MSU-AVIS dataset: Fusing Face and Voice Modalities for Biometric Recognition in Indoor Surveillance Videos

Anurag Chowdhury*, Yousef Atoum+, Luan Tran*, Xiaoming Liu*, Arun Ross*
*Michigan State University, USA
+Yarmouk University, Jordan

***Abstract*—Indoor video surveillance systems often use the face modality to establish the identity of a person of interest. However, the face image may not offer sufficient discriminatory information in many scenarios due to substantial variations in pose, illumination, expression, resolution and distance between the subject and the camera. In such cases, the inclusion of an additional biometric modality can benefit the recognition process. In this regard, we consider the fusion of voice and face modalities for enhancing the recognition accuracy. The main contribution of this work is assembling a multimodal (face and voice), semi-constrained, indoor video surveillance dataset referred to as the MSU Audio-Video Indoor Surveillance (MSU-AVIS) dataset. We use a consumer-grade camera with a built-in microphone to acquire data for this purpose. We use current state-of-art deep-learning based methods to perform face and speaker recognition on the collected dataset for establishing baseline performance. We also explore multiple fusion schemes to combine face and speaker recognition to perform effective person recognition on audio-video surveillance data. Experiments convey the efficacy of the proposed multimodal fusion scheme (face and voice) over unimodal approaches in surveillance scenarios. The collected dataset is being made available for research purposes.**

I. Introduction

A number of law enforcement operations utilize video surveillance as a tool to either pro-actively mitigate potential threats (e.g., by detecting suspicious behavior) or reactively determine causal factors of an event (e.g., identifying suspects in a crime scene). A typical *automated* surveillance tool consists of a camera operating in the visible or near infrared spectrum, that acquires a video stream pertaining to a scene, which is subsequently processed by computer vision algorithms for scene analysis and understanding. In some cases, the video surveillance tool may incorporate a biometric module for determining the identity of individuals in a scene. The face modality is a viable modality in such biometric surveillance systems since it lends itself to performing recognition-at-a-distance.

In some indoor surveillance scenarios, the in-built microphone in consumer grade cameras can potentially capture the voice of a person of interest. The availability of the voice modality enables speaker recognition to be performed on the speech sample of subjects and this could be combined with face recognition to generate a final decision. Such a biometric system, based on a multi-modal approach, is expected to be more robust to variations in the input data.

In this work, we assemble a multi-modal face and voice dataset for research purposes. We capture both video (face modality) and audio (voice modality) data in a simulated indoor CCTV surveillance setup. The motivation of this work is to generate a dataset that represents the challenges of performing face and speaker recognition in this type of a scenario.

II. Related Work

Multi-modal biometric fusion has been shown to be beneficial in different tasks [11], [25]. In biometrics, fusion [21] of different biometric modalities, including face and voice, has been explored at multiple levels such as data-level, feature-level, score-level, rank-level, and decision-level. Data-level fusion entails fusing multiple instances of raw data obtained using different sensors or the same sensor [20]. Feature-level fusion refers to the consolidation of feature sets extracted from different modalities to produce a new feature set [19]. Score-level fusion combines the match scores generated by multiple biometric matchers to render the final decision [10]. Rank-level fusion combines the multiple ranks associated with each identity in a gallery database, where each rank is computed by a different biometric identification method [2]. Decision-level fusion pools the decision output by different biometric matchers and uses techniques such as majority voting to generate the final decision [29].

Among various fusion strategies, score-level fusion has become popular since it represents a good trade-off between information availability and information entropy. On one hand, most commercial biometric systems do not provide access to the raw data nor the feature sets extracted from the data. On the other hand, while final decisions and ranks are readily accessible in most commercial systems, their entropy is rather limited compared to scores.

With the advent of hand-held and portable devices capable of capturing biometric data, research has focused toward utilizing such devices for biometric applications. In [8], the authors investigate the efficacy of existing face and speaker recognition algorithms when deployed on hand-held devices, equipped with lower quality audio/video capture hardware. The scores output by the speaker and face systems are fused using linear weighted summation, where the weights are learned using the minimum classification error principle on a training set.

Table I
SUMMARIZING THE CHARACTERISTICS OF EXISTING MULTI-MODAL FACE AND VOICE DATASETS.

