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

Anurag Chowdhury[1], Yousef Atoum[2], Luan Tran[1], Xiaoming Liu[1], Arun Ross[1]

1-Michigan State University, USA; 2-Yarmouk University, Jordan

Follow @iPRoBeLab

Introduction

- Indoor video surveillance systems primarily use the face modality for recognizing people.
- However, face recognition can suffer due to substantial variations in pose, illumination, expression
- Therefore, inclusion of an additional biometric modality, such as voice, can benefit the recognition process.
- In this work, we introduce 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 current state-of-art deep learning based face and speaker recognition algorithms on the collected dataset and explore score based fusion rules for establishing baseline performance.

Dataset Challenges

Figure 1: Sample frames from the MSU-AVIS dataset. Scan the QR code to play a sample video

We collected data from 50 subjects. Some of the major challenges observed in the MSU-AVIS dataset are described below.

- Some subjects spoke with a soft voice leading to voice activity detection challenges.
- Some subjects spoke for a short period of time, while others spoke throughout the duration of the video, thereby creating imbalanced audio data across subjects.
- Nearly 30% of the videos were collected using a poor quality microphone, thereby adding audio degradations to collected speech data.
- Large variations in facial pose and size were brought about by varying relative positioning of subjects with respect to camera.

Auxiliary Dataset

- Face recognition in the MSU-AVIS dataset was observed to suffer most due to image resolution and facial pose variation
- Voice recognition was negatively impacted by large distance between subject and microphone
- An auxilliary dataset, based on a subset of 10 subjects from the MSU-AVIS dataset, was collected to mimic the above challenges
- The auxiliary dataset helped to specifically evaluate the benefits of using multi-modal fusion in scenarios where unimodal approaches fail to perform well

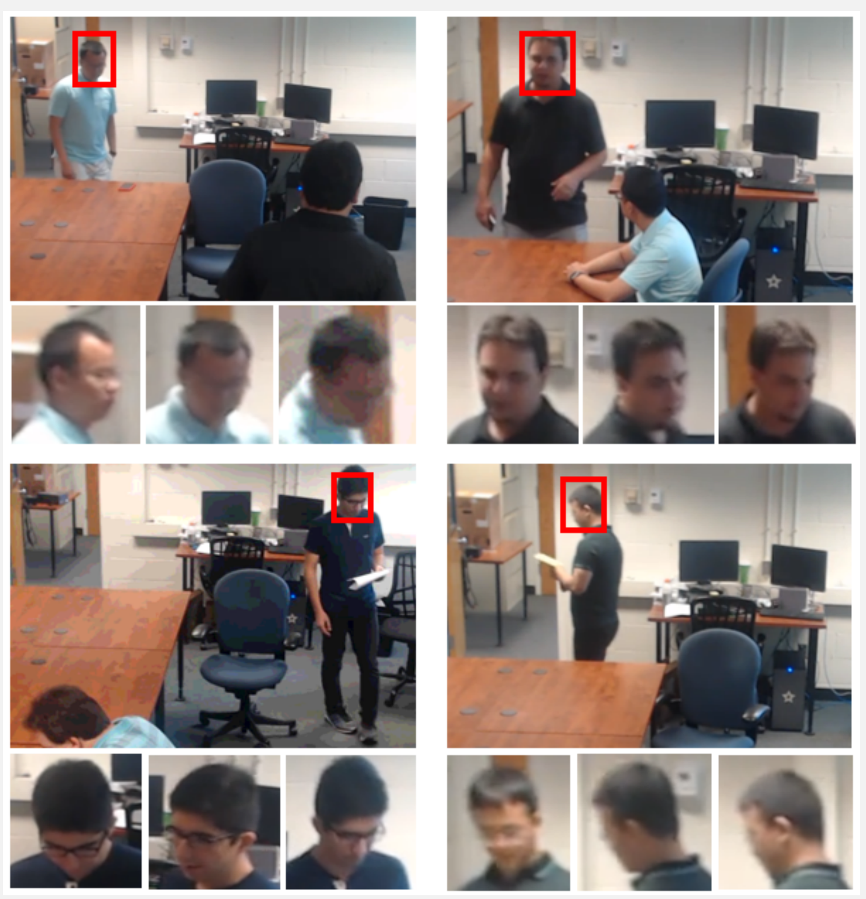

Figure 2: Face recognition failure cases in videos

Comparative Audio-Video Dataset Characteristics

Dataset	Subjects	Sessions		Samples/ Session		Data specs		Covariates
		Face	Voice	Face	Voice	Frame/Video	Audio	
XM2VTS [1]	295	4	1	2	4	576 × 720 × 3	16bit, 32kHz	Face pose variation, clean audio, text dependent
MOBIO [2]	160	6	6	5	21	64 × 80 × 1	48kHz	Frontal face, clean audio, text independent
MSU-AVIS (Proposed)	50	3	3	12	12	1920 × 1080 × 1	48kHz	Face pose-expression-distance variation, indoor, clean & degraded audio, text independent

Benchmark Results and Analysis

Methods	Description	Face Failure Subset		Auxiliary Dataset	
		Ident.	Verif.	Ident.	Verif.
Face-CNN [3]	$F_{face} = S_1$	0	**0.15**	0	0.08
Speaker-CNN [4]	$F_{spkr} = S_2$	10.98	0.06	8.49	0.02
Sum Rule	$F_{sum} = S_1 + S_2$	18.62	0.10	7.36	0.10
Product Rule	$F_{prod} = S_1 \times S_2$	**19.60**	0.12	**9.63**	**0.12**
Fusion Rule-1	$F_1 = S_1 \times S_2 \times e^{-\left(\frac{S_1 - S_2}{S_1 + S_2}\right)^2}$	18.43	0.09	7.36	**0.12**
Fusion Rule-2	$F_2 = W_1 \times S_1 + W_2 \times S_2$	14.90	*0.11*	*5.38*	*0.02*
Fusion Rule-3	$F_3 = W_1 \times S_1 \times W_2 \times S_2$	**19.60**	0.10	***9.63***	0.06
Fusion Rule-4	$F_4 = (W_1 \times S_1) \times (W_2 \times S_2) \times e^{-\left(\frac{(W_1 \times S_1) - (W_2 \times S_2)}{(W_1 \times S_1) + (W_2 \times S_2)}\right)^2}$	**19.60**	0.10	**9.63**	0.02

Figure 3: 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

Summary

- A multi-modal indoor-surveillance dataset comprising of face and voice modalities was collected.
- Face recognition experiments were performed using DR-GAN [3,4] algorithm and speaker recognition was performed using 1D-CNN [5] algorithm.
- Six different score based fusion rules were explored for establishing baseline performance on the MSU-AVIS Dataset.
- The benefit of fusing the voice and face modalities was demonstrated in scenarios where both the face and voice data suffer from extensive degradations.

Future Work

We plan to extend our work by developing methods for performing feature level fusion of face and voice modalities in the proposed dataset.

References

[1] Messer, Kieron, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. XM2VTSDB: The extended M2VTS database. In AVBPA, 1999.
[2] S. Marcel et al. On the results of the first mobile biometry (MOBIO) face and speaker verification evaluation. In ICPR, 2010.
[3] L. Tran, X. Yin, and X. Liu. Disentangled representation learning GAN for pose-invariant face recognition. In CVPR, 2017.
[4] L. Tran, X. Yin, and X. Liu. Representation learning by rotating your faces. arXiv preprint arXiv:1705.11136, 2017.
[5] 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.