www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 15(3), Sep-Dec 2024, 4-14. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2024.v15i2.289

Bimodal biometric recognition system using Convolutional Neural Networks and fusion of deep audiovisual feature vectors

Juan Carlos Atenco-Vázquez[1], Juan Carlos Moreno-Rodríguez[1], Juan Manuel Ramírez-Cortés[1], René Arechiga-Martinez[2], Pilar Gomez-Gil[3], Rigoberto Fonseca-Delgado[4]

[1] Department of Electronics, National Institute of Astrophysics, Optics and Electronics, Luis Enrique Erro 1, Sta. Maria Tonantzintla 72840, Mexico.
[2] New Mexico Tech, 801 Leroy Place, Socorro, NM 87801, USA.
[3] Department of Computer Science, National Institute of Astrophysics, Optics and Electronics, Luis Enrique Erro 1, Sta. Maria Tonantzintla 72840, Mexico.
[4] Electrical Engineering Department, Metropolitan Autonomous University, San Rafael Atlixco 186, 09340 Iztapalapa, CDMX, Mexico.

E-mails: atencovaz@inaoep.mx (corresponding author), xalatl@inaoep.mx, jmram@inaoep.mx, ene.arechiga@nmt.edu, pgomez@inaoep.mx, rfonseca@izt.uam.mx

Abstract. In recent years, interest has grown in the use biometric systems for identity authentication tasks in digital services, forensic and security applications. A unimodal system (employing a single biometric trait) with high performance is still vulnerable to falsification attacks such as spoofing. For this reason, research on multimodal biometrics (employing various biometric traits) has increased to reinforce security, increase recognition performance, and make false identity authentication more difficult. In this paper, we propose a bimodal system that combines speech and face modalities by concatenating their feature vectors, these vectors are extracted from two convolutional neural networks (CNN) and used for identity verification. The performance of unimodal CNNs was evaluated individually and compared to the bimodal system of concatenated vectors. A data augmentation scheme is used for both modalities to evaluate different operation conditions. Results were measured in terms of Equal Error Rate (EER).
Keywords: Multimodal biometrics, Speaker recognition, Face recognition, Convolutional Neural Networks, Audiovisual biometrics.

Article Info
Received Nov 8, 2021
Accepted Mar 27, 2023

1 Introduction

Traditional methods such as passwords, personal identification numbers (PINs), identification cards, etc. are used to authenticate the identity of an individual to access commercial or government digital services, or personal electronic devices. These methods have several security risks: the possibility that a third party can obtain or guess passwords or PINs, lost or stolen identification tokens, etc. Another frequent risk is that the owner forgets the password or PIN and must request it again. As a result, research on biometric applications for different applications has increased in recent years. Since biometric authentication systems use traits of an individual's body that remain with one throughout life, they are unique and distinguish one individual from any other (Dahea & Fadewar, 2018).

Biometrics is the study of measurement and analysis of both physical and behavioral biological traits with the purpose of authenticating the identity of an individual. These traits, also known as biometric traits, are considered unique for every individual in a population (Sabhanayagam, Venkatesan, & Senthamaraikannan, 2018). A biometric system aims to perform identity authentication considering the information extracted from biometric traits such as: voice, face, iris, fingerprint, gait, DNA, among others. The biometric system captures the biometric information, processes this information with mathematical methods, matches the processed information with templates previously stored in the system, issues a numerical score resulting

from the match, and decides, considering said score (Modak & Jha, 2019). Figure 1 shows the basic building blocks of a biometric system.

Although many unimodal biometric systems have high recognition performance, they also have several inherent problems such as intraclass variation, low interclass separability, vulnerability to different types of attacks, low-quality-captured biometric data, not representative models, etc.

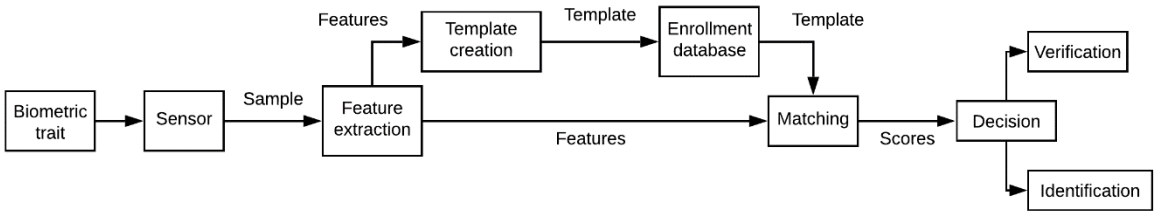

Fig. 1. Biometric system block diagram.

Biometric verification and identification

Identity authentication by a biometric system generally involves two tasks: verification and identification (Dahea & Fadewar, 2018; Sabhanayagam, Venkatesan, & Senthamaraikannan, 2018; Modak & Jha, 2019). In identification, the features of an identity claim are matched with all the templates stored in the biometric system in order to obtain the label of the stored template that most closely resembles the identity claim, it is a one-to-many match. In verification, the features of a client are matched with a specific template stored in the biometric system to verify that the identity of the client and that of the template are the same, it is a one-to-one comparison.

