Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Vegad et al., J Biom Biostat 2017, 8:5
DOI: 10.4172/2155-6180.1000377

Research Article Open Access

Audio-Visual Person Recognition Using Deep Convolutional Neural Networks

Sagar Vegad[1], Harsh Patel[1], Hanqi Zhuang[2]* and Mehul Naik[3]

[1]*Department of Computer Science and Technology, Nirma University Ahmedabad, Gujarat, India*
[2]*Department of Computer and Electrical Engineering and Computer Science, USA*
[3]*Department of Electronics Communication Engineering, Nirma University, Ahmedabad, Gujarat, India*

Abstract

Protection of data integrity and person identity has been an active research area for many years. Among the techniques investigated, developing multi-modal recognition systems using audio and face signals for people authentication holds a promising future due to its ease of use. A challenge in developing such a multi-modal recognition system is to improve its reliability for a practical application. In this paper, an efficient audio-visual bimodal recognition system which uses Deep Convolution Neural Networks (CNNs) as a primary model architecture. First, two separate Deep CNN models are trained with the help of audio and facial features, respectively. The outputs of these CNN models are then combined/fused to predict the identity of the subject. Implementation details with regard to data fusion are discussed in a great length in the paper. Through experimental verification, the proposed bimodal fusion approach is superior in accuracy performance when compared with any single modal recognition systems and with published results using the same data-set.

Keywords: CNN; Face recognition; Mel-spectrogram; Multi-modal; Speaker recognition; VGG16 model

Introduction

The protection of any data, may it be physical or logical, has always been an important matter to the society. However, in the past, due to lack of technology, study materials, and references, the issue of the data protection still exists with us. But now the scenario is entirely different. People are trying to solve the issues which they think that are necessary for the society. One can develop a system which can give access only to the authorized person and reject all other unauthorized persons thus helps in achieving data protection.

In the last decade, researchers have tried building such systems which can give access to the data only to authorized person. Some of the researchers have used faces of the person in order to recognize whether the person is authorized or not. But considering only facial information of the person have not given the satisfying results. There can be many situation where an unauthorized person may try to fool the system by just providing a precaptured image of the authorized person. Extending their research, some researchers tried to build a system which gave access to the authorized person by recognizing the speech signals. But these systems do not necessarily perform well in every cases. There may be cases in which there exists a significant amount of noise while recording the speech of the person, which can affect the system performance greatly. Also, one can feed the prerecorded speech of the authorized person to the system in order to get access by fooling the system. There are realistic situations where automatic person recognition may fail. Other biometrics proposed recently includes lip movement, retina movement, jaw movement, part of the tongue, etc. However, due to the elusiveness of such biometrics and inappropriateness of asking each person in some cases for such biometrics, these biometrics have not been popular.

Hence, to compensate the cons of one biometric, researchers moved on to multimodal biometric systems. The multimodal recognition system is the one in which the system uses more than one biometric of the person. A bimodal system is nothing but using two biometric features of the person to recognize whether he/she is authorized or not. In bimodal recognition system, speaking process may include speech as well as the face of the speaker; i.e. all the system needs, is a video of a person speaking something, from which the system can extract the facial data as well as speech data automatically and then start the processing of recognizing the person. Face and speech biometric features are the most favorable and opted one in the case of bimodal recognition systems due to availability of the video (data). One can also use the finger print features because they can be get easily too. But we need one more separate system which processes the finger prints as finger prints can't be extracted from a video.

In this paper, we present an efficient model which can lead to build a highly accurate and robust bimodal speaker recognition system. We propose an audio-visual system as shown in Figure 1 which uses Convolution Neural Networks (CNN). In this proposed approach, we first extract the audio features from the audio files. We then extract the facial features from the face images of the speakers. After extracting both the features, we feed them into two separate CNN models for training. After training, we fed our testing data into our trained models and collected the predicted outputs of both the facial and speech models. Then we fused these outputs of both the models to predict the final output. This type of fusion is called decision fusion. Other fusion types include feature fusion and sensor fusion, which happen at early stages before the models get trained. In feature fusion, both the features are combined into a single feature vector and then a single model is trained.

