy. This paper aims to explore the state-of-the-art techniques fc data fusion, dimensionality reduction, feature extraction (speech-face) and classifie Based on the above findings, we have proposed an efficient feature level fusion c speech and face cues with the deep convolutional neural network as a classifier for th VidTIMIT database. We have tested the effectiveness of the proposed approach in term of identification accuracy with different training sample sizes and numbers of users The proposed user identification approach achieves an accuracy of 97.31%, an EER c 3.62% and outperforms the unimodal biometric system for speech and face by 3.83 and 1.59 % respectively. Additionally, the proposed approach outperformed a fe existing methodologies. Thus, we can infer that even in the presence of advers conditions, such an approach can ameliorate the user identification-based solution.

1. INTRODUCTION

Features excerpted from biometric cues are a rich source of information. Moreover, the amalgamation of features permits classes to be highly separable, thereby boosting the efficiency and accuracy of the multimodal biometric system (MBS) (Abozaid et al., 2019; Oloyede & Hancke, 2016). Feature-level fusion has the benefit of monitoring correlated features obtained from several biometric methods. Such that prominent feature sets are identified to expedite the recognition accuracy. By integrating different biometric traits through a fusion approach, an MBS can significantly decrease the overlap between the feature spaces of different users (inter-class similarities).

[1] Corresponding author: Khushboo Jha
Email: kjha.phd@gmail.com

Moreover, the fused feature vectors result in a more powerful and reliable person recognition system which is difficult to forge or spoof. Among different types of fusion (Ryu et al., 2021) (such as sensor level, rank level, feature level, score level, and decision level), the most expedient is feature-level fusion. The features for fusion can be from the same modality, such as infrared and colour images of the face or from different modalities, such as speech (audio) and face (image) features. It combines features to enhance classification (Agrawal et al., 2016) and to reduce data dimensionality. Moreover, melding image-audio features emanate in a stronger and more robust recognition system. Thereby decimating the limitation of password-based as well as unimodal biometric systems (UBS) and flourishing as a trend to meet stringent security standards.

The face modality has the ability to gather data at airports and other public places without requiring the user's explicit agreement, which is its primary benefit. Speech-based recognition, particularly for those with impairments, is advantageous and reliable due to its non-contact nature and accessibility. Additional modalities necessitate user collaboration and close proximity to the device. An effective MBS should use biometric cues or modalities that can offset the weaknesses of one another in order to enhance overall performance. For example, if a legitimate user cannot be recognised by their speech or face cue due to environmental issues or low illumination respectively, then they can still be identified through another modality such as speech or face and vice versa. In nutshell, speech and face modalities together pair up to be the most reliable, hence the fused feature will result a robust and an efficient MBS, figure 1 represents a general model of a such MBS.

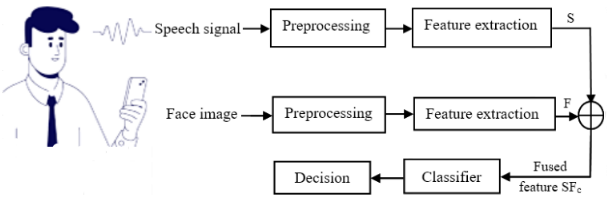

Figure 1. Multimodal biometric system.

The subsequent sections of this paper are structured in the following manner. Section 2 offers a comprehensive elucidation of data fusion techniques specifically focusing on the feature level. Section 3 provides a comprehensive overview of various dimensionality reduction techniques. Subsequently, the subsequent sections, namely Section 4 and Section 5, delve into the techniques employed for extracting features pertaining to speech and face modalities. Section 6 provides a comprehensive description of the experimental setup, encompassing the database utilised, state-of-the-art techniques employed for feature extraction, the fusion method employed, and the classifier utilised. Section 7 provides an exposition of the findings and a comprehensive analysis thereof. The conclusion of this study is presented in Section 8.

2. DIFFERENT METHODS FOR DATA FUSION AT FEATURE LEVEL

In any MBS, features from multiple sensors or biometric cues can be fused into a single feature vector using any of the methods (Nguyen et al., 2018; Dalila et al., 2020) as mentioned below. Here we consider two biometric cues, speech and face, where S=[S1…..Sn] be the speech feature vector and F= [F1……Fn] be the face feature vector then:

2.1 Merging

Data fusion is performed by placing one face feature followed by one speech feature unless all the features are placed alternatively as mentioned below:

$$SFm = [S1F1 \dots \dots \dots Sn\ Fn]. \qquad (1)$$

2.2 Multiplication

Data fusion is performed by multiplying pre-normalized speech and face features, as mentioned below.

$$SF_{mp} = [S_1{*}F_1 \dots \dots \dots S_n{*}F_n]. \qquad (2)$$

2.3 Bilinear pooling

These models (Nguyen et al., 2018) calculate the outer product of say two vectors $f1 \in Rn1$ and $f2 \in Rn2$ and result in a model W (linear), i.e.,

$$Z = W[f1 \otimes f2], \qquad (3)$$

Here $\otimes$ denote the outer product and [ ] as a linearization of the matrix in a vector. Such data fusion methods outperform simple fusion and show effectiveness in melding two vectors (say audio-visual) as it permits every element of all vectors to link with other in a multiplicative way. But, the outer product evaluation on the high-dimensional vectors such as n1 and n2 results in innumerable parameters, which makes them infeasible for subsequent learning in W.

