Multi-modal Laughter Recognition in Video Conversations

Sergio Escalera,
Eloi Puertas,
Petia Radeva,
Oriol Pujol

25/6/2009
Laughter detection
Audio features
Visual features
- Smile-laughter detection
Multi-modal fusion
- Stacked Sequential Learning
Results
Conclusions
Laughter detection

- Relaxed situations, agreement signal, welcome response, affective states, jokes, etc.
- Affective computing and Human-Computer interaction
- Audiovisual feedback is a key point
- Multi-modal fusion
Audio Features

- **Spectogram**
  Interleaved sliding window of size 256 samples with relative displacements of 128 samples.

- **Accumulated power**
  \[ \sum s(\omega) \]

- **Spectral entropy**
  \[ (-\sum s(\omega) \log s(\omega)) \]

- **Fundamental frequency**
  computed by finding the peak in the band between 20 - 500 Hz.
Visual Features – Smile/Laughter detector

Training visual features
- Face detection
  - Cascade of classifiers
- Mouth detection
  - Facial characteristics
- Learning
  - PCA and Adaboost

Testing visual features
- Face detection
  - Cascade of classifiers
- Mouth detection
  - Facial characteristics
- Mouth features
  - Mouth movement
  - Laugh label
Visual Features – Face detector

Haar-like

Integral image

Cascade of classifiers

Positive Samples

Negative Samples

D = (4 + 1) − (2 + 3)

• New York Times opinion Bloggings data set (http://video.nytimes.com/)

• 600 positive and 2500 negative samples from different speakers

• All samples are resized to a resolution of $25 \times 40$ pixels.

• PCA and saving a 99% of principal components

• 70 features per sample are obtained.

• 100 iterations of Gentle Adaboost with decision stumps.
Visual Features – Mouth historical

- Historical mouth movement

\[ F_i \in \{0, ..., 255\}^{n \times m} \]
\[ M_i \in \{0, ..., 255\}^{n/2 \times m/2} \]
\[ MM_{il} = \frac{1}{n \cdot m/4} \sum_{j=i-l}^{i-1} \sum_k |M_{i,k} - M_{j,k}| \]

\( h_{MM} \)

\( P_{MM} \)

\[ t_1 : \int_0^{t_1} P_{MM} = \frac{1}{3}, \quad t_2 : \int_0^{t_2} P_{MM} = \frac{2}{3} \]
- Non independent and identically distributed samples
- Neighboring samples within an interval have some kind of relationship

Combination by increasing the input space with features of the neighboring label
Results - settings

- **Data**
  - 18 video sequences (2 mosaic)
  - 5 min. 12 FPS : 2880 frames

- **Methods**
  - Gentle Adaboost 50 d. stumps
  - Sequential windows size of 11

- **Measurements**
  - Stratified ten-fold cross-validation
  - Accuracy, sensitivity, and specificity measures.

- **Experiments**
  - Observers inquiry – interest?
  - Multimodal laughter detection
### Results – Observer’s interest

- **40 observers**
- Sorting based on conversation preferences

<table>
<thead>
<tr>
<th></th>
<th>Video 1</th>
<th>Video 2</th>
<th>Video 3</th>
<th>Video 4</th>
<th>Video 5</th>
<th>Video 6</th>
<th>Video 7</th>
<th>Video 8</th>
<th>Video 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mosaic 1</td>
<td>5.4(1.0)</td>
<td>5.3(0.8)</td>
<td>4.3(0.9)</td>
<td>3.3(0.6)</td>
<td>2.7(0.6)</td>
<td>6.7(0.8)</td>
<td>6.4(1.0)</td>
<td>3.1(1.0)</td>
<td><strong>7.9(0.6)</strong></td>
</tr>
<tr>
<td>Laugh period</td>
<td>8</td>
<td>6</td>
<td>20</td>
<td>14</td>
<td><strong>39</strong></td>
<td>3</td>
<td>3</td>
<td>15</td>
<td><strong>0</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Video 1</th>
<th>Video 2</th>
<th>Video 3</th>
<th>Video 4</th>
<th>Video 5</th>
<th>Video 6</th>
<th>Video 7</th>
<th>Video 8</th>
<th>Video 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mosaic 2</td>
<td>3.4(0.9)</td>
<td>4.3(0.8)</td>
<td>4.8(0.9)</td>
<td><strong>7.2(1.0)</strong></td>
<td>4.2(1.2)</td>
<td>5.9(1.0)</td>
<td>4.2(1.0)</td>
<td>6.8(0.8)</td>
<td>4.3(0.9)</td>
</tr>
<tr>
<td>Laugh period</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td><strong>0</strong></td>
<td>33</td>
<td>11</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

*Ranking positions and confidence interval of dyadic interactions*
Results – Automatic laughter detection

• **Learning:** Adaboost and Sequential Stacked Learning procedures.

• **Features:**
  - **Audio cue:** sampling at 8000 Hz per second.
    Final audio sequences of 15000 positions and 134 features are obtained.
  - **Visual cue:** 3 visual features (1 smile/laughter + 2 mouth movement degree)

\[
M = \begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix}
\]

\[
ACC = \frac{TN + TP}{TN + FP + FN + TP}
\]

\[
SE = \frac{TP}{TP + FN}
\]

\[
SP = \frac{TN}{TN + FP}
\]

<table>
<thead>
<tr>
<th></th>
<th>ACC</th>
<th>SE</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaboost Audio</td>
<td>0.70</td>
<td>0.51</td>
<td>0.70</td>
</tr>
<tr>
<td>Adaboost Audio-Video</td>
<td>0.70</td>
<td>0.52</td>
<td>0.70</td>
</tr>
<tr>
<td>Sequential Audio</td>
<td>0.81</td>
<td>0.61</td>
<td>0.81</td>
</tr>
<tr>
<td>Sequential Audio-Video</td>
<td>0.77</td>
<td>0.65</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Laughter recognition results

- Unbalanced problem
Conclusions & future work

- Simple but discriminative audio/visual features for laughter detection
- Sequential Learning as a way of fusing audio-visual cues
  - Performance improvements
- Audio/visual fusion increases sensitivity of a very unbalanced problem

Future work
- Post filtering using temporal knowledge (isolated positive/negative detections)
- Complementary audio features and invariant visual features
- Comparative of multi-modal fusion methodologies: CRF, HMM, etc.


Thank you!!