We present our approach in this paper as follows: Section II gives

***Corresponding author:** Zhuang H, Department of Computer and Electrical Engineering and Computer Science, USA, Tel: +1 561-297-3000; E-mail: zhuang@fau.edu

Received October 17, 2017; **Accepted** October 25, 2017; **Published** October 30, 2017

Citation: Vegad S, Patel H, Zhuang H, Naik M (2017) Audio-Visual Person Recognition Using Deep Convolutional Neural Networks. J Biom Biostat 8: 377. doi: 10.4172/2155-6180.1000377

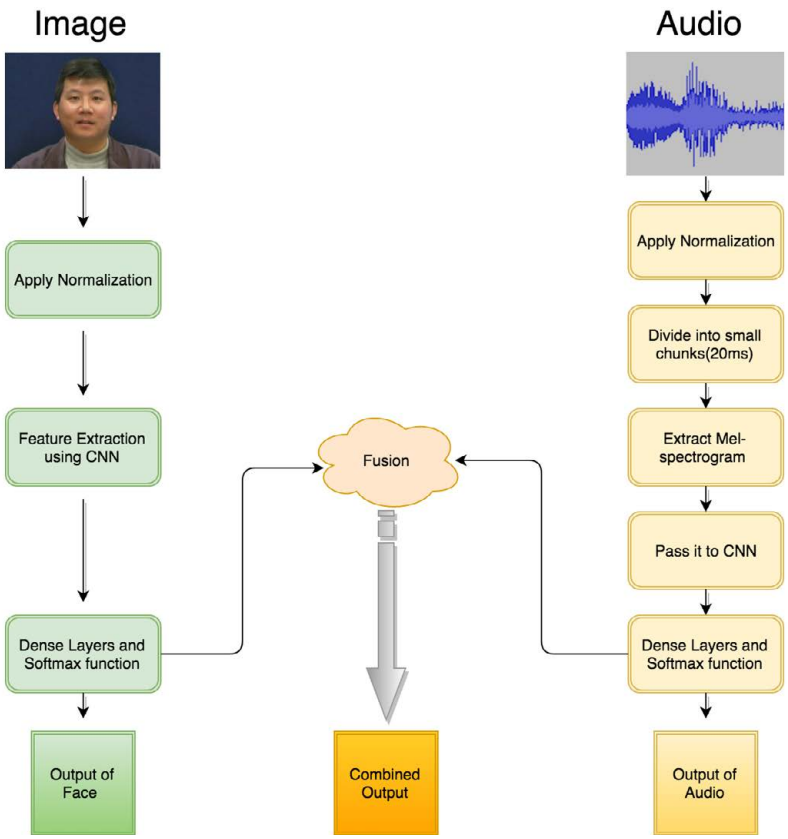

Figure 1: Overview of proposed multi-modal system.

the insights about the previous research related to this field of study. Section III delineates about the models and architecture used in the process, the information about the database, results and analysis of our research is given in Section IV, which is followed by conclusion in Section V.

Previous Works

Over the last few years, much research had been carried out regarding face recognition like Eigenface [1], Fisherface [2] and Viola and Jones face detector [3]. Wang and coworkers used the combination of HoG [4] and LBP [5] to increase the performance [6]. For extracting facial features, many researchers had tried conventional techniques like Marginal Fisher Analysis [7], and Locality Preserving Projection [8], which achieved a good performance with rooms for improvement. For speaker recognition, Gaussian mixture models (GMMs), eigenvoices [9], MFCC features [10] had shown great results. For the integration of audio and face image features, training on a single model is done by Chibelushi, et al. [11]. Multi-stream hidden Markov model (HMM) [12] has performed well for model fusion. Geng, et al. suggested a number of ways for biomodal data fusion [13]. They used CNNs for extraction for face image features; and for audio feature extraction, they used MFCC features. In feature fusion, they merged both the facial and MFCC features and fed the integrated feature vectors into the next layer. In decision fusion, they had trained two separate models for two different kinds of features and then applied softmax function separately to both the models, merged their results and predicted the final output.