Multimodal biometric systems

Multimodal biometric systems have the same basic structure as unimodal systems as presented in Figure 1. The main difference is that multimodal systems use a scheme of fusion information at some point in the basic structure of a biometric system. Multimodal systems can combine several biometric features or several sources of information for the same trait (Fierrez, Morales, Vera-Rodriguez, & Camacho, 2018; Modak & Jha, 2019). Generally, the fusion of information from a multimodal system is carried out before the classification of the biometric sample or after the classification. In (Fierrez, Morales, Vera-Rodriguez, & Camacho, 2018; Dinca & Hancke, 2017), it is detailed that pre-classification fusion can be at sensor level or at feature level, while the post-classification fusion can be found at decision, rank and score levels. The fusion levels are briefly detailed below:

- *Sensor level*: the biometric data of various sensors can be combined, either data from various sensors for a single biometric trait or one sensor per trait.
- *Feature level*: it combines the feature vectors obtained from processing the sampled data from sensors. In a similar fashion, there can be several feature vectors per trait or one vector for each trait.
- *Decision level*: it combines the identity verification decisions made by individual classifiers into a single decision. It is possible to either train various classifiers with data of a single trait, or train various classifiers (one classifier per trait) to combine their individual decisions.
- *Rank level*: it applies to identity identification and consists in that a classifier can deliver a series of ranked identity labels.
- *Score level*: the normalized scores from various classifiers are combined into a single score.

Score level fusion is generally considered the most used in multimodal systems, since the decision scores are the result of processed data by the classifier, it is widely considered that scores contain more discriminative information. Despite this, the other fusion levels are still being investigated to improve multimodal systems performance.

2 Related work

Deep Learning (DL) is a Machine Learning (ML) paradigm which consists of a feature extraction scheme using multiple layers; the more layers used, the more complex the extracted features. Currently, research on ML and DL methods has increased thanks to the increase in computing power of both Central Processing Units (CPUs) and Graphics Processing Units (GPUs) that significantly decrease computing time; it is worth mentioning the increase of software libraries specialized in training and evaluation of ML and DL models (López, 2019).

ML models, particularly those applied to biometrics, require feature extraction from biometric data through mathematical processing techniques to obtain useful information that allows models to distinguish individuals of a population; this information is fed to a classifier model which makes the final decision according to the authentication task being performed. However, it is possible that certain processing techniques are not compatible with certain biometric traits, with certain databases, or with some classifiers (Alay & Al-Baity, 2020).

DL models have generated great interest, as they have shown high recognition performance and don't have some of the limitations of ML models. Among the advantages that DL has over ML are the extraction of more complex characteristics as more layers are used, the use of non-linearities to map characteristics to an output domain appropriate to the task to be performed, and the offering of a scheme in which a model jointly learns the extracted features and performs classification and/or regression (Alay & Al-Baity, 2020; Minaee, Abdolrashidi, Su, Bennamoun, & Zhang, 2019; Irum & Salman, 2019).

In this work we considered the biometric modalities of voice and face; in the next sections we present a brief review of both modalities within the Deep Learning scheme.

Related work on voice biometrics

In this work we will refer to voice biometrics as speaker recognition since it is the most-used terminology in literature.

As mentioned above, ML methods are traditionally used for biometric authentication. In the case of speaker recognition, different characteristics of the voice signal are used. Usually, these characteristics are extracted from segments of a few milliseconds since these segments are considered stationary.

In (Mandalapu et al., 2021), the feature extraction methods are classified into four groups: cepstral coefficients such as the Mel Frequency Cepstral Coefficients (MFCC), wavelet transformations, voice frequency features obtained by means of short-time Fourier transform (STFT) and Linear Predictive Coefficients (LPC) which model the vocal tract as a filter. In (Kinnunen & Li, 2010) , other approaches are presented, such as modeling the anatomy of the vocal tract using statistical methods, modeling the prosodic content by extracting the fundamental frequency F0 and combining it with other features, and modeling high level features by differentiating individuals' speech patterns.

Classical approaches to model voice features are based on methods such as Vector Quantization (VQ), Gaussian Mixed Mixtures (GMM) with the Universal Background Models scheme (GMM-UBM), Hidden Markov Models with GMM (HMM-GMM), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) (Kinnunen & Li, 2010).

Great progress in the speaker verification task occurred with the development of i-vector, low dimensional vector representations that result from projecting the parameters of a GMM to a variability space. The extracted i-vectors are fed to a Probabilistic Linear Discriminant Analysis (PLDA) model which then decides based on a score (Sztahó, Szaszák, & Beke, 2019).