Dataset	Subjects	Sessions		Samples/Session		Data specs		Covariates
		Face	Voice	Face	Voice	Frame/Video	Audio	
M2VTS [17]	37	5	5	1	1	$286 \times 350 \times 1$	16bit, 48kHz	Face pose variation, clean audio, text dependent
XM2VTS [15]	295	4	1	2	4	$576 \times 720 \times 3$	16bit, 32kHz	Face pose variation, clean audio, text dependent
BANCA [5]	52	12	12	2	2	-	16bit, 32kHz	Frontal face, clean & degraded audio, text independent
VidTIMIT [23]	43	3	3	1	3 (approx.)	$512 \times 384 \times 1$	16bit, 32kHz	Face pose variation, clean audio, text dependent
MOBIO [13]	160	6	6	5	21	$64 \times 80 \times 1$	48kHz	Frontal face, clean audio, text independent
MSU-AVIS (proposed)	50	3	3	12	12	$1920 \times 1080 \times 3$	48kHz	Face pose-expression-distance variation, indoor, clean & degraded audio, text independent

Further extending this idea of user authentication on hand-held devices, the authors in [14] implement a bi-modal biometric authentication scheme on a smartphone. Speaker recognition is performed using the i-Vector Probabilistic Linear Discriminant Analysis (PLDA) method, while face recognition is accomplished using a histogram of Local Binary Pattern (LBP) features. The scores from both methods were then multiplied to obtain the final fused score.

Technological improvements [1] in audio-visual acquisition devices as well as processing algorithms have made video-based biometric applications feasible, where a *synchronized* audio-video stream can be used to identify a person more effectively than using a static face image and an audio sample taken separately. Further, such systems offer resilience to spoof attacks and often have better data quality leading to improved performance. Poh et al. [18] proposed a discriminative video-based score-level fusion algorithm to leverage the temporal relevance in the classification scores that vary with time across the video frames, thereby improving the performance over traditional rule-based fusion strategies in a video-based setting.

Authors in [3] use the speech and face modalities together to measure audio-visual synchrony for talking-face identity verification. While the work in their paper is aimed at improving the robustness of audio-visual biometric systems against impostor attacks, the authors also claim improved biometric verification performance by fusing the speech and face recognition systems using an SVM.

Degradations in biometric signals are known to be a major contributing factor to poor performance in biometric systems. For example, background noise can lead to poor-quality audio signals and adverse illumination can impede the usability of facial imagery. However, fusion based approaches, such as in [12], [16], [24], have leveraged the quality of the raw biometric data to control the influence of each individual biometric modality in a multi-modal fusion framework.

In [28], the authors show the effectiveness of fusing face and voice over unimodal biometrics in adverse outdoor situations, where the quality of the biometric signal varies vastly depending upon the data acquisition environment. Multi-modal face and voice biometrics is also being used in consumer applications [1] to boost security against spoof attacks, especially in scenarios where single-factor biometric solutions are more vulnerable to such attacks.

III. MSU-AVIS DATASET

The data collection conducted in this work involved acquiring the audio-visual data of subjects freely walking in an indoor environment. The resultant dataset is referred to as MSU-AVIS (MSU Audio-Visual Indoor Surveillance). The purpose of the data collection was to mimic a real-world surveillance scenario in a semi-constrained indoor environment. Such a scenario is often encountered in public buildings and grocery stores. In these scenarios, face images exhibit variations due to occlusions, pose, indoor-illumination, expressions, accessories, etc. Similarly, audio samples exhibit variations due to distance of subject from the microphone, indoor reverberations, background noise, etc. However, since this dataset is designed to mimic an indoor surveillance scenario, it excludes certain challenges that are prevalent in an outdoor surveillance scenario such as unconstrained natural illumination. Thus, we use the term *semi-constrained* to refer to our data acquisition setup.

A. Data collection scenario

An office environment is used for the data collection as seen in Fig. 1. The scenario involves two subjects. The first subject referred to as the "target", plays the role of the speaker, who talks throughout the duration of the video. The second subject plays the role of a "listener", who only listens to what the target has to say. The main role of the listener is to control the location where the speech between the two subjects occurs. For example, if the listener is seated far away from the camera, it will force the target to also be away from the camera. Our goal is to determine the target's identity by utilizing audio-visual cues obtained via the camera and microphone.

B. Data collection setup

We used a web-cam (Logitech C920 HD Pro) to collect *probe* data, by fixing it at a height of 240cm from the ground. It has built-in dual stereo microphones that are used to record the audio in the room. When acquiring some of the videos in the dataset, we replaced the built-in microphones with a lower quality microphone that picked up surrounding noise at a higher rate. This introduced some challenges in detecting speech corresponding to the target. Further, the target was asked to speak freely (text-independent) when acquiring the probe surveillance videos. For assembling the *gallery* face images and audio, we used the same web-cam, with the height of the camera set to match the height of the target subject. When collecting the gallery audio, the speech was scripted

Figure 1. Sample frames from the MSU-AVIS dataset.

(text-dependent), where the target was required to select a short script at random from a list of five scripts.