Lin and Song used eigenvalues with principal component analysis (PCA) to reduce the dimensionality of face images in ref. [14]. After that radial basis function (RBF) neural network was used to compute the recognition scores of each class which helped to detect the user's identity. They designed two stages for speaker recognition. In the first stage, the confidence measure calculated according to the recognition result helped in justifying the recognition. In the second stage, a recognition score was computed to the valid recognitions. The recognition score was measured by using Gaussian mixture model (GMM), Universal background model (UBM), and Maximum normalization. For the fusion strategy, the computed confidence index was compared with the threshold. If no confidence indices exceeded the limit, then there will be no result. If one exceeds the threshold value then that one is chosen as the result. Idf more than one exceeds the threshold values, the linear combination of the face and the speaker recognition results is used as the output.

In the next section, we will briefly overview Convolution Neural Network since it is used in our research.

Deep Convolution Neural Network (CNN)

CNN is a type of neural network which consists of neurons having weights and biases associated with it. They take inputs and multiply them with learned weights and add biases and converts them through a non-linear activation function, which is described below, in order to reduce the error of the cost function.

The underlying architecture of CNN consists of 3 layers, which are Convolution Layer, Pooling Layer, and Fully-connected Layer. A Convolution Layer calculates the dot product of the input and weight values by using a window of a particular size, and a bias value is added to it. This sliding of the window helps in detecting visuals like edges or corners.

After the Convolution Layer, an activation function is applied to produce the output for the next hidden layer, given the set of inputs of the previous layer. There are various activation functions such as sigmoid, relu, and tanh. We use the RELU activation function in our model which works by thresholding each value at 0; i.e. RELU function, $f(x)=max(0,x)$.

The next is the Pooling Layer, which will perform a downsampling operation along the spatial dimensions (width, height), resulting in the volume less than that of previous layer. At last, there is a Fully-connected Layer that calculates the output/prediction, performs the backpropagation and then updates the weight accordingly. In the next section, we will explain about the models which we have used for face recognition, speaker recognition and their integration.

Model

In our project we have applied decision fusion, i.e. building two separate models for speaker recognition, one with the facial dataset and another with the audio dataset of the speakers, followed by fusing the outputs of these different models to predict the final output. To start with the analysis of our project, we first present these two separate models and then explain the fusion process in the following text.

Person recognition using audio signals

Figure 2 shows the mel-spectrogram which is the representation of a power spectrum of an audio of a person. We normalized the audio and removed the background noise using Audacity tool. We then divided each audio into small chunks of 20 ms with 5% overlap. We calculated the log scaled mel-spectrogram of each chunk using the librosa library implementation, with the window size of 448, hop length of 32, and 60 mel-bands. These log scaled mel-spectrograms are then concatenated with their delta features, delta-delta features, and passed them as an input to the CNN model. Figure 3 shows the architecture of our model. The output of the CNN model will give the probabilities of each class, and the maximum of them would be the predicted output.

For the CNN model, we used three Convolutional layers, one Max-pooling, two fully connected layers, and an output layer.

- The first convolutional layer consisted of 64 filters with the rectangular shape of (19 × 2 size, 1 × 1 stride) followed by Max-pooling layer with pool size 2 × 2 and stride 1 × 1.
- The second convolutional layer consisted of 128 filters with the rectangular shape of (11 × 1 size, 1 × 1 stride) followed by another convolutional layer consisted of 256 filters with the rectangular shape of (3 × 3 size, 1 × 1 stride).

After flattening the input from the previous layer, two fully connected hidden layers were used with 4096 neurons and between each of them, a dropout layer with 50% probability was used.

And finally, the output layer with the softmax activation function. After each layer, relu activation function was used, and batch normalization [15] was done for careful tuning of weights and to reduce the internal covariant shift. For the first 300 epoch, we kept 0.0002 learning rate with 0.9 momentum and 1e-6 decay. And after that for 100 epochs, we kept 0.00002 learning rate for fine-tuning and used Stochastic gradient descent (SGD) optimizer to reduce the error.

Our proposed model divides an audio segment into different number of frames, and our goal is to classify the whole audio into the class which it belongs. So we have combined the 1-D list output of each frame into a 2-D list with each column representing the class number and each row representing the frame number. Further, by taking the column-wise average of the output of the frames of every audio, we have converted it back to 1-D list.

