Motivation & Objective

- Human behavior is inherently complex and multimodal
- We present a method of fusing audio and visual information for modeling human behavioral expressions in the context of couples’ problem solving discussions
- We demonstrate performance gains achieved through multimodal fusion for predicting human annotated behavioral codes

Data

Couples Therapy Corpus
- Audio-Visual recordings of 134 married couples
- Ten-minute dyadic interactions of the couples discussing a problem in their marriage
- The sessions were manually annotated for relevant behaviors at the session level
- For this work we model acceptance and blame

Methodology

Audio Feature Extraction

Visual Feature Extraction

Multiple Instance Learning

Classifier Fusion

Experimental Results

Classification accuracy (%) using audio, visual, and audio-visual fusion

<table>
<thead>
<tr>
<th>behavior</th>
<th>audio</th>
<th>visual</th>
<th>fusion early</th>
<th>late</th>
</tr>
</thead>
<tbody>
<tr>
<td>acceptance</td>
<td>70.5</td>
<td>62.5</td>
<td>64.3</td>
<td>72.3</td>
</tr>
<tr>
<td>blame</td>
<td>69.4</td>
<td>57.4</td>
<td>70.4</td>
<td>71.3</td>
</tr>
</tbody>
</table>

Late fusion improves accuracy for classification of both behaviors

Conclusions & Future Work

- We have demonstrated a method for combining audio and visual information, using a multiple instance learning approach for saliency estimation, and predicting human annotated behavioral ratings
- Late fusion outperformed early fusion for both behavioral codes
- In the future, we would like to explore how this methodology performs on other behaviors
- Study how spouses influence one another’s behaviors throughout each session
- Explore different ways to model saliency