C. Data collection statistics and challenges

We collected data from 50 subjects (among which 16 are female). Some of the major challenges observed in the MSU-AVIS dataset compared to existing multi-modal face and voice datasets are summarized in Table I and described below.

- Some subjects spoke with a soft voice leading to extensive voice activity detection challenges.
- Some subjects spoke for a short period of time, while others spoke throughout the duration of the video, thereby creating audio data imbalance across subjects.
- Nearly 30% of the videos were collected when using a microphone that recorded some of the background noise thereby making speaker recognition more challenging.
- Due to the semi-constrained nature of the data collection, the size of faces varied extensively depending upon the stand-off distance of the subject from the camera. There were also large variations in subject pose with respect to the camera.
- Some subjects spent long periods of time with the back of their heads toward the camera, thus rendering those frames unusable for face recognition and relying on speaker recognition alone.

Our goal is to overcome many of the aforementioned challenges by fusing both audio and face visual cues.

D. MSU-AVIS auxiliary dataset

Based on our preliminary results (described later), face and speaker recognition performance were observed to suffer extensively under certain scenarios. For face recognition , facial image resolution and large pose variation were among the most challenging factors impacting matching performance. On the other hand, the performance of speaker recognition was observed to degrade when the target subject was far away from the microphone.

Therefore, we assembled an auxiliary set, based on a subset of 10 subjects from the MSU-AVIS dataset, who were instructed to mimic the challenges indicated above. Each subject provided 36 video clips, with each clip being five seconds long. The auxiliary dataset helps us to specifically evaluate (results given later) the benefits of using multi-modal fusion in scenarios where unimodal approaches fail to perform well.

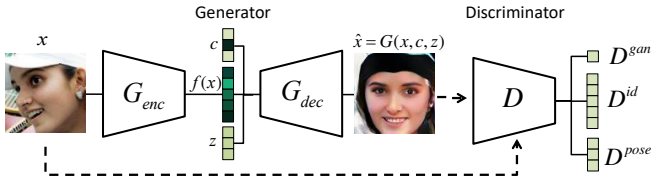


Figure 2. An illustration of the DR-GAN face recognition network.

IV. Algorithms for Face and Speaker Recognition

For recognizing the individuals in a video from the probe dataset, we used both the face and voice modalities. Detection and recognition of both face and voice modalities were done separately and then fused at the score level.

A. Face Recognition

1) Face recognition method: For recognizing the subjects in the videos based on their face, we chose the DR-GAN approach [26], [27]. Among many pose-robust face recognition algorithms [6], [31], DR-GAN is one of the latest recognition networks that can simultaneously learn pose-invariant representation and synthesize faces with arbitrary poses. As visualized in Fig. 2, the framework consists of a generator and a discriminator. The main network, generator G, consists of an encoder and a decoder. The encoder takes a face image as input and extracts a feature representation $f(x) \in \mathbb{R}^{320}$. The feature is then input to the decoder in order to synthesize face images of the same person in different poses, where the pose is specified by a pose code $\mathbf{c}$. The ability to generate images in different poses allows the pose feature to be disentangled from the identity representation, which improves face recognition performance.

In addition to adversarial loss, i.e., D^{gan} [7] that distinguishes between real and synthetic images, in DR-GAN, the discriminator D has two more tasks: Id classification D^{id} and

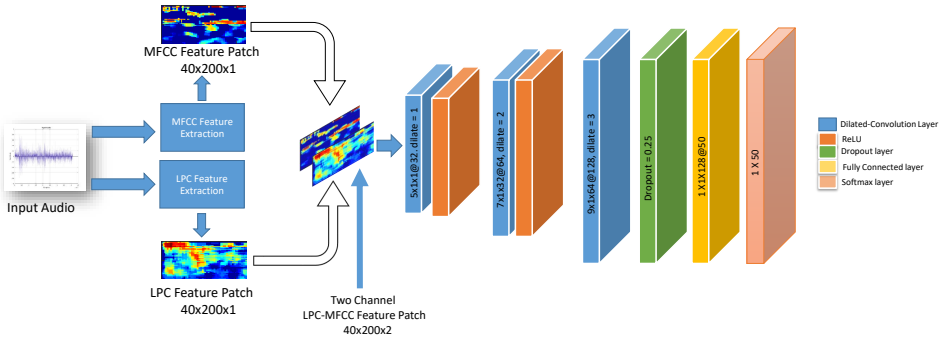


Figure 3. An illustration of the 1-D CNN based speaker recognition algorithm used in this work.

pose estimation D^{pose}. This helps to guide the generator G in synthesizing images that are not only realistic but also preserve the subject identity while manipulating the pose.

