Multimodal Biometric Verification Using Deep Neural Network

Mohammad Pournazeri (06487506)   Carlo Provenzani (06442416)

Abstract:

In this study two different Siamese networks are initially proposed for face and voice recognitions. The proposed networks consist of convolutional layers followed by fully connected layers. For face recognitions the face images data are directly fed into the model; however, for voice recognition, the voice records are converted into spectrogram images using Fast Fourier Transform (FFT) and then these images are used for training and prediction. The two models are then combined together for simultaneous voice and face recognition.

Introduction:

Personal identification is usually done by verification of the picture ID card. However, there are many situations that personal identification is either compromised or impossible. These situations could be due to different types of face occlusions such as hair, glasses, masks (during COVID-19 period), scarf or when the person is not physically present for verification (telephone banking, online shipping, etc). To prevent the above security breaches, concurrent voice and face recognition can significantly help to identify the person/user identity using camera and microphone. Siamese network is a one-shot learning algorithm and has shown significant robustness and accuracy in the area of image recognition. This technique is based on finding the level of similarity between the two images by comparing their feature vectors. The image feature vector is computed by passing the image through a convolutional neural network followed by a fully connected network. To recognize an image, its feature vector is first calculated by passing it through the trained Siamese network and then comparing it with those in our database. The absolute distance vector between the calculated feature vector and that one from the database is the passed through logistic regression model with a single output (true or false).

Related Works:

In (1) a multi-stage approach was proposed that aligns faces to a general 3D shape model. A multi-class network is trained to perform the face recognition task on over four thousand individuals. A so called Siamese network was also tested where they directly optimize the L1-distance between two face features. In (2), a unified system for face verification, face recognition and clustering was presented. The method is based on learning a Euclidean embedding per image using a deep convolutional network. The network is trained such that the squared L2 distances in the embedding space directly correspond to face similarity. In this method, the Triplet Loss minimizes the distance between an anchor and a positive image from the same person, and maximizes the distance between the anchor and a negative image of different persons.

Speaker verification systems can either be categorized as text-dependent or text-Independent. Text-dependent systems are constrained to a single key word or phrase across all users, whereas text-independent systems have no constraints on the speech content (user can speak freely), but are considered more challenging because requiring a lot longer training and testing to achieve a good performance. Siamese Networks have shown to be very useful in this area, most recent researches on Siamese networks has however been focused and proven successful on voice biometrics and speaker identification (3), (4).

In (5), Luo et al. utilized CDRNN technique initially to convert the speakers speech to a spectrogram and then they used RNN to identify the speaker. In (6), Venayagamoorthy et al. utilized digital signal processing to generate the voice pattern and then used neural network to identify the person. In (7), Chowdhury et al. explored multiple fusion schemes to combine face and speaker recognition to perform effective person recognition on audio-video surveillance data using a consumer-grade camera with a built-in microphone.

Datasets:

The database used in this study for face recognition is ColorFERET database from National Institute of Standards and Technology (www.nist.gov/humanid/colorferet). This database contains a total of 11338 facial images taken by photographing 994 subjects at various angles, facial expressions and lights over the period of 15 sessions between years 1993 and 1996. The database was created to develop, test and evaluate face recognition algorithms. The 11338 facial data was rearranged into 22676 groups of two pictures. Half of these groups include any two pictures of a same person and the second half include the pictures of two different persons as shown in Figure 2.

The dataset used for the voice recognition comes from LibreSpeech (http://www.openslr.org/12), a corpus of around 1000 hours of 16kHz derived from audiobooks. The training data consists of 120,000 pairs of audio samples from 300 plus hours of clean English speaking from more than 900 different speakers in controlled environments with no external noise. Like the training data, the test data consists 2,500 audio samples from 5 plus hours of English speaking from 20 different individuals. The length of each audio sample that is fed as input into the model is one second long, so one second long audio recordings are extracted from 360 hours datasets and have been processed and converted into spectrogram images using an FFT size of 512 to shows frequency changes over time and represent them as one channel 2D images.