With the rise of DL, Deep Neural Networks (DNN) and CNNs began to be used for speaker recognition (Sztahó, Szaszák, & Beke, 2019). The first works used DNNs to calculate statistics to calculate i-vectors with better recognition than those generated by GMMs. Other approaches train DNNs for feature extraction from the statistics calculated in the intermediate hidden layers of the networks known as Bottleneck features; these features can be used for verification with PLDA or as a basis for calculating i-vectors. In other approaches, spectro-temporal features such as MFCC or spectrograms are used to train and evaluate CNNs as in (Li et al., 2020). The scheme that has gained great notoriety in recent years are the feature embeddings, which consist of extracting the parameters of the last hidden layer of a DL network to obtain compact, low-dimensional representations; examples of these vectors are d-vectors, j-vectors, and more recently, x-vectors. In recent years, interest has grown in processing raw speech signals with DL networks; the network adapts its parameters to the characteristics of the signal. In (Muckenhirn,

Doss, & Marcell, 2018), raw speech signal is processed frame-by-frame with a 1D CNN for speaker verification. In (Ravanelli & Bengio, 2018), a 1D convolutional layer is proposed to extract frequency features by means of a filter bank modeled as parameterized sinc functions; these parameters are adapted through network training; this network was named Sincnet. Irum & Salman (2019) and Sztahó, Szaszák, & Beke (2019) offer a more extensive and exhaustive review of DL applied to speaker recognition.

Related work on face biometrics

Authentication of an individual's identity using his face features is a research topic that continues to be of great interest; different model proposals deliver a very high recognition performance for various databases and under different operation conditions.

Face biometrics, better known as face recognition, is a task that aims to have a machine recognize a human being to engage in interaction (Masi, Wu, Hassner, & Natarajan, 2018). Hence, various challenges related to the conditions in which face samples are captured as images or videos arise; typically, a recognition system for this modality requires considerable computing resources.

To perform face recognition, there are various feature extraction methods that have been widely used for many years. In Mandalapu et al. (2021), the methods for extracting features from a face image are divided into four groups: signal processing methods such as mathematical transformations to other domains such as the Fourier transform (FT), animation-based features that consider statistical animation parameters such as face shape and appearance, methods with convolutional kernels such as Viola Jones or Haar filters, methods based on texture representations that consider the information contained in the relationships of pixels with their neighbors such as Local Binary Pattern (LBP) or Gradient Histogram Oriented (HOG).

In Kortli, Jridi, Al Falou, & Atri (2020) and Andy, Bridget, Dono, & Benjamin (n.d.), face recognition systems are classified into 3 categories: local systems that either consider specific features focusing on regions of interest, or that divide the image into segments such as HOG; holistic systems that process the face image without dividing it and later convert it into feature vectors such as the Eigenfaces system; and third, hybrid systems that combine systems from the two previous categories or other approaches such as CNNs.

The above-mentioned approaches offer different recognition performances which are highly dependent on the feature extraction method. The use of deep networks became popular due to its high performance for face recognition in large databases. CNNs such as Deep-Face and FaceNet generated great interest to continue improving results using DL networks. The main idea of this scheme is to use specialized blocks to build a network, such as convolutional layers and pooling layers, and to modify the training hyperparameters to accelerate the convergence of the network. Other networks that meant a great improvement in creating different blocks for feature extraction are VGG16, ResNet, and GoogleNet (Ríos-Sánchez, Costa-da Silva, Martín-Yuste, & Sánchez-Ávila, 2020). In recent years, one goal has been to bring the high recognition of DL networks to mobile devices with limited computing resources; for example, OpenFace is a project that uses a modified FaceNet network to be retrained in different conditions.

Other great advances that have occurred in the face recognition scene with DL is the creation of new cost functions for faster network convergence; these functions increase interclass separation and reduce intraclass distance. Examples of these functions are triplet loss, contrastive loss and the family of functions derived from the combination of softmax and cross entropy such as Additive Margin softmax (AM-Softmax) (Masi, Wu, Hassner, & Natarajan, 2018; Chen, Liu, & Li, 2020).

3 Methodology

As mentioned in section 3, in order to build a multimodal biometric system, the combination or fusion of information can be performed at different levels.

In Bharathi & Sudhakar (2019), a bimodal system with fingerprint and palm veins was built; 2D Gabor filters are taken as features and matched with Euclidean Distance (ED); each modality generates a score; both are fused with a Fuzzy Inference System (FIS). In Zhang, Cheng, Jia, Dai, & Xu (2020), voice and face modalities are combined; LBP histograms are extracted from face images with ED as the matching function; in voice modality the MFCCs are extracted to train a GMM using the Maximum Posterior Likelihood (MAP) criteria for score calculation; scores are fused using a weighting function. Score level

fusion can be performed using a DL model; in Cherrat, Alaoui, & Bouzahir (2020), three CNNs are trained for fingerprint, finger veins and face modalities; each CNN outputs a single score, and two fusion experiments are performed: the first one uses a score weighted sum to obtain a final score; the second experiment uses a multiplicative function.