2) Computing the face recognition score on MSU-AVIS: We used the DR-GAN network trained on $\sim 500k$ images of celebrities from the CASIA-Webface dataset [30]. To be able to perform face recognition on the video streams, we initially conducted face detection on every frame. We used the Tiny Face Detector [9], which has been demonstrated to detect faces over a wide range of scales and poses. Each detection is accompanied by a confidence score. A threshold of 0.5 was utilized to eliminate false detections. Based on prior knowledge of the dataset, there are three possible face detection outcomes on every frame: a) Zero faces detected. None of the faces are visible in the frame, as can be seen in the second-row third-column of Fig. 1. b) One face detected. The detected face could belong to either the target or the listener, as can be seen in the first-row of Fig. 1. c) Two faces are detected. One belongs to the target walking in the room, and the other corresponds to the listener, as can be seen in the second-row first-column of Fig. 1.

For recognition purposes, we retained all faces of the target in every clip. In the case of (b) and (c), where one or two faces are detected, we need to exclude face images belonging to the listener and only consider those of the target. This was done as follows. We used the gallery images of the target to compute a feature vector representing the target's face using the DR-GAN approach [26]. By computing the similarity scores between this feature vector and those of the detected faces in the frame, we can easily determine which of the faces belongs to the target. Finally, we utilize all n face images $\{xi\}_{i=1}^{n}$ to generate a fused representation, which is the weighted average of individual representations, and the weight corresponds to the coefficient w_i estimated by the multi-image DR-GAN algorithm. The coefficient is related to image quality (see [27]):

$$f(x_1, x_2, ..., x_n) = \frac{\sum_{i=1}^{n} w_i f(x_i)}{\sum_{i=1}^{n} w_i}. \quad (1)$$

Now we can identify the target by computing similarity scores between this target and the gallery images of the 50 subjects in MSU-AVIS.

B. Speaker Recognition

1) Speech segmentation: To perform speaker recognition from videos, we first extract the audio clips from the video files. The extracted audio is then processed using *Voice Activity Detection* for segmenting out the speech frames from non-speech frames in the audio clip.

2) Speaker recognition method: For recognizing the subjects in the probe videos based on their voice, we use a 1-D CNN based speaker recognition algorithm [4]. The segmented audio from the dataset is processed to extract Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) features separately. MFCC features are widely used in audio-based biometric systems [22]. The MFCC and LPC features are then stacked together forming a two channel LPC-MFCC feature matrix. For extracting the MFCC and LPC features, a window size of $20ms$ was chosen with a stride of $10ms$. For the MFCC features, 20 cepstral coefficients (including the 0^{th} order cepstral coefficient) along with their delta features were extracted, resulting in a 40 dimensional MFCC feature vector per frame. For the LPC features, we extract 40 coefficients. The decision to extract 40 LPC coefficients (instead of the traditional 12 to 20 coefficients) was made in order to facilitate compatibility with that of the MFCC features. This was necessary for inputting the MFCC and LPC features as a two-channel matrix into the 1D-CNN. The two channel feature matrix from the input audio was then split along frames, as discussed in [4], and then into fixed-length patches. The two channel feature patches were input to the 1D-CNN as shown in Fig. 3.

3) Computing the speaker recognition score on MSU-AVIS: For training the 1D-CNN, we used the MSU-AVIS's gallery videos. MSU-AVIS dataset has one gallery video per target subject. Each video was split into multiple clips of length 5 seconds each. The 1D-CNN was trained using all the clips from the gallery set of 50 subjects. As stated in [4], to compute the match score between a probe audio X and each of the gallery speakers, the probe audio is first split into multiple overlapping patches, $x_i, i \in \{1, 2, ..., N\}$, each of length 5 seconds, where N is the number of patches. For each x_i, the trained CNN outputs match scores, $\{s_{i,j}\}, j = \{1, 2, ..., C\}$, pertaining to the C speakers in the gallery set. Here, $s_{i,j}$ is the match score assigned to the j^{th} gallery subject for the i^{th} probe patch.

Table II
FUSION RULES USED IN OUR EXPERIMENTS.

Sum Rule	$F_{sum} = S_1 + S_2$
Product Rule	$F_{prod} = S_1 \times S_2$
Fusion Rule-1	$F_1 = S_1 \times S_2 \times e^{-\left(\frac{S_1 - S_2}{S_1 + S_2}\right)^2}$
Fusion Rule-2	$F_2 = W_1 \times S_1 + W_2 \times S_2$
Fusion Rule-3	$F_3 = W_1 \times S_1 \times W_2 \times S_2$
Fusion Rule-4	$F_4 = (W_1 \times S_1) \times (W_2 \times S_2) \times e^{-\left(\frac{(W_1 \times S_1) - (W_2 \times S_2)}{(W_1 \times S_1) + (W_2 \times S_2)}\right)^2}$

The mean of the scores corresponding to all the patches extracted from the audio clip is then computed as follows:

$$S_j = \frac{1}{N} \sum_{i=1}^{N} s_{i,j}, \forall j.$$

Here, S_j is the score between the input probe audio sample and the j^{th} gallery subject.