To generalize, let n be the total number of frames for each audio and i ranges from 1 to n. The output of the audio model will be *P_audio (i)=[pi1, pi2, ..., pi42, pi43],]* where *i* ranges from [1,n]. Now, using the below equation, we can convert this 2-D output into a 1-D output.

$$P\_audio = [(\sum_{i=1}^{n} pi1)) / n, (\sum_{i=1}^{n} pi2)) / n, ..., (\sum_{i=1}^{n} pi43)) / n] \quad (1)$$

where, *pi1* is the probability predicted by the model for class 1, *pi43* is the probability predicted by the model for class 43 and *P_audio* is the list of probability for every class of one audio.

Person recognition using face images

We normalized the images by dividing it with 255. The initial size of an image was 512 × 384, so we converted that to 224 × 224 pixels and kept RGB channel. Figure 4 shows the sample images of our dataset. We passed these images to a CNN for feature extraction and then to fully connected hidden layers which gave the probabilities of each class as the output. We used VGG16 model [16] with weights pre-trained on ImageNet. We extracted features of the images from the top 18 VGG16 layers and then passed that to the two dense layers and one output layer to compute the probabilities of each class. Figure 5 shows an overview of our model. We kept the same parameters for the first 18 layers of VGG16 architecture and for the rest of the layers. We kept 512 filters with 50% probability dropout for the next two fully connected hidden layers with relu activation function. We kept learning rate 0.0005 with 0.9 momentum, 1e-6 decay and used SGD optimizer.

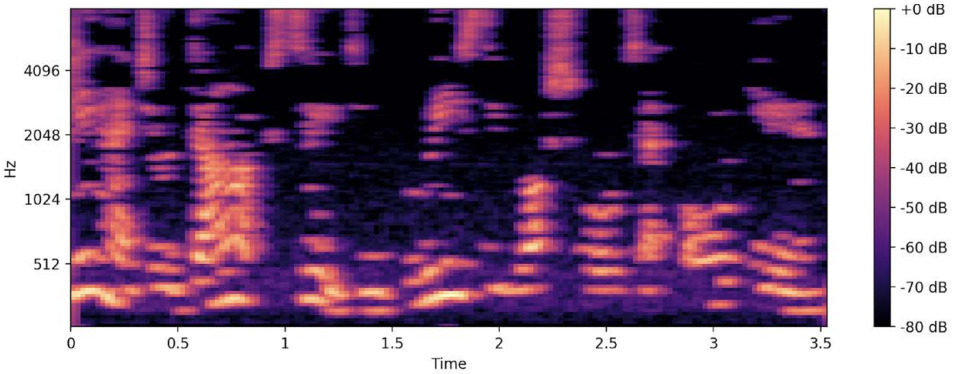

Figure 2: Mel-spectrogram of audio.

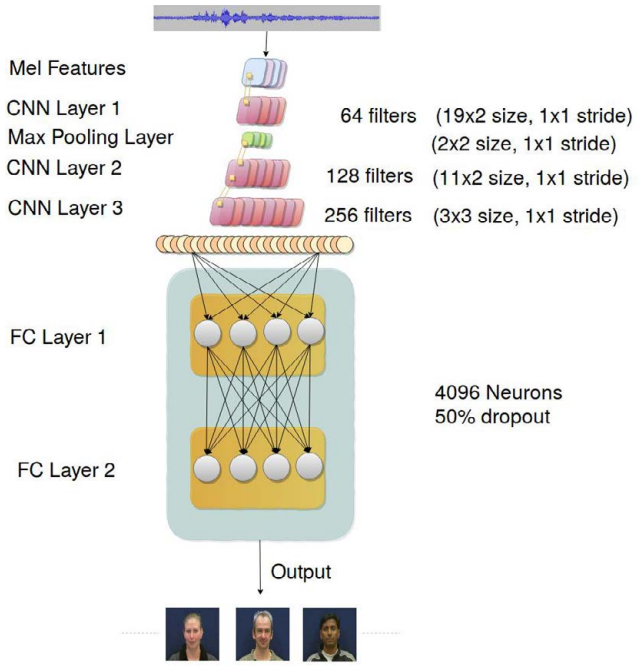

Figure 3: Architecture of our model for speaker recognition.