2.4 Concatenation

It serves as an excellent litmus test since it is the simplest fusion method for constructing a joint representation of feature vectors from multiple modalities and quickly determining the effectiveness of the new approach.

$$FVc = [F1 \dots \dots \dots Fn\ V1 \dots \dots \dots Vn] \qquad (4)$$

For this reason, concatenation is the baseline for most research work and ours too. Data fusion is performed by appending the two feature vectors as mentioned above.

3. DIMENSIONALITY REDUCTION TECHNIQUES

As a consequence of data fusion at the feature level, the resultant feature vectors with larger dimensionality often led to the curse of dimensionality problem (Mehraj & Mir, 2021) i.e., the error increases with the increase in the dimension. Moreover, working with high dimensional data, it becomes very difficult to identify meaningful features and it also deteriorates the accuracy as well as the computational speed of the recognition system. Given the disparity in dimensionality, performing dimensionality reduction (DR) helps the combination by purging the low informative portion of every embedding. Therefore, to decimate this issue, DR techniques (Mehraj & Mir, 2021) serve two purposes; smoothen the neural network result over the training dataset to permit a significant generalization and feature fusion of different cues. A few techniques are discussed below:

3.1 Truncated Singular Value Decomposition (TSVD)

SVD (Neto et al., 2023) is a way of matrix analysis that represents a high-dimensional matrix as low-dimensional. When applied to a matrix, SVD can help us build a faithful representation by filtering out irrelevant details. Because of this, it generates a close approximation of the presentation in the number of dimensions specified. For M be an m x n matrix having rank r as:

$$M = U\sum Vt\ , \qquad (5)$$

Where $\sum$ is a diagonal matrix while U is an orthogonal m x r matrix. Singular values of M are samples of $\sum$ and V is an orthogonal n x r matrix. Only the columns of U with respect to the t greatest singular values are useful when employing SVD for DR because the remaining are discarded in the reduction process. Finding the t columns of U that correlate to the highest solitary values is thus more cost-effective. A technique called Truncated SVD (Freire-Obregon et al., 2021) can be used to accomplish this. Since the original matrix is imperfectly decomposed, the DR can be carried out significantly more quickly than the standard SVD.

3.2 t-Distributed Stochastic Neighbour Embedding (t-SNE)

t-SNE (Meyer et al., 2022) is a multifaceted approach that is often used for visualising high-dimensional information generated using a CNN. It seeks to discover local and global trends in data by grouping similar data together and pushing distant samples away. Perplexity is a significant t-SNE hyper-parameter. This hyper-parameter specifies the amount of global structure must be kept over the local structure. Whilst t-SNE is useful for reducing dimensionality in existing data, it cannot modify new data. It is because t-SNE does not train a transformation function; rather, directly improves the low-dimensional representation.

3.3 Uniform Manifold Approximation and Projection (UMAP)

UMAP (Oskolkov N., 2022) like t-SNE, is a manifold approach. Its purpose is to discover a transformation function that can convert high-dimensional data to a low-dimensional plane. The number of neighbours used when creating the weighted graph and the minimum distance between data points are the most essential hyper-parameters for UMAP. Lowering (or eliminating) the minimum distance lowers the limits on point placement in terms of distance between neighbours. In the low-dimensional plane, a greater minimum distance usually results in more sparse data points. UMAP is suitable for transforming fresh data points since it indirectly optimises the embedding through the optimization of the weighted graph.

3.4 Principal Component Analysis (PCA)

It is an extensively used DR technique, which constructs new attributes known as principal components (PCs), to linear groupings of the original variables. These PCs must satisfy a few criteria: linear combinations of the original parameters that are orthogonal to each other and capture a huge amount of data variance. Generally, the data variance is captured by less number of PCs, as a consequence PCA (Dalila et al., 2020; Freire-Obregon et al., 2021) can accomplish large DR with low noise than the given input. Thus, it is regarded as the state-of-the-art technique.

4. SPEECH MODALITY

Speech being a physiological (e.g., vocal tract shape, larynx size, et.) as well as behavioural (e.g., speaking style, pronunciation pattern, etc.) biometric trait, is the most natural, non-invasive, contactless and hence widely accepted fundamental way of distinguishing person among all biometric traits. Speech based systems can equip a smart home with a speaker recognition (SR) (Kinnunen & Li 2010; Jha et al., 2023a) system that includes features to assist visually challenged and senior citizens in controlling devices. Therefore, utilizing robust and distinct acoustic features, the SR system aims to extract, analyse, characterize, and identify information about the speaker's identity.