V. MULTI-MODAL FUSION EXPERIMENTS

For combining the scores from the face recognition (S_1) and speaker recognition (S_2) algorithms, we first normalize the individual scores in the range $[0, 1]$. Further, quality values for both face images (W_1) and audio data (W_2) were extracted based on the degree of usable face data and speech data available in each clip. Here, we define a video frame to be usable when the face detector detects a face image and the voice activity detector detects some speech. The face quality, W_1, is computed as the average of the coefficients w produced by DR-GAN on all detected faces in the video clip. The speech quality, W_2, is computed as the proportion of the audio clip where speech was detected by the voice activity detector.

For establishing baseline performance for score-based fusion, we chose the sum rule and product rule. We further explore four more score-based fusion rules - Fusion Rule-1 to Fusion Rule-4 - as shown in Table II.

1) Fusion Rule-1: In this rule, we introduce an exponential weighting factor in the product rule. This weighting factor is inversely proportional to the squared difference between face and voice scores. The rationale behind this weighting mechanism is to assign a lower weight to the fused scores in cases where the mutual confidence of the modalities is low.
2) Fusion Rule-2: In this rule, we use the quality values for face images (W_1) and audio data (W_2) along with their corresponding match scores (S_1) and (S_2), in a weighted sum rule scheme.
3) Fusion Rule-3: In this rule, we use the quality values for face images (W_1) and audio data (W_2) along with their corresponding match scores (S_1) and (S_2), in a weighted product rule scheme.
4) Fusion Rule-4: In this rule, we introduce the quality values for face images (W_1) and audio data (W_2) along with their corresponding match scores (S_1) and (S_2), in Fusion Rule-1.

VI. RESULTS AND ANALYSIS

Both identification and verification experiments were conducted. The results of the identification and verification experiments are given in Tables III and IV. Corresponding CMC and ROC curves are given in Fig. 4 and 5. Rank-1 identification accuracies (in %) are reported for the identification experiments and True Match Rates (TMR) at a False Match Rate (FMR) of 0.1 are reported for the verification experiments. The experimental results corresponding to the entire MSU-AVIS dataset using different methods are recorded in Tables III and IV under the 'Overall' column. Further, experiments were performed on several subsets, based on different data and subject characteristics, and the corresponding results are given in Tables III and IV with the columns labeled accordingly.

Session-based experiments: Each subject, during the probe data collection, participated in three sessions. The main difference across the sessions is the distance between the subject and the camera. The distance in sessions 1 through 3 progressively reduced. The results are given in Table III and IV. The general system performance degrades with large standoff distances, due to the lower resolution of the captured face images and the degraded quality of the recorded speech audio. It was observed that the fusion results has the highest impact in session 1, where both modalities have lower identification and verification results.

Audio quality-based experiments: A total of 17 videos in the MSU-AVIS dataset used a low quality microphone that introduced high noise artifacts, making speaker recognition very challenging, as can be seen from the results presented in Tables III and IV. Due to the poor performance of speaker recognition in these cases, the fusion methods had low impact compared to the face-only performance.

Gender-based experiments: The total number of female subjects was 16 out of a total of 50 subjects. The results of gender-based performance can be seen in Tables III and IV. It was observed that speaker identification for females was higher than that of males. Fusion results also reflect the same pattern, resulting in higher performance for females.

Race-based experiments: The Asian race is maximally represented in the MSU-AVIS dataset, with a percentage of 28%. The results of Asian subjects versus non-Asian subjects can be found in Tables III and IV. It was observed that face recognition has lower performance on Asian subjects compared to non-Asians. This is not surprising, since DR-GAN was trained on celebrity faces in the CASIA-Webface dataset, which does not have many Asian subjects. However, the fusion methods improve the results on Asian targets compared to the face-only performance.