Figure 4: Sample images for face Dataset.

Fusion of both models

Output of any model can be categorized [17,18] into three parts:

1. Abstract level output: The output of the model produces the label values for each of the testing data.
2. Rank level output: The output of the model generates a subset of label values having ranks associated with them for every testing data.
3. Measurement level output: The output of the model consists of a list/vector of length equal to the number of classes and the values of this list/vector represents the probabilities of the testing data to which class it belongs.

Both the output of our models used in above two subsections produce the measurement level output. The fusion process for such type of outputs can be done through two ways namely Class-conscious and Class-indifferent as described in ref. [19]. The former consists of many different methods to combine the probabilistic outputs; examples of these methods are Sum rule, Max rule, Min rule, Average rule, Product rule, Weighted Average rule, etc. The later consists of two methods namely Decision Template and Dempster-shafer combination. We have applied the Class-conscious method for our fusion process which includes the Weighted Average Rule. In such a method, we just compare the output list of both the models and pick the output accordingly to make the final output list. The Weighted Average Rule can be visualized in Figure 6. In this figure Face model block has been described in Figure 5 and Audio model block has been shown in Figure 3. In the Fusion block, eqn. (2) is applied, which is described below. The

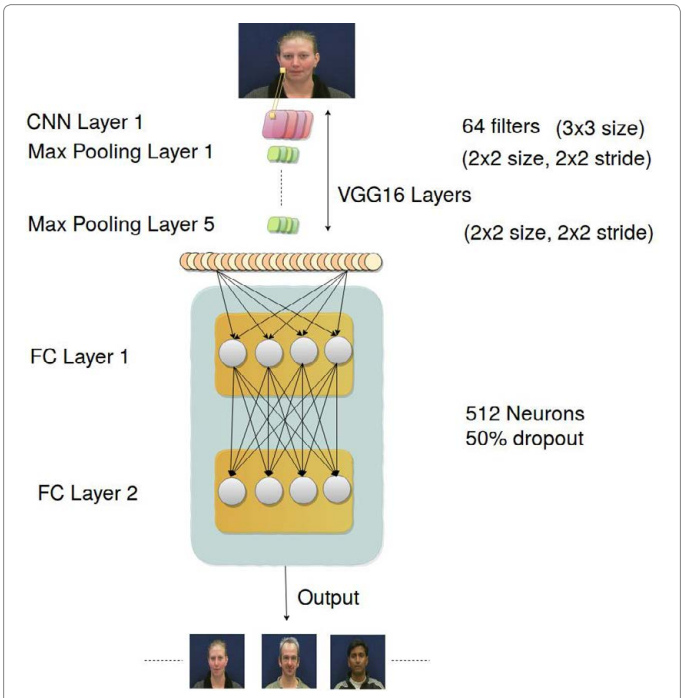

Figure 5: Architecture of our model for face recognition.

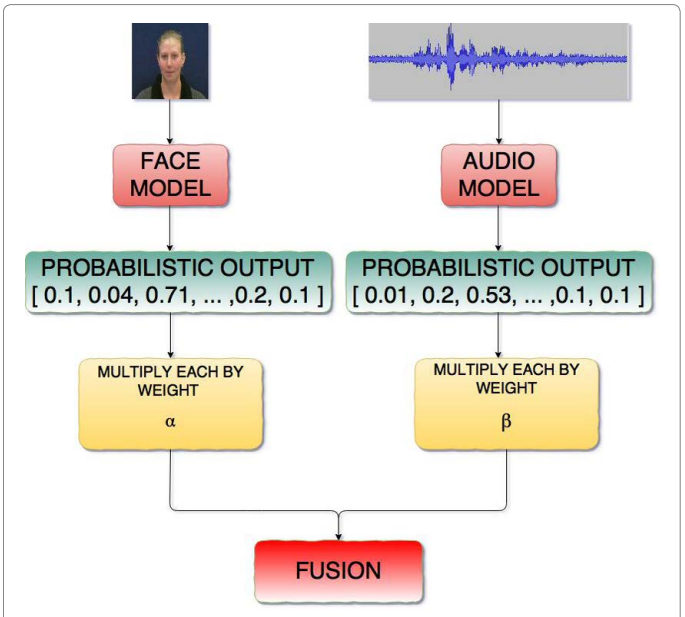

Figure 6: Diagram of our model for Fusion.

class with the maximum probability will be considered the final output of this multimodal system.