4.1 Feature extraction of speech signal

The conversion of a raw speech signal into a series of acoustic feature vectors that include speaker-specific details is called feature extraction. It is a crucial step, as it strongly influences the recognition rate. The linear predictive coding (LPC) (Srivastava et al., 2019) technique reflects the vocal tract and extracts robust features from low-bit-rate audio signals. The perceptual linear predictive (PLP) technique, works best in low-dimensional analysis and the disparity between voiced and unvoiced input is lessened. It is superior to LPC, as it efficaciously suppresses speaker-dependent information. The linear prediction cepstral coefficients (LPCC) technique is obtained from the LP method of the speech signal. Also, the cepstral analysis eliminates data with significant similarity and is more reliable than the LPC method. The mel-frequency cepstral coefficients (MFCC) (Srivastava et al., 2019) technique, has been the most preferred acoustical feature as it is based on the known fluctuation of the human ear's significant bandwidth with frequency. It represents spectral characteristics of an utterance of speech, which is one of the prominent features same as the human auditory perception system. But it is not resilient to noise. Moreover, the spectral-temporal receptive fields (STRF) based features when fused with conventional MFCCs (Wang et al., 2017), resulting in a robust and efficient acoustical feature for the SR system. The STRF concept is based on the physiological model of the mammalian auditory system. It is responsible for the generation of the cortical response, which is then transformed into effective STRF features. Thus, we suggest STRF-MFCCs (Wang et al., 2017) features as an effective and robust acoustical feature set that will work efficiently even in noisy environments. Also, in this work, we recommend a text-independent SR system (Kinnunen & Li 2010; Jha et al., 2023a), since the user need not remember the passcode uttered during the training phase i.e., it is not constrained by text and hence more efficient than the text-dependent system.

MFCC includes the following steps:

It starts with sliding the analysis window including the speech signal. The average energy around the center frequency of each triangle is then computed using a filter bank of triangular weighted filters. The distribution of filters is based on a Mel scale that mimics the behaviour of the human auditory system. The mel-scale is given by:

$$\mathrm{Mel}(f)=2595^{*}\log_{10}(1+(f/700)) \qquad (6)$$

Here, f denotes the frequency. Lastly, MFCCs are specified as the Discrete Cosine Transform (DCT) for the logarithmic filter bank energies.

STRF includes the following steps:

It is a multi-resolution analysis covering temporal and spectral area. It is a mathematical auditory assessment approach, motivated by psycho-acoustical as well as neuro-physiological research on the central and early phases of the mammalian auditory system. The low and high scales in STRF are used to represent the formant and harmonics, respectively. Using the above scale information, STRF oriented feature is generated. The frequency band of the auditory spectrum is represented by scale values ranging from 2^{-3} to 2^{3}, with periods of 0.5 cycle per octave. For a given audio spectrogram of the form y(t, f), we can derive the cortical representation of the spectral-temporal response STRF(t, f, Ω, ω, φ, θ), as follows:

$$\mathrm{STRF}(t, f,\Omega, \omega,\varphi,\theta)=y\,(t, f)*_{t\,f}[h_S(f, \omega, \theta)\cdot h_T(t, \Omega,\varphi)] \quad (7)$$

Here $*_{tf}$ denotes the convolution operation in the frequency and time domain. At first, S(t, ω) is generated by adding all the frequencies and rates in the STRF magnitude notation.

$$S(t, w) = \textstyle\sum_f \sum_\Omega |\mathrm{STRF}\,(t, f, \Omega, w, 0, 0)|, \qquad (8)$$

Where N_w is the scale value and w = 1,2,...N_w. The characteristic phases θ and φ in Eq.(8) are assigned to zero. Then, a non-linear logarithmic operation is applied to S(t, ω) to produce $S_L(t, \omega)$.

$$S_L(\,t, \omega) = \log\,(S\,(t, \omega)), \quad \omega = 1, 2, \ldots N_\omega. \qquad (9)$$

When DCT is employed to $S_L(t, \omega)$ we get the required STRF based feature as $S_{DL}(t, k)$ mentioned below.

$$s_{DL}\,(\mathrm{t,k}) = \sum_{w=1}^{Nw} S_L(\,t,\omega)\cos = 1,2,\ldots N_k \qquad (10)$$

where N_k is equal or smaller to N_ω is considered as feature dimension of $S_{DL}(t, k)$. In nutshell, 13 features of S_{DL}-MFCC are recommended as the robust features set for SR.

5. FACE MODALITY

Face recognition (FR) (Jha et al., 2023b; Taskiran et al., 2020) is a biometric method that can recognise a person from a digital photograph or a video frame captured by an external source. It evolved as an expedient biometric modality for individual identification, analyzing facial features' unique shape, pattern, and positioning. The face has distinctive landmarks, approximately 80 nodal points, i.e., peaks and valleys that comprise the various facial features. These nodal points are evaluated to generate a numeric code called face print, a string of numbers corresponding to a face in the database. However, only 12 to 22 nodal points are practical to accomplish the recognition process, such as distance between the eyes, jawline, cheekbones, depth of eye socket, etc. FR is advantageous in applications such as video surveillance, human-computer interface, identity validation for crime detection, social welfare, helps in mass identification, computer security, auto-screening at airports, ATM access, etc. An advantage of using the facial cue, as opposed to other methods, is that data can be collected (via photos or video) without requiring the user's consent or awareness, such as in airports or other public locations. Whilst, other cues or biometric traits (fingerprint, iris, retina, etc.), for surveillance cannot be acquired without the user's collaboration.