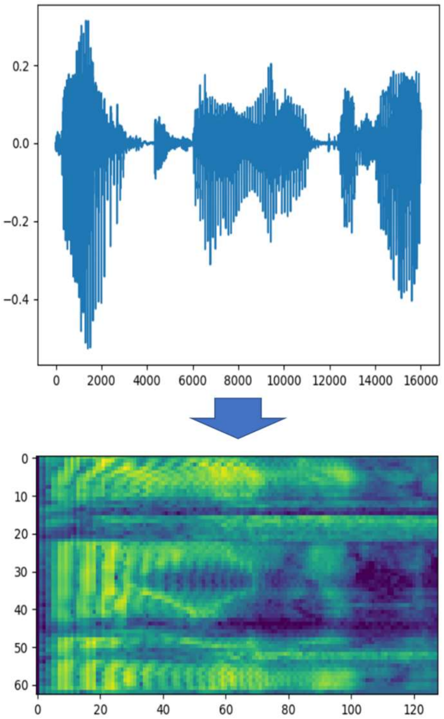

Figure 1: Conversion of voice data to 2d image data using Fast Fourier Transform with size of 512

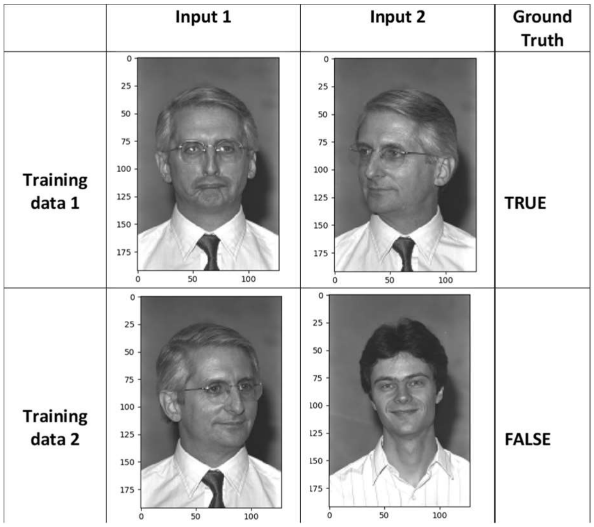

Figure 2: An example of the training dataset for this study

Method:

In this study, as shown in Figure 3, an end-to-end Siamese network is used for face and voice recognition. In this approach two parallel convolutional neural networks followed by a fully connected network are used to determine the feature vector for each image input. The two parallel networks are identical and share the same parameters. The absolute difference (still a vector) between the the two feature vectors is then calculated and passed through another fully connected layer with sigmoid activation function. The model output is in fact the probability of the two images that are from the same person.

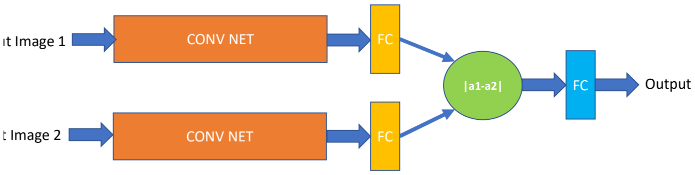

Figure 3: Siamese Network schematic used in this study

In this study the face recognition model architecture and its hyperparameters are shown in Table 1 and Table 2. Since the proposed model is kind of logistic regression model, cross entropy loss function is used.

.Table 1: Suggested Face Recognition Model Architecture

Input Image	192 x 128 x 3	Activation Function
Convolution Layer 1	5 x 5 x 32 (stride 1 and no padding)	ReLu
Max Pool Layer 2	2 x 2 x 32 (stride 2)	
Convolution Layer 3	5 x 5 x 32 (stride 1 and no padding)	ReLu
Max Pool Layer 4	2 x 2 x 32 (stride 2)	
Convolution Layer 5	5 x 5 x 64 (stride 1 and no padding)	ReLu
Max Pool Layer 6	2 x 2 x 64 (stride 2)	
Convolution Layer 7	5 x 5 x 64 (stride 1 and no padding)	ReLu
Max Pool Layer 8	2 x 2 x 64 (stride 2)	
Flatten Layer 9		
Fully Connected Layer 10	128 neurons	Sigmoid
Absolute Feature Difference Layer 11		
Fully Connected Layer 12	Scalar Output (0-1)	Sigmoid