We have taken the weighted average of the output probabilities of both the models i.e. of speaker and face recognition model. We did the linear combination of the output probabilities of each model, multiplied with their adjusted weights and then combined them.

$$P\_comb(i) = (P\_audio[i] * \alpha + P\_face[i] * \beta) / (\alpha + \beta) \quad (2)$$

where, i ranges from [1,class length], is the adjusted weight for audio, is the adjusted weight for face, *P_audio[i]* is the probability list for *ith*-audio, *P_face[i]* is the probability list of *ith*-face and *P_comb[i]* is the final probability list of the *ith*-audio and face after combining both of them. For each class, the index of the maximum of all the probabilities is considered to be the final output.

Database and Experimental Analysis

We have chosen VidTIMIT database [20] for our research project. This dataset was made by taking videos of 43 people speaking short sentences and also with some head movements. The dataset was primarily taken in 3 sessions, having an average delay of 7 days and 6 days between session 1 and session 2, and session 2 and session 3 respectively. Each of the 43 speakers speaks exactly 10 sentences, and they were distributed into 6 sentences for session 1, next 2 sentences for session 2 and the remaining 2 sentences for the last session. Also, every first and second statement of every speakers were identical and the remaining 8 statements were different. As mentioned previously, each speaker rotated their head to the left, right, back to the center and then down and finally back to the center in each session.

The dataset was recorded in an office milieu with the help of the broadcast quality digital video camera. Both the audio and image files from each video of all the speakers werte extracted. The images of faces have a resolution of 512 × 384, and the audio is stored as mono, 16 bit, 32 kHz WAV file.

Analysis for audio

In the experiment, we shuffled the data and divided it into 80% training and 20% testing. We used the text-independent approach, which means that we tested our examples without any constraint on speech content. We compared our results to those obtained in ref. [21]. The authors [21] also worked on VidTIMIT dataset and used MFCC and DT-CWPT methods. Table 1 shows the accuracies when we passed only mel-spectrogram to CNN, mel-spectrogram combined with its delta (**Δ**) features as a 2-channel input and mel-spectrogram combined with delta and delta-delta (**Δ-Δ**) features as a 3-channel input.

Analysis for face

We divided the data into 50% training and 50% validation. We compared our results [21]. Table 2 shows the accuracies of the face recognition attained by our method and approaches like Eigenface PCA, 2D-DWT PCA, and DT-CWT PCA described in detail [21].

Method	Features	Accuracy	Accuracy, (Mel-band=128)
-	MFCC	60%	-
-	DT-CWPT	65%	-
1-ch input	Mel-spect+CNN	79.49% (Mel-band=60)	75.25%
2-ch input	Mel-spect+$_{\Delta}$+CNN	80.66% (Mel-band=60)	76.35%
3-ch input	Mel-spect+$_{\Delta}$+$_{\Delta}$-$_{\Delta}$+CNN	80.40% (Mel-band=60)	75.53%

Table 1: Accuracy for speaker recognition.

Method	Accuracy
Eigenface PCA	70%
2D-DWT PCA	75%
DT-CWT PCA	91%
Our method	93.53%

Table 2: Accuracy for face recognition.

Method	W_Audio	W_Face	Combined Accuracy
Cluster	-	-	87%
Cluster	-	-	90%
-ch audio+face	0.2	0.8	94.84%
-ch audio+face	0.2	0.8	95.18%
-ch audio+face	0.2	0.8	94.99%
-ch audio+face	0.3	0.7	95.60%
-ch audio+face	0.3	0.7	96.05%
-ch audio+face	0.3	0.7	95.79%
-ch audio+face	0.4	0.6	96.03%
2-ch audio+face	0.4	0.6	97.33%
-ch audio+face	0.4	0.6	96.17%
-ch audio+face	0.5	0.5	95.86%
-ch audio+face	0.5	0.5	96.12%
-ch audio+face	0.5	0.5	95.01%
-ch audio+face	0.6	0.4	90.52%
-ch audio+face	0.6	0.4	90.44%
-ch audio+face	0.6	0.4	90.67%

Table 3: Accuracy for fusion.