Thus, FR systems benefit from being a non-intrusive, passive method of verifying personal identity naturally and courteously. In order to achieve optimal performance and reliability in facial feature recognition, it is imperative to address various challenges such as low illumination, pose variation, and emotional expressions. Images captured under low-illumination conditions possess significant facial characteristics that hold valuable information and thus should not be disregarded. Hence, we propose the utilisation of Dynamic Histogram Equalisation (DHE) (Abdullah-Al-Wadud et al., 2007) as a method to enhance the illumination of each raw face image. This technique aims to improve the quality of the face image by mitigating any potential adverse effects on the original input image.

5.1 DHE technique for enhanced illumination

This technique has three aspects as discusses below:
Partitioning – It divides the image histogram into sub-histograms unless no dominant section can be found in any of these sub-histograms.

Allocating Gray Level (GL) – The decisive criterion for GL range of ith sub-histogram for input image is determined as:

$$\text{factor}_i = \text{span}_i * (\log F_i)^x \quad (11)$$

where, $\text{span}_i = m_{i+1} - m_i$, denotes dynamic GL range and F_i is the total frequency of gray levels in ith sub-histogram. The value of x indicates how much focus should be placed on Fs. For grubby images, the dynamic range of gray levels is minimal. For such case, it is acceptable to use span alone (x=0).

Applying histogram equalization (HE) – Compared to conventional HE, DHE avoids any region of the image from being over or under enhanced. Allocation of non-overlapping grey level ranges to sub-histograms ensures that no two GLs from the distinct sub-histograms map to the same GL, avoiding considerable loss of image features. The consecutive assignment and independence from the dominance of any section guarantee that there is no uncomfortable leap in nearby GLs. By using DHE, the output histogram is guaranteed to include only one mapping for each grey level, i.e., it ensures no obstruction effect. With the increase of x, the low illumination image gets enhanced, increased brightness and sharpening the edges without causing any distortions

5.2 Feature extraction of speaker’s face

The aforesaid enhanced speaker’s facial image is fed into the feature extraction module to extract robust feature for face recognition. The scale invariant feature transform (SIFT) (Ketab et al., 2023) technique is used for extracting image characteristics. The SIFT feature is highly distinct in nature, achieving accurate matching on maximum pairs of feature points having greater probability between a huge database and the test data. As a result, it is mathematically complicated and computationally heavy. The Histogram of oriented gradient (HOG) (Jha et al., 2023b) is a type of feature descriptor used to extract features from image data and detect small-scaled images with less computational power.

But in the case of large-scale images, computational speed is low. The local binary patterns (LBP) (Jha et al., 2023b) technique is a robust texture features extraction technique with the goal to summarise an image's local structure by comparing each pixel to its surroundings. But susceptible to surface issues such as blurring. The Eigen-face (Mahmood & Kurnaz, 2023) approach is simple and performs well in a constrained environment with almost nil illumination variations i.e., it is sensitive to illumination differences. The Gabor features (Yaermaimaiti & Kari, 2023) are less vulnerable to changes in lighting, expression and pose than holistic features like Eigen face. While the Gabor filter is effective at extracting features, it is time consuming as well as sensitive to non-linear distortions and rotations. The aforesaid traditional feature extraction techniques, such as SIFT, HOG, LBP, Gabor, etc. have limitations in precision and often achieve unsatisfactory recognition results.

Active Appearance Model (AAM)

With the advent of AAM (Ning et al., 2023) the accuracy and speed of FR have made great strides. The state-of-the-art AAM is one of the most effective model-assisted objects tracking as well as feature extraction algorithm. In particular, the AAM creates texture and shape of the object (20 landmark points) to form a collection of instantaneous and real images (i.e., detailed texture feature based on pattern of intensity and color). The difference among a new image and one that has been synthesized is minimized by the AAM model, which treats interpretation as an optimization issue. The difference vector δI is defined in equation below:

$$\delta I = I_i - I_m \quad (12)$$

where I_i depicts the grey level value vector in the preprocessed input face image I_p^F and I_m depicts grey level vector of present model parameter. By adjusting the model parameter, we can find the model and image I_p^F feature that matches each other the best by minimizing the size of difference vector as $\Delta = \delta I^2$.

6. EXPERIMENTAL SETUP

The audio-image fusion approach can significantly reduce the overlap between the feature spaces of different users (inter-class similarities). For illustration, two people of the same family, can have the same facial feature but would not have the same speech trait, and vice versa. Also, the audio-image combination is unaffected by each challenge such as the acquisition of audio in low illumination or the acquisition of facial image in a noisy environment. The data acquisition of speech-face cues can be done concurrently in less time by using a camera with an embedded microphone.