Table 2: Suggested Face Recognition Model Hyperparameters

Hyperparameter	**Value**
Learning Rate	1e-4
Mini Batch Size	200
Optimization	Adam, 0.9, 0.99

For voice recognition, the one-second long audio samples wave are used to train a simple Siamese Neural Network as follows:

- 2D convolution with 32 filters (5x5) followed by max pooling and dropout.
- 2D convolution with 64 filters (4x4) followed by max pooling and dropout.
- 2D convolutions with 128 filters (3x3) each followed by max pooling and dropout.
- 2D convolutions with 128 filters (2x2) each followed by max pooling and dropout.
- Dense layer of size 1024

Each Convolution layer in this network uses the ReLU activation function while the dense layers use sigmoid. The two identical weight sharing networks are then joined together by another layer for calculating the Euclidean distance between the two samples of size 1024.

Preliminary Models:

Both face recognition and voice recognition models are created using sequential modeling using Tensorflow Keras. The proposed face recognition model is fed with 19000 training data and around 1600 validation data each consist of two images. The model is run for 100 epochs and the results are shown in Figure 4. As shown in confusion matrix, Figure 5, the model validation accuracy is around 92.7%.

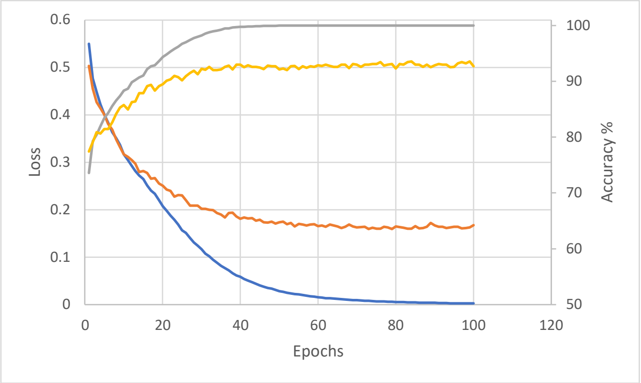

Figure 4: Training and validation for proposed Face Recognition network

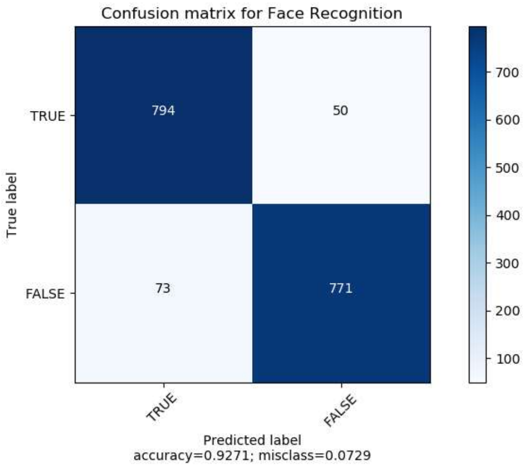

Figure 5: Confusion Matrix for Face Recognition Model

For voice recognition, the Siamese Network was trained in batches of size 32, each batch consisting of two spectrograms. Half of the batch being samples from the same speaker and the other half being samples from two different speakers. The model was trained for 10 epochs using Adam optimization with a learning rate of 0.00005. Figure 6 shows the loss and accuracy for both training and validation data.