Analysis for fusion

Table 3 shows the combined accuracies with different weights for speaker and face models having mel-bands=60. In this table we also compared our results [21], where the feature fusion method was also applied. The authors [21] used the K-means algorithm for 32 clusters and 64 clusters.

For the purpose of verification of our model, confusion matrices are constructed. A confusion matrix is a tabular representation where the values of correct predictions and incorrect predictions are shown with rows labeled as correct output and columns labeled as predicted output. A Confusion Matrix is also sometimes referred to as an error matrix. Confusion matrices for person recognition using audio, person recognition using face images and person recognition with fusion are shown in the Figures 7-9 respectively.

The interpretation of confusion matrices goes like this. Let's say cell (1,1) has a value equal to x. Then, one can say that there are x cases in which the correct output was class 1 and predicted output was also class 1. Hence, this is a favorable cell. Similarly, for cell (1,2), one can say that there are some cases where correct output was class 1, but the model predicted the output to be class 2. Hence, this is an unfavorable cell. On generalizing the above statements, one can easily assert that all cells having row number equal to column number or cells *(i, i)*, where *i* ranges from [1, class length], are all favorable cells while the rest are unfavorable cells.

As shown in the Figures 7-9, we observe that the values in the cells *(i, i)*, where *i* ranges from (1, 43) in the confusion matrix for fusion are higher than that for the audio and face individually. So, we can assert that the fusion process helped us in increasing the accuracy of our model.

A ROC curve is just another method to evaluate any model in machine learning. ROC stands for Receiver Operating Characteristic.

Act/Pred	1	2	3	..	..	..	41	42	43
1	0.757	0.015	0.03	..	..	..	0	0.015	0
2	0	0.702	0	..	..	..	0	0	0
3	0	0.021	0.782	..	..	..	0	0	0
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
41	0	0	0	..	..	..	0.547	0	0
42	0	0	0	..	..	..	0	0.893	0
43	0	0	0	..	..	..	0	0.022	0.667

Figure 7: Confusion matrix for audio model.

Act/Pred	1	2	3	..	..	..	41	42	43
1	1	0	0	..	..	..	0	0	0
2	0	0.851	0	..	..	..	0	0	0
3	0	0	0.608	..	..	..	0	0	0
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
41	0	0	0	..	..	..	0.809	0	0
42	0	0	0	..	..	..	0	1	0
43	0.044	0	0	..	..	..	0	0	0.822

Figure 8: Confusion matrix for face model.

Act/Pred	1	2	3	..	..	..	41	42	43
1	1	0	0	..	..	..	0	0	0
2	0	0.978	0	..	..	..	0	0	0
3	0	0	0.804	..	..	..	0	0	0
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
41	0	0	0	..	..	..	0.88	0	0
42	0	0	0	..	..	..	0	1	0
43	0	0	0	..	..	..	0	0	0.977

Figure 9: Confusion matrix for fusion model.

A ROC curve is a graph between True Positive Rate on the y-axis and False Positive Rate on the X-axis. The area under the curve (AUC) for a ROC curve should be as high as possible. The maximum possible value of AUC can be equal to 1. Hence, the ideal point on any ROC curve is the top-left-most point. The ROC curve for our proposed fusion model is shown in Figure 10 and has obtained 0.98 value for AUC.