Hence, the above combination outperforms others in terms of being non-invasive, user-friendly, and contactless entailed to resist the spread of the COVID-19 (Corona virus disease 2019 and its variants) pandemic and high user acceptance, especially for disabled people. As a corollary, we have chosen a combination of speaker's speech and face cues for our research work. A systematic block diagram for the proposed multimodal biometric identification system (Abozaid et al., 2019; Dalila et al., 2020), consisting of the training and testing phase is shown in figure 2. In the training phase of the MBS, the user will enrol themselves. A short video is captured. The audio/speech signal is used to extract the speech features and the image is extracted from the same video for facial feature extraction. The resultant feature vectors say will be fused at the feature level (Gofman et al., 2016) using the concatenation method. These fused feature vector templates of individual speaker models are trained via CNN classifier, a deep learning model and saved in the database. The testing phase is similar to the training phase, except that the unknown claimed user's fused biometric template is matched with the trained reference N models, saved in the database. A user is identified or rejected based on the 1:N comparison of the fused template i.e., the claimed user's fused template is matched against N registered templates stored in the database. The decision is taken based on the degree of similarity between the query and the reference data.

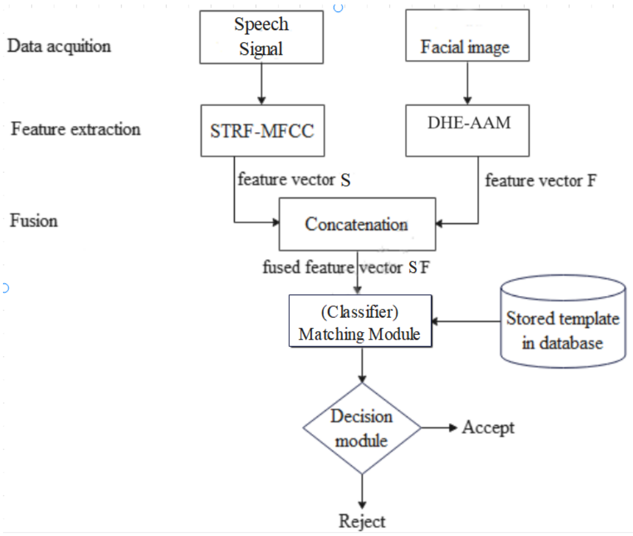

Figure 2. Block diagram of the proposed identification system

6.1 Database

VidTIMIT (Sanderson, 2002) is a publicly available multi-modal database found suitable for simulating the proposed approach for user identification. The collection consists of 43 video recordings along with their corresponding audio (10 for each), reciting short sentences selected from the TIMIT corpus. Each individual's video is kept as a series of JPEG images with a resolution of 512×384 pixels. The audio samples are saved as a .wav file that is mono, 16 bits and 32 kHz.

6.2 Proposed features and its feature level fusion

At first, all audio/speech samples are pre-processed through standard Python library functions such as noisereduce. Speech signals are sampled at 16kHz and divided into frames of 30ms with 20ms overlap and hamming window is used. Then as discussed in section 4, thirteen SDL-MFCC features are extracted for each speech (text-independent) frame and stored in a .csv file. For face cues, the image stream (of the respective speaker) as input consists of grayscale facial images with dimensions of 128 × 128 pixels. And as discussed in section 5, the image enhancement technique followed by feature extraction DHE-AAM (Abdullah-Al-Wadud et al., 2007; (Ning et al., 2023)) is applied. The speech-face cues being fused at the feature level is explored to be the most useful approach. Also, as features contain raw information, thus the resultant single feature vector will be prevented from emulation and will make the system free from any adversary attacks. Therefore, the aforesaid resultant audio-image feature vector says, $S = [S_1....S_n]$ and $F = [F_1.... F_n]$ are fused at the feature level using the concatenation method as:

$$SF_c = [S_1....S_n \;\; F_1 .... F_n] \tag{13}$$

The feature level fusion enables a rapid assessment of the efficacy of multiple cues and the simplest of implementation among all methods. For DR, the state-of-the-art PCA technique is used. This will evict the raw biometric data noise, thereby potentially ameliorating the recognition rate and accuracy (even in noisy/low illumination conditions).

6.3 Classifier

In this work, the state-of-the-art 2-Dimensional Deep Convolutional Neural Network (2D CNN) with three layers is used as a classifier. It is implemented in Python using Keras with TensorFlow (Jose, 2019) as backend, to make the most of predictive accuracy while automatically avoiding over-fit of the data even in adverse conditions. The deep CNN classifier contains of one input layer with one output layer, three convolutional layers and three pooling layers. The different parameters of convolutional and pooling layers are discussed in Table 1.

