Analysis and Compensation of the Reaction Lag of Evaluators in Continuous Emotional Annotations

Soroosh Mariooryad and Carlos Busso

Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas

Sep 4, 2013
Emotional Descriptors

• Emotional labels at sentence level
  • One descriptor assigned to a segment
    • sentence, turn, chunk, word
  • Long segments: variations are not captured

• Continuous labels
  • Track emotional content continuously over time
  • They capture localized emotional behaviors
  • Facilitate emotion analysis at different resolutions
Continuous Emotional Labels

- Record position of a cursor controlled by user
- Examples of these GUIs:
  - FEELTRACE [Cowie et al. 2000] and Gtrace [Cowie et al., 2012]
  - MoodSwings [Kim et al., 2008] and EmuJoy [Nagel et al., 2007]
Discrete Classification Problem

- **Approach:** estimate mean across evaluators

Segment $\in \begin{cases} 
\text{Class 1} & \text{if } \bar{a} < 0.5 \\
\text{Class 2} & \text{if } \bar{a} \geq 0.5 
\end{cases}$

(Or regression model)
Motivation

- Emotional classification results in naturalistic database was very low - SEMAINE [McKeown et al., 2012]
- Challenging task - spontaneous emotions

AVEC 2011

Is there any other reason for low performance?

Source: http://sspnet.eu/avec2011
Evaluator Reaction Lag

- Emotion assessment
  - Sense the stimuli, appraise the emotional message, define their judgment, moving the cursor

Underlying emotional profile

Thursday, September 5, 13
Problem Formulation

• How to formulate the estimation of the reaction lag?
  • Constant reaction lag or time-variant
  • Annotator-dependent or annotator-independent

• Assumptions in this work
  • Constant reaction lag across time
  • Annotator-independent (mean across evaluators)
    • Preliminary results on annotator-dependent
Estimating Reaction Lag

- Proposed approach based on mutual information (MI)
  - Capture the dependency between two random variables
- Find the optimal reaction lag

\[
\hat{\tau} = \arg_{\tau} \max I[EMO; ANN^\tau]
\]

- \( EMO \) = emotional content of the stimulus
- \( ANN^\tau \) = shift version of emotional annotation
Estimation of Emotional Content

\[ \hat{\tau} = \arg_{\tau} \max I[EMO; ANN^\tau] \]

- \textit{EMO} represented by facial features capturing the deviations from neutral behaviors (\(EMO^F\))
- Why acoustic features are not included?
  - During silence, speech features are not available
  - Single frame does not convey enough emotion cues
- Distributions are estimated with k-means
  - \(P(EMO^F)\)
  - \(P(ANN^\tau)\)
  - \(P(ANN^\tau, EMO^F)\)
SEMAINE Database

- Emotionally colored interactions
- Annotations: FEELTRACE (activation, valence)
- 44 sessions, 9 unique speakers (users)
  - Sessions with annotations and correctly extracted facial features

Source: McKeown et al. (2012)
Facial Features

- Facial features extracted with CERT [Bartlett et al. 2006]
- Action Units from FACS (deviation from neutral faces)
- Head rotation (Jaw, Yaw and Pitch)

<table>
<thead>
<tr>
<th>AU</th>
<th>Description</th>
<th>AU</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU 1</td>
<td>Inner Brow Raise</td>
<td>AU 15</td>
<td>Lip Corner Depressor</td>
</tr>
<tr>
<td>AU 2</td>
<td>Outer Brow Raise</td>
<td>AU 17</td>
<td>Chin Raise</td>
</tr>
<tr>
<td>AU 4</td>
<td>Brow Lower</td>
<td>AU 18</td>
<td>Lip Pucker</td>
</tr>
<tr>
<td>AU 5</td>
<td>Eye Widen</td>
<td>AU 20</td>
<td>Lip stretch</td>
</tr>
<tr>
<td>AU 6</td>
<td>Cheek Raise</td>
<td>AU 23</td>
<td>Lip Tightener</td>
</tr>
<tr>
<td>AU 7</td>
<td>Lids Tight</td>
<td>AU 24</td>
<td>Lip Presser</td>
</tr>
<tr>
<td>AU 9</td>
<td>Nose Wrinkle</td>
<td>AU 25</td>
<td>Lips Part</td>
</tr>
<tr>
<td>AU 10</td>
<td>Lip Raise</td>
<td>AU 26</td>
<td>Jaw Drop</td>
</tr>
<tr>
<td>AU 12</td>
<td>Lip Corner Pull</td>
<td>AU 28</td>
<td>Lips Suck</td>
</tr>
<tr>
<td>AU 14</td>
<td>Dimpler</td>
<td>AU 45</td>
<td>Blink/Eye Closure</td>
</tr>
</tbody>
</table>