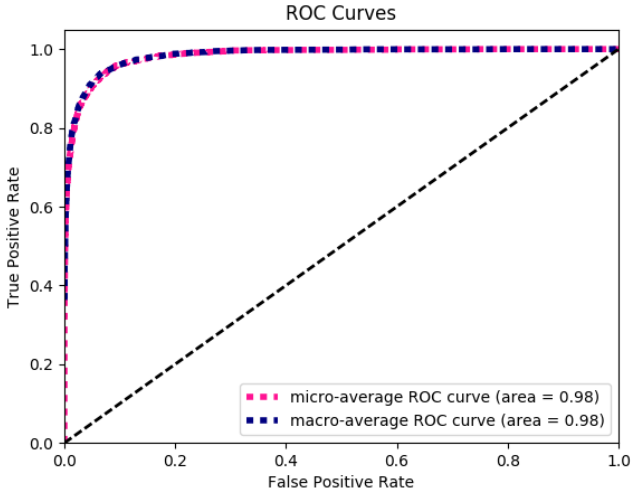

Figure 10: ROC curve for fusion.

Conclusion and Future work

In this paper, we have presented a bimodal person recognition system which attained better accuracy compared with a single model audio or face recognition system. We described the architecture of our models for audio, face and integrating both of them using a Convolution Neural Network (CNN). By using only audio recognition system, we obtained an accuracy of 80.66%; by using only face recognition system, we obtained an accuracy of 93.53%, and by fusing both, we obtained an accuracy of 97.33% on VidTIMIT dataset. Our future work is aimed at improving the multi-modal system by integrating more biometric modals, say iris scanning data. We hope that the proposed approach will make person identification system more robust and practical.

References

1. Turk M, Pentland A (1991) Eigenfaces for recognition. Journal of cognitive neuroscience 3: 71-86.
2. Kwak KC, Pedrycz W (2005) Face recognition using fuzzy fisherface classifier. Pattern Recognition 38: 1717-1732.
3. Viola P, Jones M (2001) Rapid object detection, using a boosted cascade of simple features.
4. Dalal N, Triggs B (2005) Histogram of oriented gradients for human detection.
5. Zhang H, Gao W, Chen X, Zhao D (2006) Object detection using spatial histogram features. Image and Vision Computing 24: 327-341.
6. Wang X, Han TX, Yan S (2009) An HoG-LBP human detector with partial occlusion handling.
7. Yan S, Xu D, Zhang B, Zhang HJ (2005) Graph embedding: A general framework for dimensionality reduction. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2: 830-837.
8. He X, Yan S, Hu Y, Niyogi P, Zhang HJ (2005) Face recognition using laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 27: 328-340.
9. Thyes O, Kuhn R, Nguyen P, Junqua JC (2000) Speaker Identification and verification using eigenvoices.
10. Murty K, Yegnanarayana B (2006) Combining evidence from residual phase and mfcc features for speaker recognition. IEEE Signal Processing Letters 13: 52-55.
11. Chibelushi CC, Mason JS, Deravi F (1997) Feature-level data fusion for bimodal person recognition. IEEE International Conference on Image Processing and Its Applications 3: 399-403.
12. Gunawan Sugiarta YB, Bambang R, Suhardi H (2010) Feature Level Fusion of Speech and Face Image based Person Identification System. Second International Conference on Computer Engineering and Applications.
13. Geng J, Liu X, Cheung Y (2016) Audio-visual Speaker Recognition via Multi-modal correlated Neural Networks. IEEE/WIC/ACM International Conference on Web Intelligence Workshops.
14. Lin C, Song KT (2012) User Identification Design by Fusion of Face Recognition and Speaker Recognition. International Conference on Control, Automation and Systems in ICC, Jeju Island, Korea.
15. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. Computer Science.
16. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Vision and Pattern Recognition ICLR arXiv: 1409.1556.
17. Xu L, Krzyzak A, Suen CY (1992) Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans Systems Man and Cybernetics 22: 418-435.
18. Kuncheva LI (2004) Combining Pattern Classifiers: Methods and Algorithms. Wiley Inter-science.
19. Mangai UG, Samanta S, Das S, Chowdhury PR (2010) A Survey of Decision Fusion and Feature Fusion Strategies for Pattern Classification. IETE Technical Review.
20. Sanderson C, Lovell BC (2009) Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. Lecture Notes in Computer Science (LNCS). 5558: 199-208.
21. Dupont S, Luettin J (2000) Audio-visual speech modelling for continuous speech recognition. IEEE Transactions on Multimedia 2: 141-151.