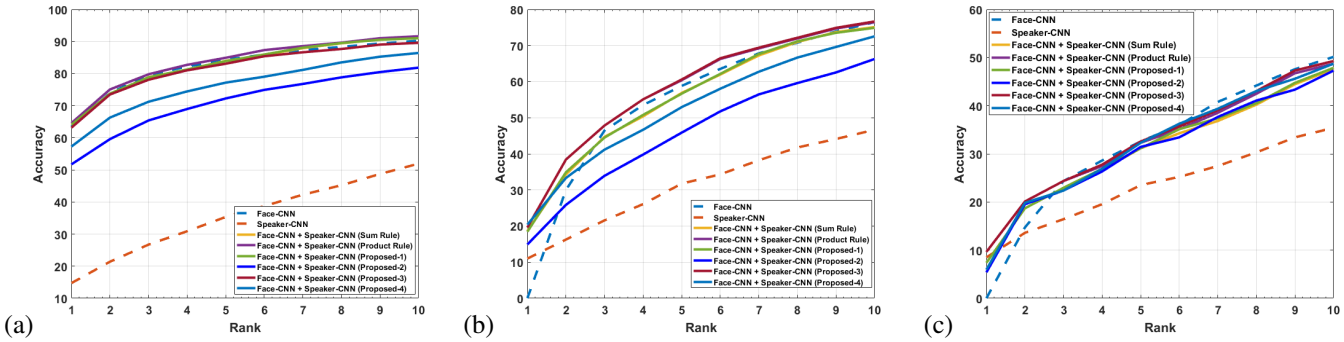
Experiment on cases where face recognition fails: In this experiment, we considered a subset of the MSU-AVIS dataset where face recognition failed, as seen in Fig. 6. This amounted to almost 30% of the entire MSU-AVIS dataset. We analyzed the face images in this subset and found that most of the face images are either of poor resolution or are side-profile faces. We performed fusion experiments on this subset and found

Table III
IDENTIFICATION RESULTS (RANK 1, IN %) ON THE MSU-AVIS DATASET ACROSS DIFFERENT BASELINE METHODS.

Methods	Video			Audio Quality		Gender		Race		Overall
	1	2	3	Good	Bad	Male	Female	Asian	Non_Asian	
Face -CNN	36.50	**65.67**	**88.33**	63.06	**76.07**	**64.30**	74.13	63.97	**74.80**	63.50
Speaker -CNN	10.50	14.33	19.33	20.57	14.96	15.52	23.09	18.06	21.43	14.72
Fusion -Sum Rule	37.00	64.10	85.80	66.24	67.64	61.19	**77.56**	66.14	69.35	64.04
Fusion -Product Rule	**38.12**	65.52	84.45	**66.50**	66.18	62.38	75.11	**67.26**	68.05	**64.59**
Fusion -1	37.00	64.10	85.99	66.24	67.64	61.49	**77.56**	66.14	69.35	64.04
Fusion -2	32.74	46.45	70.44	54.85	56.36	49.11	61.33	57.21	56.88	51.71
Fusion -3	38.12	64.30	81.57	64.73	66.18	60.99	73.11	65.77	66.49	63.15
Fusion -4	34.98	54.36	73.70	59.83	59.64	53.86	70.00	61.30	60.00	57.26

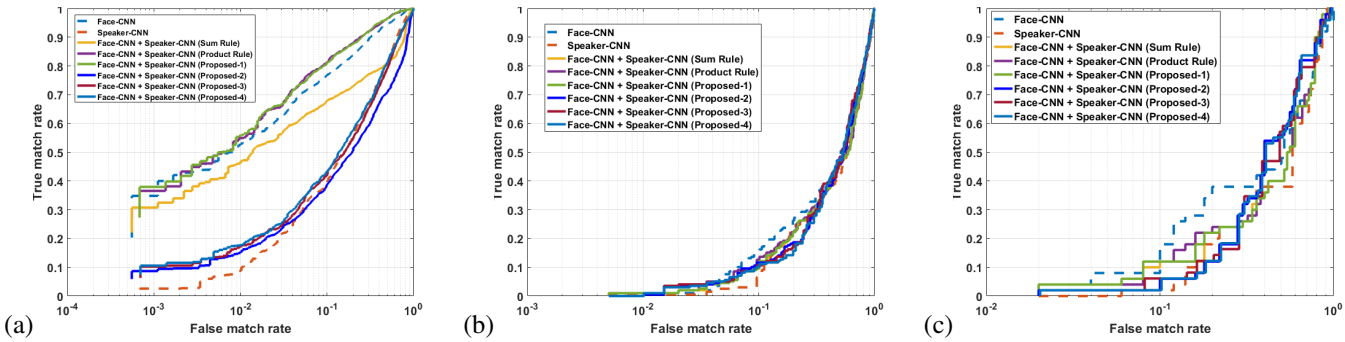
Table IV
VERIFICATION RESULTS (TMR@FMR = 0.1) ON THE MSU-AVIS DATASET ACROSS DIFFERENT BASELINE METHODS.