Feature level fusion has also been extensively investigated given the large amount of discriminant information contained in feature vectors. In Xin et al. (2018), feature vectors of fingerprint, finger veins and face modalities are obtained; these vectors are concatenated; Fisher vectors are calculated from the concatenated vector; the Fisher vectors are subsequently used to train and evaluate different classifiers. In Olazabal et al. (2019), voice and face modalities are combined; for voice, the MFCCs are extracted, and for face, HOG and LBP features are extracted; these three vectors are combined using Discriminant Correlation Analysis (DCA), K Nearest Neighbors (KNN) is used to classify the combined feature vector. Regarding DL, there are also many proposals that use deep networks for feature level fusion. In Leghari, Memon, Dhomeja, Jalbani, & Chandio (2021), fingerprint and digital signature modalities are combined by training a two input CNN; each modality is processed in parallel; finally, the features extracted with the convolutional layers for each modality are fed to fully connected layers to create feature vectors, and then concatenated within the network. A similar approach is found in Luo, Li, & Zhu (2021); iris and periocular modalities are fused with a CNN with two inputs; the generated features vectors are combined with spatial and channel attention mechanisms. In Alay & Al-Baity (2020), three CNNs are trained for iris, face and finger veins; these networks are trained individually, and feature vectors are extracted from each network; the extracted vector are concatenated and fed to a softmax classifier for identification purposes.

In our work, we build a bimodal voice and face biometric verification system using CNNs; we follow a scheme similar to (Alay & Al-Baity, 2020). We train a CNN for each modality, and the vectors of the last hidden layer of each network were extracted; this pair of vectors is concatenated to build the template vector of each client of the system. We employed a multimodal database that contains data of voice, face and electroencephalogram (EEG) modalities. Since the database is small, we use a data augmentation scheme to generate more information to train and evaluate the system; the augmented data allowed us to perform evaluations under different conditions for both modalities; this will be explained in the next sections.

Databases

Biometric data from BIOMEX-DB (Moreno-Rodriguez et al., 2021) was used for our experiments. This database has information from 51 volunteers and consists of audio recordings of the pronouncing of isolated digits, and EEG signals recorded during pronunciations; only 39 volunteers had video recordings made of their faces during the digit-pronunciations.

For this study, we use both video and audio recordings as face and speech data respectively. The database is divided into 2 sets: the first one is composed of audio and video recordings of 10 strings of 10 digits each; the second set consists of recordings of 10 strings of 4 digits each.

Since there is more speech information than face information, to compensate for the missing face information we included face images taken from the Yaleface database (Belhumeur, Hespanha, & Kriegman, 1966). The 12 individuals from BIOMEX-DB who only have speech data were matched with 12 subjects from the Yaleface database. How the data will be used in the verification task will be explained later.

Speech data preprocessing

The BIOMEX-DB database contains noiseless speech recordings; the speech data was augmented similarly to Snyder, Garcia-Romero, Sell, Povey, & Khudanpur (2018); however, only background noise from the MUSAN database(Snyder, Chen, & Povey, 2015) was added to the original speech samples. Unlike Snyder, Garcia-Romero, Sell, Povey, & Khudanpur (2018), the noise samples were added randomly. For each original recording there are four copies with added noise; the noise of each copy has a Signal to Noise Ratio (SNR) value in dB; the SNR values considered were: 0, 5, 10 and 15 dB. Finally, all the audio data was normalized; the values were in the range (-1,1).

Face data preprocessing

The face images were extracted from the videos of the BIOMEX-DB and Yalefaces databases. Since the recording protocol of BIOMEX-DB focused on EEG, in many video frames the volunteer has his/her eyes closed in a state of relaxation; these frames are not useful due to the low variability of information they contain; only frames corresponding to moments when digits were being pronounced were extracted, so the images contain information on various facial expressions.

After the image extraction, we used the Scikit-image library (Boulogne, Warner, & Yager, 2014) face detector based on LBP cascade classifiers; the detected faces were cropped to generate new images of 100x100 pixels. Once the face images were extracted, geometric transformations were applied for data augmentation; these transformations were: flip them horizontally and random rotation with angles between -45° to 45°. Additionally, the brightness of the images was modified by adding random offset values to the pixel value; this results in lighter or darker images. A fourth set of augmented images consisted in combining the three previous modifications. All images were normalized to 8 bits grayscale. Finally, the LBP operator was applied to all the generated images. The LBP operator describes the local texture characteristics, usually applied in tasks of texture classification, face analysis and even face recognition through histograms. Among the advantages of this operator is that it is rotation and grayscale invariant (Zhang, Qu, Yuan, & Li, 2017). The LBP operator encodes the pixels of the original image and generates a new image whose pixels contain information about the relationships they have with their surrounding neighbors; for this work the LBP operator was used in its uniform variant using the Scikit-image library.

