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

- 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 auxiliary 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.

### Comparative Audio-Video Dataset Characteristics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Subjects</th>
<th>Sessions</th>
<th>Samples/Session</th>
<th>Data specs</th>
<th>Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>XM2VTS [1]</td>
<td>295</td>
<td>4</td>
<td>1</td>
<td>576 x 720 x 3</td>
<td>16bit, 32kHz, face pose variation, clean audio, text dependent</td>
</tr>
<tr>
<td>MOBIO [2]</td>
<td>160</td>
<td>6</td>
<td>6</td>
<td>64 x 80 x 1</td>
<td>48kHz, frontal face, clean audio, text independent</td>
</tr>
<tr>
<td>MSU-AVIS (Proposed)</td>
<td>50</td>
<td>3</td>
<td>12</td>
<td>1920 x 1080 x 1</td>
<td>148kHz, face pose-expression-distance variation, indoor, clean &amp; degraded audio, text independent</td>
</tr>
</tbody>
</table>

Benchmark Results and Analysis

<table>
<thead>
<tr>
<th>Methods</th>
<th>Description</th>
<th>Face Failure Subset</th>
<th>Auxiliary Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face-CNN [3]</td>
<td>$F_{\text{face}} = S_1$</td>
<td>0</td>
<td>0.15</td>
</tr>
<tr>
<td>Speaker-CNN [4]</td>
<td>$F_{\text{spkr}} = S_2$</td>
<td>10.98</td>
<td>0.06</td>
</tr>
<tr>
<td>Sum Rule</td>
<td>$F_{\text{sum}} = S_1 + S_2$</td>
<td>18.62</td>
<td>0.10</td>
</tr>
<tr>
<td>Product Rule</td>
<td>$F_{\text{prod}} = S_1 \times S_2$</td>
<td>19.60</td>
<td>0.12</td>
</tr>
<tr>
<td>Fusion Rule-1</td>
<td>$F_1 = S_1 \times S_2 \times e^{\frac{(S_1 - S_2)^2}{(S_1 + S_2)^2}}$</td>
<td>18.43</td>
<td>0.09</td>
</tr>
<tr>
<td>Fusion Rule-2</td>
<td>$F_2 = W_1 \times S_1 + W_2 \times S_2$</td>
<td>14.90</td>
<td>0.11</td>
</tr>
<tr>
<td>Fusion Rule-3</td>
<td>$F_3 = W_1 \times S_1 \times W_2 \times S_2$</td>
<td>19.60</td>
<td>0.10</td>
</tr>
<tr>
<td>Fusion Rule-4</td>
<td>$F_4 = (W_1 \times S_1)(W_2 \times S_2) \times e^{\frac{(W_1 \times S_1)(W_2 \times S_2)}{(W_1 \times S_1)(W_2 \times S_2)}}$</td>
<td>19.60</td>
<td>0.10</td>
</tr>
</tbody>
</table>

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