Table 1. Description of layers used in DCNN classifier

Convolutional layer 1	
Filter count	16
Kernel size	1 x 1
Activation	ReLU
Pooling layer 1	
Pool size	1 x 1
Dropout rate	0.2
Convolutional layer 2	
Filter count	32
Kernel size	1 x 1
Activation	ReLU
Pooling layer 2	
Pool size	1 x 1
Dropout rate	0.2
Convolutional layer 3	
Filter count	64
Kernel size	1 x 1
Activation	ReLU
Pooling layer 3	
Pool size	1 x 1
Dropout rate	0.2
Loss	Sparse categorical cross entropy
Optimizer	Adam
Batch size	100

The multi-layers in CNN (Qian et al., 2021; Nithya & Sripriya, 2022) based classifier permit the models to become more efficient at learning complex features and perform a more intensive computational task. ReLu (Jha et al., 2023a; Jha et al., 2023b) is used as the activation function, with Adam optimizer (Jha et al., 2023a; Jha et al., 2023b) to lower the cost function. The performance of the CNN classifier is evaluated using accuracy (Jha et al., 2023a; Jha et al., 2023b) and Equal Error Rate (EER) (Abozaid et al., 2019; Freire-Obregon et al., 2021) in percentage. The user identification accuracy (%) of the aforesaid classifier is evaluated as:

$$Accuracy = \frac{(total\ test\ samples - error)}{total\ test\ samples} x\ 100\% \qquad (14)$$

The EER value implies that the proportion of false acceptances (FA) and false rejections (FR) are proportionally equal. If the EER is minimal, then the user identification system is said to be reliable and efficient. Also, the VidTIMIT database is divided into 80% training data and 20% as testing data for user identification tasks.

7. RESULTS AND DISCUSSION

The research is conducted using an HP Laptop equipped with a 12th generation Intel CoreTM i5-1240P processor and 16 GB of RAM, operating on the Windows 11 platform. The experiment was conducted utilising the Python programming language, specifically version 3.8. The TensorFlow environment is used to achieve the desired result in deep learning. At first, we have implemented the individual unimodal identification system for speech and face respectively. Then we have implemented the feature level fusion of the aforesaid cues to see their combined effectiveness for user identification task. The aim of this study is to examine the efficacy of the proposed feature level fusion of speech and facial cues as MBS with a CNN classifier across various training biometric samples and number of users. As a result, Table 2 depicts the comparative performance of the proposed MBS in terms of accuracy (%) for user identification task with respect to UBSs for various percentage of training sample (100% and 80%) with 20% testing sample.

Table 2. Performance of the proposed MBS using CNN classifier with various percentage of training data

Biometric System	**Feature Extraction Technique**	**Training dataset**	
		100%	80%
UBS for Face	AAM	96.12 %	**95.72 %**
UBS for Speech	S_{DL}-MFCC	94.98 %	**93.48 %**
Proposed MBS	Feature level Fusion of AAM and S_{DL}-MFCC features	98.52 %	**97.31 %**

According to the aforementioned data, we can conclude that better results are obtained when 100% of the test samples are trained, or when the outcomes are sample dependant. On the other hand, the efficacy of any system can only be regarded credible when it is tested using a small number of samples that have not been trained, also known as sample independence. Additionally, it aims to evaluate the effectiveness of the proposed user identification method in accurately classifying samples that have not been previously observed or trained. In order to provide further clarification on the findings presented in Table 2, a comparative analysis (%) of the proposed MBS with UBSs for different training data size is presented in figure 3.

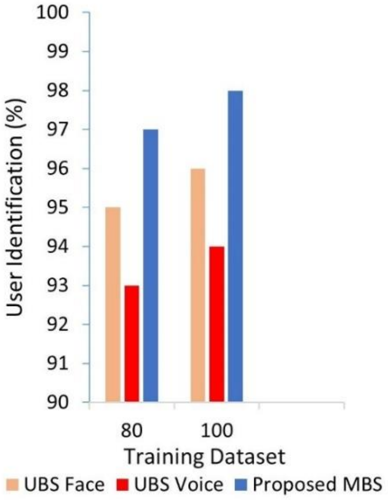

Figure 3. Comparative performance of the proposed MBS and UBSs with various percentage of training dataset

And Table 3 represents the performance of the proposed MBS and UBS in terms of accuracy (%) with different number of users, with database divided into 8:2 ratio for training and testing respectively. It also represents the equal error rate (EER in %) evaluated for UBSs and the proposed MBS. In order to provide further clarification on the findings presented in Table 3, a comparison (%) of the proposed MBS with UBSs for various number of users is shown in figure 4.

Table 3. Performance of the proposed (in %) MBS using CNN classifier with various number of users

Biometric System	Feature Extraction Technique	No. of User			EER
		10	25	43	
UBS for Face	AAM	89.4	92.3	**95.72**	**9.53**
UBS for Speech	S_{DL}-MFCC	86.0	91.4	**93.48**	**11.10**
Proposed MBS	AAM and S_{DL}-MFCC	92.1	95.1	**97.31**	**3.62**

From Table 3, we can infer that the performance of the CNN classifier depends on the size of the dataset. With large amount data, CNN achieve significant result. To further elucidate our work, we have done a comparative study of few existing feature level fusion based methodologies in Table 4. The performance of the methodologies is based on the accuracy and EER percentage.