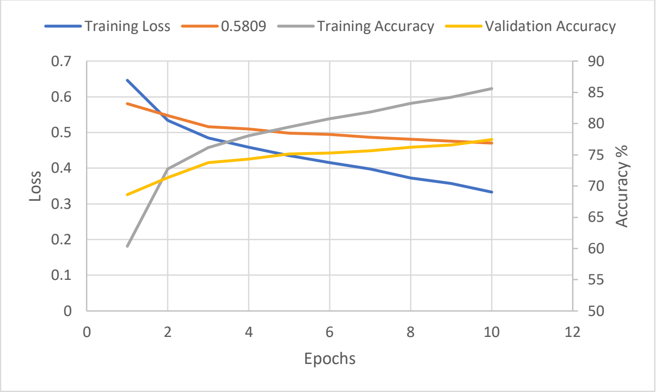

Figure 6: Voice Recognition During Training

Further Improvement in Face Recognition Model:

One of the main problems with the dataset used for training and testing the discussed face recognition algorithm is that all the photos in the dataset were all taken in the studio and as the result they are all from the same distribution. Therefore, when the trained algorithm was tested with some other photos from different distribution, the recognition performance was significantly lower than when tested with the same distribution of photos used during training. The new set of photos were tested using the proposed model trained with original dataset and only 60% of these new images were identified correctly.

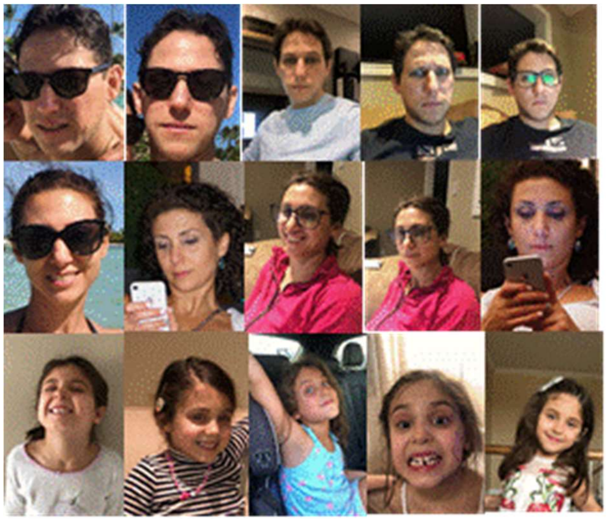

Figure 7: Sample of photos from different distribution

To mitigate this problem following strategies were used:

- A set of photos taken in the wild was added to both training and testing data.
- Dropout with ratio of 20% was added to every convolutional layer to prevent overfitting
- All the faces in the training photos were cropped and the cropped version of images were used for training

The face recognition model was trained with the above changes and the training and validation results in Figure 8 show slightly lower variance (less overfitting) but higher bias.

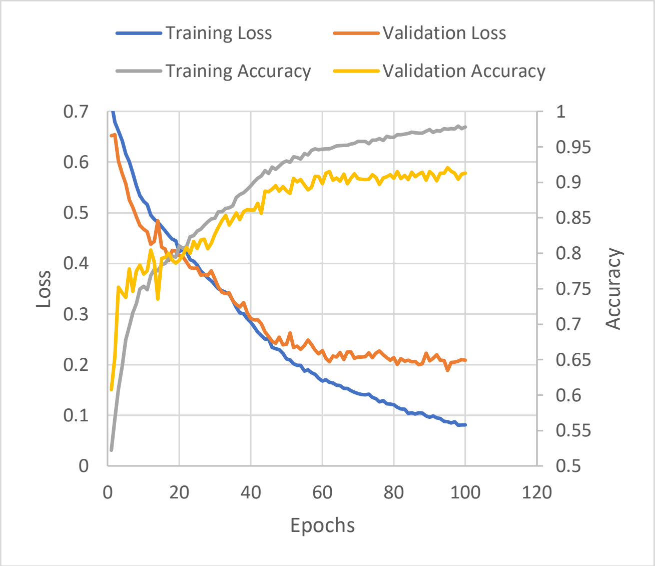

Figure 8: Training and validation of the Face Recognition Network with 20% dropout and added dataset from different distribution

Face Detection:

One of the steps in face recognition is detecting the face in the image using bounding box. The detected face is then cropped using the bounding box dimensions and the cropped face is then used for face recognition.