Methods	Video			Audio Quality		Gender		Race		Overall
	1	2	3	Good	Bad	Male	Female	Asian	Non_Asian	
Face -CNN	0.57	0.83	0.94	0.76	0.10	0.15	0.08	**0.14**	0.40	0.77
Speaker -CNN	0.36	0.39	0.44	0.44	0.11	0.11	0.10	0.09	0.23	0.40
Fusion -Sum Rule	0.52	0.71	0.74	0.83	0.12	0.17	0.08	0.09	0.36	0.68
Fusion -Product Rule	0.66	**0.85**	**0.95**	**0.81**	**0.13**	**0.18**	0.10	0.08	**0.42**	**0.81**
Fusion -1	**0.67**	0.84	**0.95**	0.80	0.12	**0.18**	0.09	0.11	0.40	**0.81**
Fusion -2	0.28	0.43	0.50	0.42	0.08	0.16	0.12	0.11	0.23	0.39
Fusion -3	0.27	0.50	0.53	0.42	0.07	0.16	**0.15**	0.05	0.25	0.42
Fusion -4	0.32	0.51	0.55	0.45	0.07	0.17	**0.15**	0.05	0.25	0.43



(a) (b) (c)

Figure 4. CMC curves on (a) the entire MSU-AVIS dataset, (b) the subset where face recognition fails, and (c) the MSU-AVIS-auxiliary dataset



(a) (b) (c)

Figure 5. ROC curves on (a) the entire MSU-AVIS dataset, (b) the subset where face recognition fails, and (c) the MSU-AVIS-auxiliary dataset

that fusion with speaker recognition gives a substantial boost to the overall recognition performance. The results are given in Table V and Figs. 4 and 5.

Experiment on the auxiliary dataset: We performed fusion experiments on the auxiliary dataset to confirm our earlier findings on the subset of the MSU-AVIS dataset where face recognition failed. We found that speaker recognition positively aids in improving the recognition performance in cases where face images are of degraded quality and exhibit large pose variations. The results are given in Table V and Figs. 4 and 5.

However, it is interesting to note that the performance gain

Table V
IDENTIFICATION (RANK 1) AND VERIFICATION (TMR@FMR=0.1) RESULTS ON A SUBSET OF THE MSU-AVIS DATASET WHERE THE FACE MODALITY FAILS AND ON THE MSU-AVIS-AUXILIARY DATASET.

Methods	Face Failure Subset		Auxiliary Dataset	
	Ident.	Verif.	Ident.	Verif.
Face-CNN	0	**0.15**	0	0.08
Speaker-CNN	10.98	0.06	8.49	0.02
Fusion-Sum Rule	18.62	0.10	7.36	0.10
Fusion-Product Rule	**19.60**	0.12	**9.63**	**0.12**
Fusion-1	18.43	0.09	7.36	**0.12**
Fusion-2	14.90	0.11	5.38	0.02
Fusion-3	**19.60**	0.10	**9.63**	0.06
Fusion-4	**19.60**	0.10	**9.63**	0.02

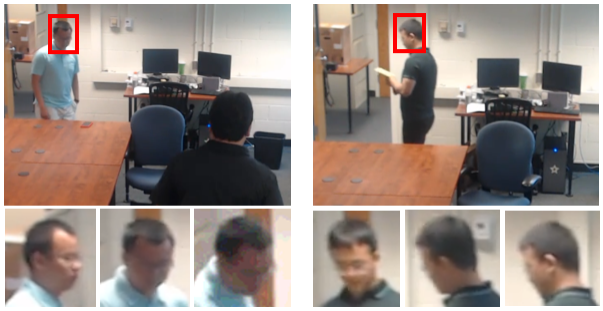


Figure 6. Examples of video clips where face recognition fails. The larger images are sub-regions from a frame obtained from the clip. The smaller images are some of the other faces obtained from the same clip.

offered by fusion in the auxiliary dataset is not as large as was seen in the cases where face recognition failed on the original MSU-AVIS dataset. This could be attributed to the fact that, unlike in the MSU-AVIS dataset, face images in the auxiliary dataset are strictly non-frontal and captured at a large stand-off distance. This makes face recognition even more challenging and, hence, the impact of the face recognition algorithm on fusion is further reduced.

VII. SUMMARY

In summary, the following are the primary contributions of this work:

(i) A multi-modal indoor-surveillance dataset comprising of face and voice modalities was collected.

(ii) The performance of CNN-based face recognition and speaker recognition was evaluated on this dataset.

(iii) The benefit of fusing the voice and face modalities was demonstrated in scenarios where both the face and voice data suffer from extensive degradations.

VIII. DATASET AVAILABILITY

The dataset is available for research purposes at http://cvlab.cse.msu.edu/msu-avis-dataset.html

REFERENCES

[1] Blending biometric facial and voice recognition systems advances security. https://www.electronicproducts.com/Biotech/Research/.