- For $EMOF^F$, we use $K \in \{2, 4, 6, 8, 10, 16, 20\}$ over the joint feature space
Analysis of the Reaction Lag

- Activation

Clusters (ANN)

K = 2

K = 3

K = 4

Lag analysis

\[ P(ANN^\tau) \]
Analysis of the Reaction Lag

- Valence

Clusters (ANN)

Lag analysis

K = 2

K = 3

K = 4

Thursday, September 5, 12
Analysis of the Reaction Lag

• [Activation, Valence]

Clusters (ANN)

K = 2

K = 3

K = 4

Lag analysis
Experimental Setting

- The optimal delay is defined as the first time the mutual information does not increase
- Priority to shorter reaction lag

<table>
<thead>
<tr>
<th>Attribute</th>
<th>K=2</th>
<th>K=3</th>
<th>K=4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std</td>
<td>mean</td>
</tr>
<tr>
<td>Act</td>
<td>2.27</td>
<td>0.82</td>
<td>2.84</td>
</tr>
<tr>
<td>Val</td>
<td>3.48</td>
<td>0.66</td>
<td>3.68</td>
</tr>
<tr>
<td>Act-Val</td>
<td>3.61</td>
<td>0.52</td>
<td>4.98</td>
</tr>
</tbody>
</table>
Validation with Emotion Recognition

- 1049 turns (at least 300ms long) - 9 subjects
- SVM with 9-fold speaker independent cross-validation
- Evaluation settings
  - Activation, valence, and [activation, valence]
  - Discrete emotional labels with $K=2, 3, 4$ classes
  - Reaction lag: 0, 1, 2 and 3 sec + optimal delay
- Facial features
  - [AUs+head] x 6 statistics (e.g., quantiles, mean and std)
- Acoustic features
Acoustic Features

• **openSMILE 4368 features** [Eyben et al. 2010, Schuller et al. 2011]

  - **Spectral**
    - Rasta-style filtered auditory spectrum bands
    - MFCCs
    - Spectral energy 25-60Hz, 1k-4KHz
    - Spectral roll-off point 0.25 0.50 0.75 0.90
    - Spectral Flux, entropy, variance, skewness, kurtosis, slope

  - **Energy**
    - Sum of auditory spectrum (loudness)
    - Sum of Rasta-style filtered auditory spectrum
    - RMS Energy
    - Zero-Crossing Rate

  - **Voice**
    - F0
    - Probability of voicing
    - Jitter (local, delta)
    - Shimmer

• Feature selection with CFS (~ 99 features)
Recognition Experiments - Activation

**K = 2**

**K = 3**

**K = 4**

Face

Speech

Thursday, September 5, 13
Recognition Experiments - Valence

K = 2

K = 3

K = 4

Face

Speech

Thursday, September 5, 13
Recognition Experiments - [Act, Val]

Face

Speech

K = 2

K = 3

K = 4

Thursday, September 5, 13
Recognition Experiments - [Act,Val]

The optimal delay gives statistically significant improvements in accuracy (p-value<0.008) and F-score (p-value<0.045) for 

K = 2 and K = 4

The optimal delay gives statistically significant improvements in F-score (p-value<0.007) for K = 2 and K = 4
Recognition Experiments - Average

- Across all settings
  - Act, Val, [Act-Val]
  - K = 2, 3, 4

- Optimal delay estimated from training set yields the best performance across all settings on the test set
Experiments – Pre-Aligning the Annotations

- Evaluator dependent lag
  - Assumption: phase between two annotators is fixed and is less than 1 sec
  - Pre-Aligning the labels of multiple annotator to maximize the correlation between them within [-1, 1] seconds
  - F-score improves 1.06% (face) and 0.26% (speech)
Conclusions

• The mutual information analysis unveils and quantifies the reaction lag with respect to facial features
• Compensating for the reaction lag improves the performance of both facial and vocal emotion recognition systems
• Shift-delayed emotional annotations achieved statistically significant improvements

We are using the wrong labels!
Future Work

• Reaction lag analysis with respect to speech features

• Reaction lag analysis in evaluator-dependent fashion
  • Find optimum delay per annotation

• Considering time-variant reaction lag
  • Time warping methods e.g., dynamic probabilistic canonical correlation with time warping (DPCTW) [Nicolaou et al., 2012]
Multimodal Signal Processing (MSP)

Thanks!

Work funded by Samsung Telecommunications America and NSF

http://msp.utdallas.edu/