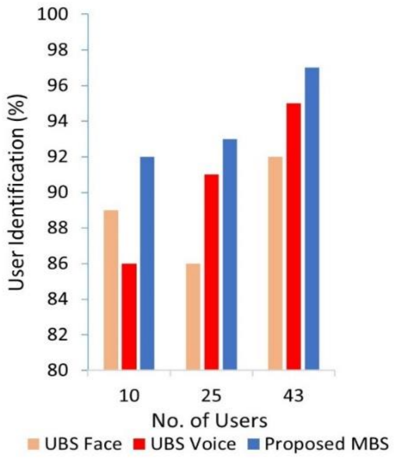

Figure 4. Comparative performance of the proposed MBS and UBSs with various number of users

Table 4. Comparison of few existing feature level fusion methodologies.

Ref.	Methodology	Performance
Gofman et al., 2016	Facial feature – Discrete cosine transform (DCT). Speech feature – MFCC Data fusion – Concatenation Classifier – Hidden Markov Model.	EER of the proposed fusion method with 11.87% result into higher accuracy compared to the individual unimodal system (face 16.05% and speech 21.58%). A larger database would have been used for better performance evaluation.
Abozaid et al., 2019	Facial feature – Eigenface and principal component analysis (PCA) Speech feature – cepstral and statistical coefficients Classifier – Gaussian Mixture Model (GMM).	The proposed method achieved an EER equal to 2.81%.
Dalila et al., 2020	Facial feature – DCT and PCA. Speech feature – MFCC and vector quantization (VQ). Data fusion – Concatenation. Classifier – Artificial neural network (ANN).	The proposed method attained lowest error using ANN having 1.67% against K-nearest-neural classifier as 8.1%. Also, the recognition rate under clean environment for ANN gave better results than K-NN. An enhanced method should be proposed for noisy environment.
Freire-Obregon et al., 2021	Facial feature – ResNet50 and PCA Speech feature – Triplet and PCA Data fusion – Concatenation Classifier – Deep learning architecture.	An EER of the proposed model, 13.35% was achieved, which outperforms the unimodal biometric system (face 13.38% and speech 32.07%). Compared to other above mentioned methods, EER is considerably high.
Proposed MBS	Facial feature – AAM Speech feature – S_{DL}-MFCC Data fusion – Concatenation Classifier – DCNN.	The proposed feature level based MBS achieved an identification accuracy of 97.31%. Also, it outperforms UBS for face and speech by 1.59 % and 3.83% respectively. An EER of 3.62% is achieved by the proposed methodology, which surpass all the above methodologies.

Hence, as per the above findings (Table 2, 3 and 4), we can infer that the proposed feature level fusion of speech and face cues using deep CNN classifier, outperforms the UBSs and existing MBS based methodologies.

8. CONCLUSION

This paper explores different techniques for data fusion at the feature level, including dimensionality reduction in MBS, feature extraction and classifier, to implement feature level fusion using speech and face cues as an efficient MBS. Based on these, we have formulated and proposed a methodology for an effective MBS for user identification tasks. For robust feature extraction of speech cues, a state-of-the-art SDL-MFCC acoustical feature set is proposed, and for facial features, DHE with AAM with a 2D CNN classifier is proposed. We have investigated the effectiveness of the proposed MBS and UBS using the DCNN classifier, over different training data sizes, number of users, accuracy and EER. The benefits and drawbacks of various approaches are also discussed and compared. Based on the findings, we infer that the proposed MBS approach surpassed both UBSs and existing methodologies. In a nutshell, the proposed approach can be considered an efficient real-time solution for user identification.

References:

Abdullah-Al-Wadud, M., Kabir, M. H., Dewan, M. A. A., & Chae, O. (2007). A dynamic histogram equalization for image contrast enhancement. *IEEE transactions on consumer electronics*, *53*(2), 593-600. doi: 10.1109/TCE.2007.381734

Abozaid, A., Haggag, A., Kasban, H., & Eltokhy, M. (2019). Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion. *Multimedia tools and applications*, *78*, 16345-16361. doi:10.1007/s11042-018-7012-3

Agrawal, S. S., Jain, A., & Sinha, S. (2016). Analysis and modeling of acoustic information for automatic dialect classification. *International Journal of Speech Technology*, *19*, 593-609. doi:10.1007/s10772-016-9351-7

Dalila, C., Saddek, B., & Amine, N. A. (2020). Feature level fusion of face and voice biometrics systems using artificial neural network for personal recognition. *Informatica*, *44*(1). doi:10.31449/inf.v44i1.2596

Freire-Obregon, D., Rosales-Santana, K., Marín-Reyes, P. A., Penate-Sanchez, A., Lorenzo-Navarro, J., & Castrillón-Santana, M. (2021). Improving user verification in human-robot interaction from audio or image inputs through sample quality assessment. *Pattern Recognition Letters*, *149*, 179-184. doi:10.1016/j.patrec.2021.06.014

Gofman, M. I., Mitra, S., & Smith, N. (2016, November). Hidden markov models for feature-level fusion of biometrics on mobile devices. In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)* (pp. 1-2). IEEE. doi: 10.1109/AICCSA.2016.7945755

Jha, K., Jain, A., & Srivastava, S. (2023, March). An Efficient Speaker Identification Approach for Biometric Access Control System. In *2023 5th International Conference on Recent Advances in Information Technology (RAIT)* (pp. 1-5). IEEE. doi: 10.1109/RAIT57693.2023.10127101

Jha, K., Srivastava, S., & Jain, A. (2023, March). Integrating Global and Local Features for Efficient Face Identification Using Deep CNN Classifier. In *2023 International Conference on Device Intelligence, Computing and Communication Technologies,(DICCT)* (pp. 532-536). IEEE. doi: 10.1109/DICCT56244.2023.10110170

Jose, R. (2019). A convolutional neural network (cnn) approach to detect face using tensorflow and keras. *International Journal of Emerging Technologies and Innovative Research*, *ISSN*, 2349-5162.