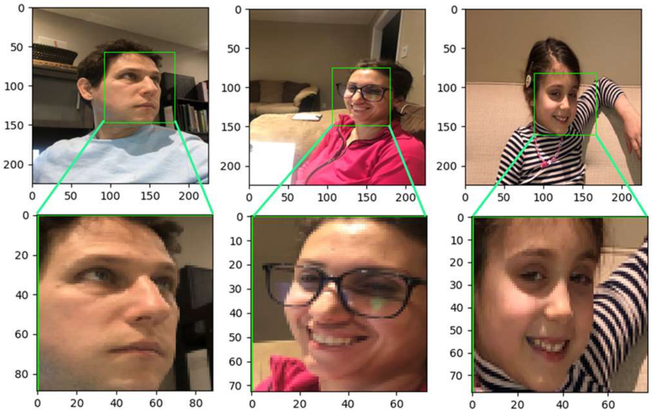

Figure 9: Face detection and cropping using Haar object classifier in openCV

To correctly detect and identify faces, HAAR Cascade Classifier (OpenCV) has been used. This classifier was proposed by Paul Viola and Michael Jones in their paper, "Rapid Object Detection using a Boosted Cascade of Simple Features" in 2001. It is an algorithm that needs a lot of positive images of faces and negative images without faces to train the classifier. It works by extracting and considering adjacent rectangular regions at a specific location in a detection window (HAAR features), summing up the pixel intensities in each region and calculating the difference between these sums.

Further Improvement in Voice Recognition Model:

Too improve the developed voice recognition algorithm, the model was then retrained with three-second-long audios, down sampled by a factor of three, making the spectrogram the same size as the previous training attempt of one-second-long audios. On this attempt, additionally, noise (non-speaking) was introduced to the training data, using samples from the Clotho audio captioning dataset, which contains various recorded audio files accumulating to about 3.4GB in size.

The model, however, remained equivalent to the previous trials and was retrained on randomly selected batches of size 32, the same as before. This time, half similar and half dissimilar pairs for 30 epochs were used, each epoch containing 300 batches, and validation data consisted of 100 batches of size 32. Adam optimization was used with a learning rate of 0.0005. As a result, this Siamese classifier trained on three-second-long audios outperforms the classifier trained on only in second-long audios, as hoped. With about 88% Validation accuracy, optimal performance is not yet achieved, and with additional test effort and tuning time significant further improvements are realistically possible. The model, however, is already performing well enough to be used in low-security applications.

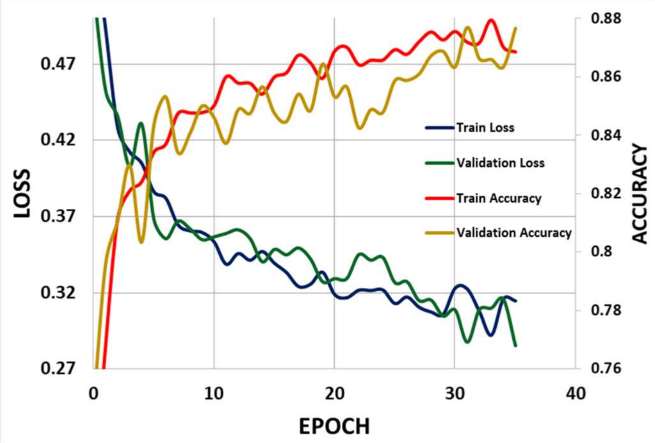

Figure 10: Training and validation of the Voice Recognition Network with modified voice recording time and sampling and also added background noise to the data

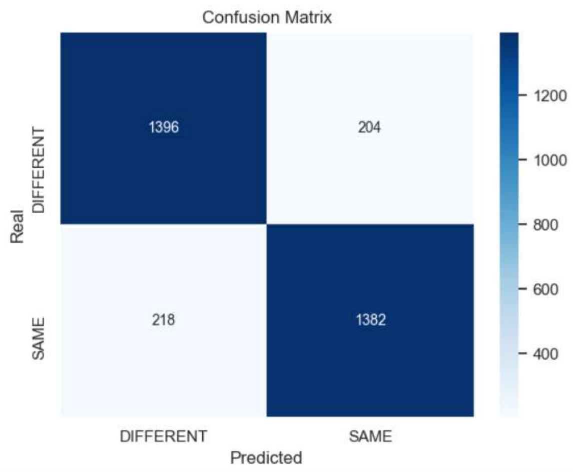

Figure 11: Confusion matrix for the improved voice recognition algorithm

Combined Face and Voice Recognition:

Figure 12 shows the combined face and voice recognition. The image frames are captured from the camera at every 100ms. The face in the image frame is then detected and cropped using Haar classifier. The cropped image is then compared with those in the database using face recognition algorithm and the image of the person with maximum similarity to the cropped face from the camera is selected as the recognized face. At the same time, the audio data from mic is recorded for three (3) seconds and after that the spectrogram of the recorded data is determined and compared to the spectrogram of the pre-recorded voices in the database using voice recognition algorithm. Similarly, to the voice recognition, the voice of the person with maximum similarity to the recorded audio from the mic is selected as the recognized voice.

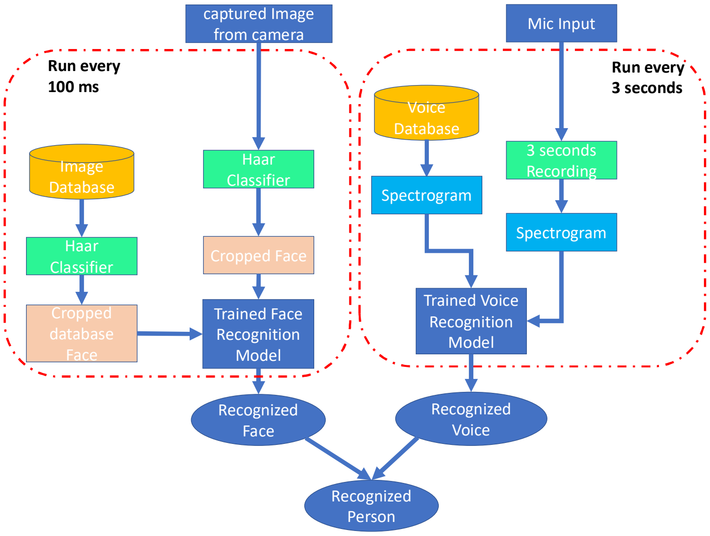

Figure 12: Person recognition using combined face and voice recognition algorithms

To test the combined voice and face recognition experimentally, the final algorithm was coded in Python and it was linked to the computer microphone and camera. The code was run in real and the algorithm performance was tested.

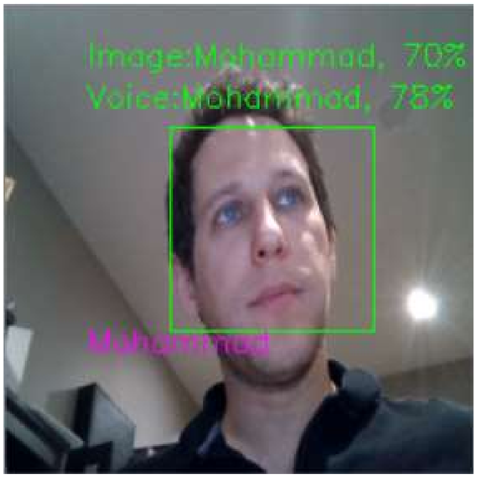

Figure 13: The snapshot of the experiment in the real time

Conclusions and Future Works:

In this study, a bimodal biometric identification model was developed. The model contains a face recognition and a voice recognition model. These two models were developed independently based of Siamese network. The models' architectures were improved through several iterations to increase validation and testing accuracy while keeping the training accuracy high. To improve face recognition algorithm, extra dataset from different distribution was added to original dataset and to prevent model overfitting a dropout with ratio of 20% was added to every convolutional layer. For better face recognition, a trained Haar cascade classifier was added upstream of the face recognition model for face detection and image cropping. To improve voice recognition model, the length of the recording and also the number of samples were increased to improve the spectrogram resolution. Also, to reduce voice recognition model sensitivity to noise, a background noise was added to some of the training data and the model was trained again. The two trained models were combined together for final bimodal biometric identification.