[2] A. Abaza and A. Ross. Quality based rank-level fusion in multibiometric systems. In *IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6, September 2009.

[3] H. Bredin and G. Chollet. Audio-visual speech synchrony measure for talking-face identity verification. In *ICASSP*, 2007.

[4] A. Chowdhury and A. Ross. Extracting sub-glottal and supra-glottal features from MFCC using convolutional neural networks for speaker identification in degraded audio signals. In *IJCB*, 2017.

[5] J. Czyz, S. Bengio, C. Marcel, and L. Vandendorpe. Scalability analysis of audio-visual person identity verification. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, 2003.

[6] C. Ding and D. Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on intelligent systems and technology (TIST)*, 7(3):37, 2016.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[8] T. Hazen, E. Weinstein, R. Kabir, A. Park, and B. Heisele. Multi-modal face and speaker identification on a handheld device. In *Proceedings of the Workshop on Multimodal User Authentication*, 2003.

[9] P. Hu and D. Ramanan. Finding tiny faces. In *CVPR*, 2017.

[10] A. Jain, K. Nandakumar, and A. Ross. Score Normalisation in Multimodal Biometric Systems. *Pattern Recognition*, 38(12), 2005.

[11] A. K. Jain and A. Ross. Multibiometric systems. *Communications of the ACM, Special Issue on Multimodal Interfaces*, 47(1), January 2004.

[12] J. Kittler, N. Poh, O. Fatukasi, K. Messer, K. Kryszczuk, J. Richiardi, and A. Drygajlo. Quality dependent fusion of intramodal and multimodal biometric experts. In *Proc. of SPIE*, 2007.

[13] S. Marcel et al. On the results of the first mobile biometry (MOBIO) face and speaker verification evaluation. In *International Conference on Recognizing Patterns in Signals, Speech, Images, and Videos*, 2010.

[14] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernockỳ, N. Poh, J. Kittler, A. Larcher, C. Levy, et al. Bi-modal person recognition on a mobile phone: using mobile phone data. In *ICME Workshops*, 2012.

[15] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database.

[16] N. Ozay, Y. Tong, F. W. Wheeler, and X. Liu. Improving face recognition with a quality-based probabilistic framework. In *CVPRW*, 2009.

[17] S. Pigeon and L. Vandendorpe. The M2VTS multimodal face database (release 1.00). In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 403–409. Springer, 1997.

[18] N. Poh, J. Kittler, and F. Alkoot. A discriminative parametric approach to video-based score-level fusion for biometric authentication. In *ICPR*, 2012.

[19] A. Ross and R. Govindarajan. Feature Level Fusion Using Hand and Face Biometrics. In *Proc. SPIE Conf. on Biometric Technology for Human Identification II*, volume 5779, pages 196–204, Orlando, 2005.

[20] A. Ross, S. Shah, and J. Shah. Image Versus Feature Mosaicing: A Case Study in Fingerprints. In *Proceedings of SPIE Conference on Biometric Technology for Human Identification*, volume 6202, pages 1–12, Orlando, USA, April 2006.

[21] A. A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of multibiometrics*, volume 6. Springer Science & Business Media, 2006.

[22] J. Roth, X. Liu, A. Ross, and D. Metaxas. Investigating the discriminative power of keystroke sound. *IEEE Transactions on Information Forensics and Security*, 10(2):333–345, February 2015.

[23] C. Sanderson and K. Paliwal. Fast features for face authentication under illumination direction changes. *Pattern Recognition Letters*, 24(14):2409–2419, 2003.

[24] Y. Tong, F. W. Wheeler, and X. Liu. Improving biometric identification through quality-based face and fingerprint biometric fusion. In *CVPRW*, 2010.

[25] L. Tran, X. Liu, J. Zhou, and R. Jin. Missing modalities imputation via cascaded residual autoencoder. In *CVPR*, 2017.

[26] L. Tran, X. Yin, and X. Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *CVPR*, 2017.

[27] L. Tran, X. Yin, and X. Liu. Representation learning by rotating your faces. *arXiv preprint arXiv:1705.11136*, 2017.

[28] H. Vajaria, T. Islam, P. Mohanty, S. Sarkar, R. Sankar, and R. Kasturi. Evaluation and analysis of a face and voice outdoor multi-biometric system. *Pattern recognition letters*, 28(12):1572–1580, 2007.

[29] K. Veeramachaneni, L. Osadciw, A. Ross, and N. Srinivas. Decision-level fusion strategies for correlated biometric classifiers. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, June 2008.

[30] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[31] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017.