Ketab, F., Russel, N. S., Selvaraj, A., & Buhari, S. M. (2023). Parallel deep learning architecture with customized and learnable filters for low-resolution face recognition. *The Visual Computer*, 1-12. doi: 10.1007/s00371-022-02757-y

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, *52*(1), 12-40. doi:10.1016/j.specom.2009.08.009

Mahmood, B. A., & Kurnaz, S. (2023). An investigational FW-MPM-LSTM approach for face recognition using defective data. *Image and Vision Computing*, *132*, 104644. doi: 10.1016/j.imavis.2023.104644

Mehraj, H., & Mir, A. H. (2021). A multi-biometric system based on multi-level hybrid feature fusion. *Herald of the Russian Academy of Sciences*, *91*(2), 176-196. doi:10.1134/S1019331621020039

Meyer, B. H., Pozo, A. T. R., & Zola, W. M. N. (2022). Global and local structure preserving GPU t-SNE methods for large-scale applications. *Expert Systems with Applications*, *201*, 116918. doi:10.1016/j.eswa.2022.116918

Neto, E. D. A. L., & Rodrigues, P. C. (2023). Kernel robust singular value decomposition. *Expert Systems with Applications*, *211*, 118555. doi:10.1016/j.eswa.2022.118555.

Ning, X., Nan, F., Xu, S., Yu, L., & Zhang, L. (2023). Multi-view frontal face image generation: a survey. Concurrency and Computation: Practice and Experience, 35(18), e6147. doi:10.1002/cpe.6147

Nithya, B., & Sripriya, P. (2022). Feature-Level Fusion of Multimodal Biometric for Individual Identification by Training a Deep Neural Network. In *Inventive Communication and Computational Technologies: Proceedings of ICICCT 2021* (pp. 145-159). Springer Singapore.

Nguyen, D., Nguyen, K., Sridharan, S., Dean, D., & Fookes, C. (2018). Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Computer vision and image understanding*, *174,* 33-42. doi:10.1016/j.cviu.2018.06.005

Oloyede, M. O., & Hancke, G. P. (2016). Unimodal and multimodal biometric sensing systems: a review. *IEEE access*, *4*, 7532-7555. doi:10.1109/ACCESS.2016.2614720

Oskolkov, N. (2022). Dimensionality Reduction: Overview, Technical Details, and Some Applications. *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, 151-167.

Qian, Y., Chen, Z., & Wang, S. (2021). Audio-visual deep neural network for robust person verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 1079-1092. doi: 10.1109/TASLP.2021.3057230

Ryu, R., Yeom, S., Kim, S. H., & Herbert, D. (2021). Continuous multimodal biometric authentication schemes: a systematic review. *IEEE Access*, *9*, 34541-34557. doi:10.1109/ACCESS.2021.3061589

Sanderson, C. (2002). *The vidtimit database* (No. REP_WORK). IDIAP.

Srivastava, S., Chandra, M., & Sahoo, G. (2019). Speaker identification and its application in automobile industry for automatic seat adjustment. *Microsystem Technologies*, *25*, 2339-2347. doi: 10.1007/s00542-018-4111-z

Taskiran, M., Kahraman, N., & Erdem, C. E. (2020). Face recognition: Past, present and future (a review). *Digital Signal Processing*, *106*, 102809. doi: 10.1016/j.dsp.2020.102809

Wang, J. C., Wang, C. Y., Chin, Y. H., Liu, Y. T., Chen, E. T., & Chang, P. C. (2017). Spectral-temporal receptive fields and MFCC balanced feature extraction for robust speaker recognition. *Multimedia Tools and Applications*, *76*, 4055-4068. doi: 10.1007/s11042-016-3335-0

Yaermaimaiti, Y., & Kari, T. (2023). Research on a feature fusion-based image recognition algorithm for facial expression. *International Journal of Information and Communication Technology*, *22*(2), 133-146. doi: 10.1504/IJICT.2023.128712

Khushboo Jha
Birla Institute of Technology Mesra,
Ranchi-835215,
India
kjha.phd@gmail.com
ORCID 0000-0003-1062-8128

Aruna Jain
Birla Institute of Technology Mesra,
Ranchi-835215,
India
arunajain@bitmesra.ac.in

Sumit Srivastava
Birla Institute of Technology Mesra,
Ranchi-835215,
India
sumit.srivs88@gmail.com
ORCID 0009-0003-6880-2958