Automatic detection of features of emotional state

...or ‘what analysis can provide in terms of standards’

Kostas Karpouzis

Humaine WP10 workshop, Vienna, November 6-8, 2006

the HUMAINE Research Area

HUMAINE WP10 workshop, Vienna, November 6-8, 2006
so, where is WP4 in all this?

• a very bold statement:
  - *there is little sense in trying to standardize analysis, especially in the visual channel*
  - a single computer vision (CV) algorithm is fine-tuned to work better on specific sequences
  - e.g. uses morphological elements with discrete, hand-picked size or other kinds of thresholds

• so any claims to provide ‘universal’ CV algorithms are unrealistic
  - and, therefore, cannot be part of a standard

still, what WP4 can do for you?

• provide automatic annotation of input audiovisual data
  - use for training and recognition
  - use to drive ‘emotion cloning’ in ECAs

• lower-level vs high-level annotation depends on application
visual channel

• so, you’ve heard of FAPs...
  - used extensively in driving ECA animation
  - MPEG-4 compatible like Greta
• now, we also have BAPs!
  - the same thing, but for hand and body part rotation
  - minor technical detail: impossible to calculate 3D data from a single video, so we depend on kinematics

more visual channel

• in cases of naturalistic data
• e.g. EmoTV
  - FAP estimation is impossible for inter-ocular distances less than 50 pixels
  - or when color information is hampered, e.g. VHS tapes
• however, analysis can provide other cues
  - general head, hand or body movement
  - can be related to expressivity with promising results (>75%)
aural channel

- The FAU prosody module I: integrated into end-to-end system, developed within VERBMOBIL 1992-2000, multi-functional, omnibus (recognition of accents, boundaries, disfluencies, off-talk, emotions, ...)
- A prosody module II based on voiced segments, developed within SmartKom 1999-2003, somehow inspired by the other one

aural channel (module I)

- Input to the prosody module: recognition result and word hypotheses graph (WHG)
- Computation of prosodic features on word level or across several words
- Local features from silent and filled pauses, signal energy, word duration, F0 (here: 95 features)
- Global features from jitter, shimmer, voiced/voiceless decisions (here: 15 features)
aural channel (module II)

- unit not word, but voiced segment, e.g. > 50 msec
- pauses, duration, energy, F0, jitter/shimmer, FFT co-efficients
- mean/max/min values computed for whole turn
- 219 features
- for any time domain/unit, computation of features (i.e. structured features such as mean for basic features such as F0) possible without much scripting/programming

producing emotion labels

- quoted from Marc and Hannes’ presentation:
  - One design principle for EARD was that simple cases should look simple. For example, annotating text with a simple “pleasure” emotion results in a simple structure
- automated results are usually very low-level
  - e.g. FAPs describe ‘atomic’ facial movements
- human-readable representations should be higher-level
  - e.g. emotion labels or dimensions
different time scales

- `<emotion category="pleasure">Hello!</emotion>`
- `<emotion start="0.4" end="1.3" category="pleasure"/>`
- possible for visual, aural or multimodal recognition

additional information

- **dimensions**
  - `<emotion category="pleasure" regulation="simulate" intensity="0.9"/>`
- **confidence**
  - my favorite!
  - since it bridges the gap with recognition
  - `<emotion category="pleasure" modality="face" confidence="0.5"/>`
more complex information

• e.g. mixture of emotions
  - `<complex-emotion`
    `<emotion category="pleasure" intensity="0.7"/>`
    `<emotion category="worry" intensity="0.5"/>`
  `</complex-emotion>`
• *is* possible, but with ‘proper’ training
  - which essentially means ‘proper’ annotation
  - or ‘what constitutes worry?’

more complex information

• multimodality
  - `<complex-emotion start="0.4" end="1.3">`
    `<emotion category="pleasure" modality="face"/>`
    `<emotion category="worry" modality="voice"/>`
  `</complex-emotion>`
• the main issue here is that *usually* speech and face features work on different time frames
  - visual information is calculated per frame
  - aural information is calculated per turn, word or arbitrary time window
  - maybe try to meet in the middle?