Various works have shown high performance when training and evaluating a CNN with images generated by LBP operator such as (Zhang, Qu, Yuan, & Li, 2017; Ke, Cai, Wang, & Chen, 2018). In Wang, Wang, & Li (2017), good results are shown when training a CNN with LBP images in small databases.

CNNs architectures

The CNN for speaker recognition is based on the convolutional Sincnet layer described in Ravanelli & Bengio (2018); this layer processes raw audio frames of several milliseconds with a filterbank of bandpass filters, where each filter is modeled as a sinc function. The network implemented with Sincnet has shown rapid convergence; it needs to learn fewer parameters; the resulting feature maps have an interpretation in the frequency domain and the parameters of the bandpass filters are adapted to the frequency features of the system clients' voice. The Sincnet network architecture used in this work is shown in Table 1.

Table 1. Sincnet network architecture for speaker recognition.

Layers	Filters/Neurons	Size	Activation fcn
Sincnet	120	251	LeakyReLU
Maxpooling1D	-	5	-
Batch Normalization	-	-	-
Convolution 1D	32	5	LeakyReLU
Maxpooling 1D	-	5	-
Batch Normalization	-	-	-
Convolution 1D	64	5	LeakyReLU
Maxpooling 1D	-	5	-
Batch Normalization	-	-	-
Fully connected	512	-	LeakyReLU
Batch Normalization	-	-	-
Fully connected	39	-	Softmax

For face modality, the CNN architecture is shown in Table 2.

Table 2. CNN architecture for face recognition.

Layers	Filters/Neurons	Size	Activation fcn
Convolution 2D	32	3x3	LeakyReLU
Maxpooling 2D	-	2x2	-
Batch Normalization	-	-	-
Convolution 2D	64	5x5	LeakyReLU
Maxpooling 2D	-	2x2	-
Batch Normalization	-	-	-
Fully connected	512	-	LeakyReLU
Batch Normalization	-	-	-
Fully connected	39	-	Softmax

Both networks have an output layer of 39 neurons since that is the number of individuals from the BIOMEX-DB database that have data from both modalities and were considered as the clients whose identity must be verified as true.

CNNs training

In the network training stage, the Keras library was used. This library is included in the Tensorflow platform. For the Sincnet network, the training was performed for 50 echos; the network was fed with speech frames of 200 ms; 75 batches of 128 frames each were used as training data; the validation data also consisted of 75 batches of 128 frames each; training and validation batches were randomly generated at training time. The learning rate was 0.001 and it was decreasing exponentially after epoch 30.

The speech training data was taken from the recordings of 10-digit pronunciations from the BIOMEX-DB database, by using the information from the time labels; the silences between pronounced digits were removed. Training data contained both noisy and noiseless data; noise of all SNR values were considered.

The face network was trained for 35 epochs; training set consisted of 50 batches of 80 images each; the validation set had 50 batches of 30 images; both sets of batches were randomly generated during training time. The learning rate was 0.001 and decreased exponentially after epoch 30. The training images were extracted from the videos in which the volunteers spoke 10 digits as mentioned above. The training set contained LBP images both original and augmented.

Unimodal biometric systems evaluation

To execute the verification task, a d-vector configuration was used; this means that the templates of genuine clients and test feature vectors of both genuine clients and imposters are extracted from the last hidden layer of the networks; in our case these vectors have a length of 512. To generate these vectors, the trained networks are fed with test samples. This is illustrated in Figure 2.

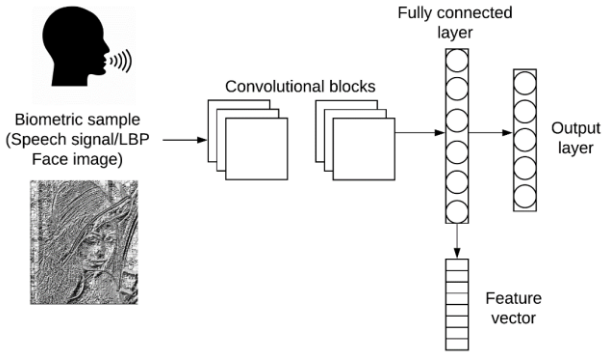

Fig. 2. Test feature vector extraction.

The voice evaluation set consisted of the audios with pronunciations of 4-digit strings; time labels were used to remove silences; likewise, the face evaluation set consisted of LBP images extracted from the videos of 4-digit-string pronunciations.

To create the voice template and store it in the biometric system, a noiseless audio sample and a noisy audio sample of each SNR value were randomly selected. Each audio sample was segmented into 200-ms frames with a 100-ms overlap; then these frames were fed to the network and the resulting vectors were averaged. This way, a single feature vector is generated for each 4-digit audio; this is illustrated in Figure 3.