Due to time and computational limitation, the models could not be optimized in terms of their architecture and hyperparameters. So, models architectural improvement could be a potential continuation of this project. Also, both models could be combined into a single end to end model. This could be another interesting future work for this study.

References

1. **Taigman, Yaniv, et al.** DeepFace: Closing the Gap to Human-Level Performance in Face Verification.

2. **Schroff, Florian , Kalenichenko, Dmitry and Philbin, James .** FaceNet: A Unified Embedding for Face Recognition and Clustering.

3. **Salehghaffari, S H.** Speaker verification using convolutional neural network. s.l. : arXiv preprint arXiv:1803.05427, 2018 - arxiv.org.

4. *Seq2Seq Attentional Siamese Neural Networks for Text-dependent Speaker Verification,.* **Y. Zhang, M. Yu, N. Li, C. Yu, J. Cui and D. Yu.** Brighton, UK : ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

5. *Research and Application of Voiceprint Recogntion uing Recurrent Neural Network.* **K. Luo, L. Fu.** Chongqing : s.n. Automatic Control, Mechatronics and Industrial Engineering.

6. *Voice Recognition using Neural Networks.* **Ganesh K Venayagamoorthy, Viresh Moonasar, Kumbes Sandrasegaran.** 1998. IEEE.

7. *MSU-AVIS dataset: Fusing Face and Voice Modalities for Biometric Recognition in Indoor Surveillance Videos.* **A Chowdhury, Y Atoum, L Tran, X Liu, A Ross.** Beijing : s.n., 2018. 24th International Conference on Pattern Recognition (ICPR).

8. *Face Recognition Using the SR-CNN Model.* **Yu-Xin Yang, Chang Wen, Kai Xie, Fang-Qing Wen, Guan-Qun Sheng.** 2018, Sensors.

9. *Augmenting remote multimodal person verification by embedding voice characteristics into face images.* **Su Wang, Roland Hu, Huimin Yu, Xia Zheng and R. I. Damper.** San Jose : s.n., 2013. IEEE International Conference on Multimedia and Expo Workshops (ICMEW).

10. *Multimodal biometric authentication based on voice, face and iris.* **T. Barbu, A. Ciobanu and M. Luca.** Iasi : s.n., 2015. E-Health and Bioengineering Conference (EHB).

11. *Face and Speech Recognition Based Smart Home.* **A. Munir, S. Kashif Ehsan, S. M. Mohsin Raza and M. Mudassir.** Lahore : s.n., 2019 . International Conference on Engineering and Emerging Technologies (ICEET).

12. *An Efficient Android-Based Multimodal Biometric Authentication System With Face and Voice.* **X. Zhang, D. Cheng, P. Jia, Y. Dai and X. Xu.** 2020, IEEE Access, Vol. 8, pp. 102757-102772.

13. **Yang Song, Zhifei Zhang.** UTKFACE - Large Scale Face Dataset. [Online] https://susanqq.github.io/UTKFace/.

14. **Chowdhury, Aruni Roy.** FDDB: Face Detection Data Set and Benchmark. [Online] University of Masachusetts Amherst. http://vis-www.cs.umass.edu/fddb/.

15. **Pete Warden, Google Brain Team.** [Online] August 24, 2017. https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html.

16. **Keyword Spotting. [Online] https://paperswithcode.com/task/keyword-spotting.**

17. ***EFFICIENT KEYWORD SPOTTING USING DILATED CONVOLUTIONS AND GATING.*** **Alice Coucke, Mohammed Chlieh, Thibault Gisselbrecht, David Leroy,. 2018.**

18. ***Design and Evaluation of a Real-Time Face Recognition System Using Convolutional Neural Network.*** **Manikandan, KB Pranav and J. s.l. : Third International Conference on Computing and Network Communications, 2020.**

19. ***MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.*** **Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam. s.l. : arXiv:1704.04861, 2017.**

20. ***Robust Real-time Object Detection.*** **Paul Viola, Michael Jones. Vancouver : s.n., 2001.**