Multimodal affect recognition

Tiina Lindh-Knuutila
Jaakko Väyrynen
14.3.2005

Multimodal affect recognition


Images for the data collection task: Japanese Female Facial Expression (JAFFE) database

Outline

- Introduction
- Emotions
- Problem domain
- Affect-recognition from single modalities
- Affect-recognition from multiple modalities
- Conclusions
- Data collection task for exercise

Background

- Psychological theories of affective states
- Emotional intelligence measures communication skills
  - recognition of affective states
  - interpersonal social communication
- Based on nonverbal communicative cues
Affective computing

- Target: emotionally intelligent human-computer interaction (HCI)
- Tasks
  - sensing
  - tracking
  - analysis
  - affect arousal

Motivation

- More human-like interaction
  - natural
  - trustworthy
  - efficacious
  - persuasive
  - may cause problems
- Benefits in surveillance, monitoring, interpreting, indexing, ...

Views on emotions (1)

- Classical view
  - basic expressions of emotions
    - happiness, anger, sadness, surprise, disgust, fear
    - hardwired into specific neural structures
  - recognized cross-culturally

Views on emotions (2)

- Russell
  - multidimensional affect space
  - critique of experiment design
- Ortony and Turner
  - components of emotions are linked with communicative displays
- Social constructivists (Averill)
  - interpretation and response to classes of situations
  - do not explain the genuine feeling
**Multimodal emotional cues**

- Multimodal analysis of multiple communication channels
- Modalities
  - sight, hearing, touch
- Cues from different modalities
  - e.g. vocal intonation, facial expression
- The modalities support each other
- Recognition depends on many factors

**Emotions: summary**

- No consensus of
  - basic emotions
  - expressions of emotions
- Limited set of emotions
- Display of emotions most likely culturally dependent

**Fundamental research questions**

- What is an affective state?
- What kinds of evidence warrants conclusions about affective states?
- How can various kinds of evidence be combined to generate conclusions about affective states?

**Technical questions**

- How should emotions be recognized?
  - different modalities
  - obtrusive methods
- Human-like performance
  - human-like sensors?
  - human-like recognition level?
Methodological questions

- What are the appropriate channels?
- How to combine the information conveyed by the channels?
- How to handle temporal aspects?
- How to make them context-sensitive?

Ideal system?

Affect-recognition from single modalities

- Choice of selected moods application dependent
- Context is not taken into account
- Modalities: haptic, visual, audio
  - single tactile-based affect recognition study
    - data collection not comfortable
    - signals measured
      - electromyogram from jaw, blood volume pressure, skin conductivity, respiration and heart rate
  - audio- and visual data based recognition next
**Single modality: Face/visual**

- Three subproblems
  - finding the face
  - detecting facial features
  - classifying data to affect categories
- Various classification techniques
- Focused at attempts to recognize a small set of posed prototypic facial expressions of basic emotions

**Features with emotional correlation**

<table>
<thead>
<tr>
<th></th>
<th>Happiness</th>
<th>Anger</th>
<th>Fear</th>
<th>Sadness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pitch</strong></td>
<td>increase</td>
<td>increase in</td>
<td>increase in</td>
<td>decrease in</td>
</tr>
<tr>
<td></td>
<td>mean, range, variability</td>
<td>range, variability</td>
<td>mean, range</td>
<td>mean, range</td>
</tr>
<tr>
<td><strong>Intensity</strong></td>
<td>increased</td>
<td>increased</td>
<td>normal</td>
<td>decreased</td>
</tr>
<tr>
<td><strong>Duration (speech rate)</strong></td>
<td>increased rate, increased</td>
<td>increased, reduced rate, reduced rate</td>
<td>normal</td>
<td>decreased</td>
</tr>
<tr>
<td><strong>Pitch contour</strong></td>
<td>descending line, stressed syllables ascend frequently &amp; rhythmically, irregular up &amp; down inflection</td>
<td>disintegration in pattern great number of changes in the direction</td>
<td>descending line</td>
<td></td>
</tr>
</tbody>
</table>
Chen et al.
- Rule-based method
- Speech: pitch, intensity, pitch contours
- Video: facial features e.g., raising/lowering the eyebrows
- No separate test set
- Quantification of the recognition rate is not reported

De Silva and Ng
- Rule-based method
- Speech: pitch, pitch contours
- HMM-based classification into emotion classes
- Video: displacement and velocity of e.g., mouth corners with the optical flow method
- nearest neighbor classification into emotion classes
- 72% recognition rate for a reduced data set

Yoshitomi et al.
- Hybrid method
- Speech: pitch, intensity, pitch contours
  - HMM classification into emotions
- IR and VR images of maximal intensity for the syllables in the word 'Ta-ro'
  - extraction of regions of interests (mouth, eyebrow...)
  - differential image based on 'neutral' images
  - DCT of differential IR and VR images fed to an ANN
- Summing of classifications for the final decision
- 85% recognition rate for a reduced data set

Chen and Huang
- Set of methods
- Speech: pitch, intensity, speech rate
  - classification using Gaussian distributions
- Video: facial motion tracking
  - piecewise Bezier volume deformation model (3D)
  - 12 predefined facial muscle actions estimated
  - classification by a sparse network of winnows with naive Bayes output nodes
- 79% person-dependent recognition rate
- 53% person-independent recognition rate
**Challenges: visual**

- Scale
- Resolution
- Pose
- Occlusion
- Changing illumination
- Movement, tracking

**Challenges: audio**

- Unconstrained continuous speech
- naturally spoken
- rather meaningful than semantically neutral content
- range of speakers and languages
- Development of better affective state features

**Challenges: multimodal input (1)**

- Handling partial, missing and erroneous data
- methods: HMM, SVM
- Unsupervised learning of human behavioral grammar
- application, user and context-dependent grammars
- Integration of modalities at the feature level
- context dependent models
- methods: Bayesian inference, ...

**Challenges: multimodal input (2)**

- Affect-sensitive interpretation of multimodal input
- Context sensitivity
- Multiple-emotion categories
- Other than 'basic' emotions
- Unsupervised learning for the interpretation
**Validation issues**

- Proposal of a commonly used audio-visual database for the validation of the results

**Conclusions**

- Perceiving emotions has a multimodal nature
- State-of-the-art systems not quite mature yet
  - most use only a single modality
  - context is not taken into account
- Future

**Data collection for the exercise**

- Please fill in the distributed forms