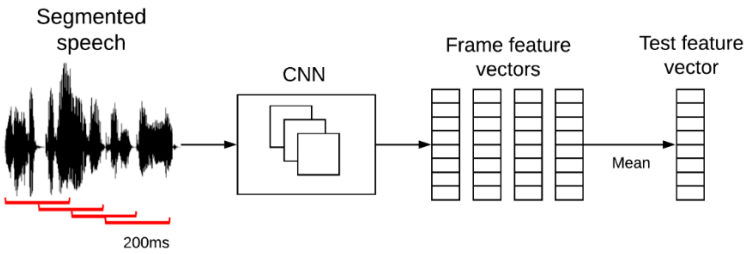

Fig. 3. Test feature vector extraction from a sentence.

The previous process was performed for the 5 selected audios; the 5 averaged vectors obtained are averaged again to obtain the template that contains information on the noise levels considered. The template generation is shown in Figure 4.

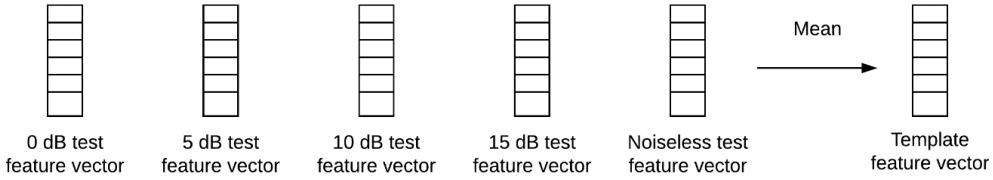

Fig. 4. Voice template feature vector generation.

The face template was generated in a similar way; an untransformed LBP image and an LBP image of each transformation described in section 7.3 were randomly selected. These images were fed to the trained face network and each vector extracted from the last hidden layer was averaged to obtain the template that contains information on the considered transformations. The test vectors are generated in the same way, one per image, without averaging with the vectors from other images. This process is shown in Figure 5.

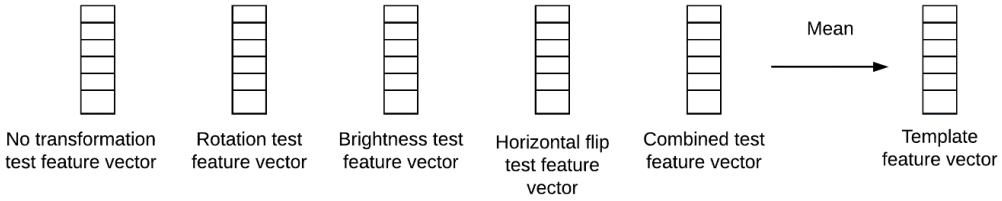

Fig. 5. Face template feature vector generation.

In the first stage of evaluation, verification tests were performed for each modality; the verification was made by matching the template vectors with the test vectors using cosine distance to generate a score. For our experiments, the templates of genuine clients were matched with their corresponding test vectors to evaluate the ability of the system to correctly verify their identity, the second match was made with the templates of genuine clients with the imposter test vectors to measure the system's ability to reject impostors.

As mentioned at the end of section 7.4, 39 legitimate clients were considered, corresponding to individuals who have speech and video information. The other 12 individuals from the database that only have audio data were taken as impostors, to compensate for the missing face information, images from 12 individuals of the Yalefaces database were paired with the impostors of BIOMEX-DB; Yalefaces data was preprocessed as described in section 7.3.

The verification results of each modality were measured in terms of Equal Error Rate (EER), which represents the point where True Positive Rate (TPR) and False Positive Rate (FPR) measurements are equal. The TPR is the number of times the identity of a legitimate client was successfully verified divided by the number of verification tests classified as positive. The FPR is the number of times an imposter was mistakenly verified as a legitimate client divided by the total number of tests classified as negative. More information on these measurements can be consulted in Fawcett (2006).

Tables 3 and 4 show the verification results of the individual modalities. Both were divided into different conditions considered in the data augmentation schemes.

Table 3. Verification results for speaker recognition.

SNR (dB)	EER (%)
0	15.48 +- 0.36
5	9.03 +- 0.24
10	4.53 +- 0.21
15	4.53 +- 0.21
Noiseless	4.36 +- 0.18

Table 4. Verification results for face recognition.

Transformation	EER (%)
No transformation	2.46 +- 2.3
Brightness	4.25 +- 2.89
Horizontal flip	3.09 +- 3
Rotation	8.95 +- 2.69
Combined	11.25 +- 1.99

In the second stage of the evaluation, the templates and test feature vectors of both modalities were concatenated; each client has a template and test vectors of length 1024. Templates and test vectors were compared with cosine distance. For this bimodal experiment, all the conditions considered in the first stage were combined; the results are shown in Table 5.

Table 5. Bimodal verification results.

Condition	EER (%)				
Transformation / SNR (dB)	No transformation	Brightness	Horizontal flip	Rotation	Combined
0	3.14 +- 0.67	3.87 +- 0.86	3.44 +- 0.63	4.14 +- 0.7	5.1 +- 0.45
5	2.34 +- 0.65	2.6 +- 0.55	2.64 +- 0.68	3.1 +- 0.38	3.62 + 0.63
10	2.03 +- 0.6	2.38 +- 0.59	2.16 +- 0.64	2.7 +- 0.36	3.08 +- 0.47
15	1.92 +- 0.59	2.25 +- 0.61	2.1 +- 0.7	2.3 +- 0.39	2.89 +- 0.45
Noiseless	1.82 +- 0.7	2.15 +- 0.55	2.09 +- 0.72	2.4 +- 0.38	2.88 +- 0.45

4 Discussion

In speaker recognition, the results show that the most difficult case is the SNR of 0 dB; the EER is approximately 15.48; as the SNR value increases the EER decreases. The best-case scenario corresponds to noiseless samples, and the EER is approximately 4.36. For face recognition, the rotation transformation has the highest EER value with approximately 8.95, while the other two transformations and the original images have a notably lower EER with 2.46 being the minimum value. When combining the 3 transformations, the EER has the highest value with 11.25.

It is observed that the results of the voice modality do not fluctuate significantly; the largest deviation is 0.36. In face recognition the results can be considered better than speaker recognition; however, the results have higher deviations. This allows us to conclude that the voice modality delivers the most stable results.

In the bimodal experiment the results were improved considerably compared to the individual modalities. The most noticeable improvement is shown in the most difficult conditions of evaluation i.e. 0 dB and combination of transformations. For 0 dB tests the EER is considerably lower than the individual test, decreasing from 15.48 to 3.14. The combination of transformation tests also showed a considerable decrease of EER, from 11.25 to a low of 2.88. The rest of the bimodal results showed improvement compared to unimodal results. Due to the bimodal combination, face recognition results became more stable.

5 Conclusions

Two biometric systems were created based on voice (speaker recognition) and face (face recognition) modalities by training two CNNs. Data from the BIOMEX-DB database was used for training and evaluation; we used some subjects from the Yalefaces database to create a set of impostors for evaluation purposes. A data augmentation scheme was implemented; we considered a specific set of conditions for each modality; augmented data was employed to train and test the CNNs.

Unimodal biometric verification was performed extracting feature vectors from the last hidden layer of the CNNs, then a template for each genuine client was created by averaging a specific set of feature vectors; in the final step, we used cosine distance as a matching function between templates and test vectors; EER was used to measure the results. To create the bimodal system, templates and test feature vectors from both modalities were concatenated and matched in the same way as the unimodal cases.

In the first stage of evaluation, unimodal systems were tested under different conditions contemplated in the data augmentation scheme. Both modalities showed good results in terms of EER; for speaker recognition the highest value was 15.48 and the minimum was 4.36; in face recognition the highest value was 11.25 and the lowest, 2.46. Face recognition obtained the best results, while the speaker recognition results are more stable with a lower deviation value.

In the second stage, the multiple conditions of voice and face modalities were combined. The results showed significant improvements compared to the first stage, even for the most difficult conditions. The EER with the highest value was 5.1 and the lowest 1.82, which are lower than those obtained in the first evaluation. The stability of the results improved compared to unimodal face modality.

Acknowledgements

Atenco-Vazquez and Juan Carlos Moreno-Rodriguez acknowledge the financial support from the Mexican National Council for Science and Technology (CONACYT) to pursue doctoral studies.

References

Alay, N., & Al-Baity, H. H. (2020). Deep learning approach for multimodal biometric recognition system based on fusion of iris, face, and finger vein traits. *Sensors, 20*(19), 5523.
Andy, E. E., Bridget, M. O., Dono, O. K., & Benjamin, A. B. (n.d.). State of the art on face recognition-a.
Bashbaghi, S., Granger, E., Sabourin, R., & Parchami, M. (2019). Deep learning architectures for face recognition in video surveillance. In *Deep Learning in Object Detection and Recognition* (pp. 133-154). Springer.
Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1966). Recognition using class specific linear projection. In *European conference on computer vision* (pp. 43-58). Springer.
Bharathi, S., & Sudhakar, R. (2019). Biometric recognition using finger and palm vein images. *Soft Computing, 23*(6), 1843-1855.
Boulogne, F., Warner, J. D., & Yager, E. N. (2014). Scikit-image: Image processing in Python. *PeerJ, 2*, 453.
Chen, D., Liu, F., & Li, Z. (2020). Deep learning based single sample per person face recognition: A survey. *arXiv preprint arXiv:2006.11395*.
Cherrat, E. M., Alaoui, R., & Bouzahir, H. (2020). Convolutional neural networks approach for multimodal biometric identification system using the fusion of fingerprint, finger-vein and face images. *PeerJ Computer Science, 6*, e248.
Dahea, W., & Fadewar, H. S. (2018). Multimodal biometric system: A review. *International Journal of Research in Advanced Engineering and Technology, 4*(1), 25-31.

Dinca, L. M., & Hancke, G. P. (2017). The fall of one, the rise of many: A survey on multi-biometric fusion methods. *IEEE Access, 5*, 6247-6289.
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874.
Fierrez, J., Morales, A., Vera-Rodriguez, R., & Camacho, D. (2018). Multiple classifiers in biometrics. Part 1: Fundamentals and review. *Information Fusion, 44*, 57-64.
Irum, A., & Salman, A. (2019). Speaker verification using deep neural networks: A. *International Journal of Machine Learning and Computing, 9*(1).
Ke, P., Cai, M., Wang, H., & Chen, J. (2018). A novel face recognition algorithm based on the combination of LBP and CNN. In *2018 14th IEEE International Conference on Signal Processing (ICSP)* (pp. 539-543). IEEE.
Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication, 52*(1), 12-40.
Kortli, Y., Jridi, M., Al Falou, A., & Atri, M. (2020). Face recognition systems: A survey. *Sensors, 20*(2), 342.
Leghari, M., Memon, S., Dhomeja, L. D., Jalbani, A. H., & Chandio, A. A. (2021). Deep feature fusion of fingerprint and online signature for multimodal biometrics. *Computers, 10*(2), 21.
Li, R., Jiang, J. Y., Liu, J. L., Hsieh, C. C., & Wang, W. (2020). Automatic speaker recognition with limited data. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 340-348).
López, A. B. (2019). Deep learning in biometrics: A survey. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 8*(4), 19-32.
Luo, Z., Li, J., & Zhu, Y. (2021). A deep feature fusion network based on multiple attention mechanisms for joint iris-periocular biometric recognition. *IEEE Signal Processing Letters, 28*, 1060-1064.
Mandalapu, H., Reddy, P. N. A., Ramachandra, R., Rao, K. S., Mitra, P., Prasanna, S. R. M., & Busch, C. (2021). Audio-visual biometric recognition and presentation attack detection: A comprehensive survey. *IEEE Access, 9*, 37431-37455.
Masi, I., Wu, Y., Hassner, T., & Natarajan, P. (2018). Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)* (pp. 471-478). IEEE.
Minaee, S., Abdolrashidi, A., Su, H., Bennamoun, M., & Zhang, D. (2019). Biometrics recognition using deep learning: A survey. *arXiv preprint arXiv:1912.00271*.
Modak, S. K. S., & Jha, V. K. (2019). Multibiometric fusion strategy and its applications: A review. *Information Fusion, 49*, 174-204.
Moreno-Rodriguez, J. C., Atenco-Vazquez, J. C., Ramirez-Cortes, J. M., Arechiga-Martinez, R., Gomez-Gil, P., & Fonseca-Delgado, R. (2021). Biomex-DB: A cognitive audiovisual dataset for unimodal and multimodal biometric systems. *IEEE Access, 9*, 111267-111276.
Muckenhirn, H., Doss, M. M., & Marcell, S. (2018). Towards directly modeling raw speech signal for speaker verification using cnns. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4884-4888). IEEE.
Olazabal, O., Gofman, M., Bai, Y., Choi, Y., Sandico, N., Mitra, S., & Pham, K. (2019). Multimodal biometrics for enhanced IOT security. In *2019 IEEE 9th annual computing and communication workshop and conference (CCWC)* (pp. 0886-0893). IEEE.
Ravanelli, M., & Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 1021-1028). IEEE.
Ríos-Sánchez, B., Costa-da Silva, D., Martín-Yuste, N., & Sánchez-Ávila, C. (2020). Deep learning for face recognition on mobile devices. *IET Biometrics, 9*(3), 109-117.
Sabhanayagam, T., Venkatesan, V. P., & Senthamaraikannan, K. (2018). A comprehensive survey on various biometric systems. *International Journal of Applied Engineering Research, 13*(5), 2276-2297.
Snyder, D., Chen, G., & Povey, D. (2015). MUSAN: A Music, Speech, and Noise Corpus. *arXiv preprint arXiv:1510.08484v1*.
Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5329-5333). IEEE.
Sztahó, D., Szaszák, G., & Beke, A. (2019). Deep learning methods in speaker recognition: A review. *arXiv preprint arXiv:1911.06615*.
Wang, M., Wang, Z., & Li, J. (2017). Deep convolutional neural network applies to face recognition in small and medium databases. In *2017 4th International Conference on Systems and Informatics (ICSAI)* (pp. 1368-1372). IEEE.
Xin, Y., Kong, L., Liu, Z., Wang, C., Zhu, H., Gao, M., Zhao, C., & Xu, X. (2018). Multimodal feature-level fusion for biometrics identification system on iomt platform. *IEEE Access, 6*, 21418-21426.
Zhang, H., Qu, Z., Yuan, L., & Li, G. (2017). A face recognition method based on LBP feature for CNN. In *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)* (pp. 544-547). IEEE.
Zhang, X., Cheng, D., Jia, P., Dai, Y., & Xu, X. (2020). An efficient android-based multimodal biometric authentication system with face and voice. *IEEE Access, 8*, 102757-102772.