Person Detection from Audio

“Speaker Diarization / Segmentation”: given multi-party audio data (possibly with background noise):

→ who talks when?

• Typically 3 steps:
  -- segmentation into speech / non-speech
  -- detection of speaker transitions
  -- clustering of speaker segments (+ classification of speaker)

• Segmentation into speech / non-speech:
  -- Generate features:
    ++ digital signal (pre-) processing
      (involving e.g. sub-division signal into overlapping samples of typically several ms, Fourier-transform etc.)
    ++ MEL filters → MEL cepstrum coefficients
    ++ Further Fourier- and other transformations
    ++ additional features: zero-crossing rates, energy statistics etc.
Person Detection from Audio

“Speaker Diarization / Segmentation”: given multi-party audio data (possibly with background noise):
 → who talks when?

• Typically 3 steps:
  -- segmentation into speech / non-speech
  -- detection of speaker transitions
  -- clustering of speaker segments (+ classification of speaker)

• Segmentation into speech / non-speech:
  -- Generate features:
    ++ digital signal (pre-) processing
    (involving e.g. sub-division signal into overlapping samples of typically several ms, Fourier-transform etc.)
    ++ MEL filters → MEL cepstrum coefficients
    ++ Further Fourier- and other transformations
    ++ additional features: zero-crossing rates, energy statistics etc.

Person Detection from Video

• First step: Face detection
  Naive approach: simple pixel based binary classifier.
  Problem: too many possibilities for non-faces

• Other approaches:
  • detect correct relatively positioned patches of skin, eyes or other face elements. Advantage: relatively robust against rotations
  • Approach [6]: Use special features instead of pixels (advantage: domain knowledge can be encoded into features), Intelligent feature selection / combination of simple binary classifiers that work on single features (AdaBoost)

  (optional second step: face recognition (e.g. via Eigenfaces (via PCA) [5])
Person Detection from Video

First step: Face detection
- Naive approach: simple pixel based binary classifier. Problem: too many possibilities for non-faces
- Other approaches:
  - detect correct relatively positioned patches of skin, eyes or other face elements. Advantage: relatively robust against rotations
- Approach [6]: Use special features instead of pixels (advantage: domain knowledge can be encoded into features), intelligent feature selection / combination of simple binary classifiers that work on single features (AdaBoost)
  - (optional second step: face recognition (e.g. via Eigenfaces (via PCA) [5])

Person Detection from Video

First step: Face detection
- Naive approach: simple pixel based binary classifier. Problem: too many possibilities for non-faces
- Other approaches:
  - detect correct relatively positioned patches of skin, eyes or other face elements. Advantage: relatively robust against rotations
- Approach [6]: Use special features instead of pixels (advantage: domain knowledge can be encoded into features), intelligent feature selection / combination of simple binary classifiers that work on single features (AdaBoost)
  - (optional second step: face recognition (e.g. via Eigenfaces (via PCA) [5])
Person Detection from Video

- **First step: Face detection**
  - Naive approach: simple pixel based binary classifier.
    Problem: too many possibilities for non-faces
  - Other approaches:
    - detect correct relatively positioned patches of skin, eyes or other face elements. Advantage: relatively robust against rotations
  - Approach [6]: Use special features instead of pixels (advantage: domain knowledge can be encoded into features), intelligent feature selection / combination of simple binary classifiers that work on single features (AdaBoost)
  - (optional second step: face recognition(e.g. via Eigenfaces (via PCA) [5])

Detecting Social Signals: Gestures and Posture

- **Human figure detection**
  - Main problem: too many options (clothes, accessories) → pixels as features won’t work
  - Approaches:
    - features: histograms of directions of detected edges

- **Gesture recognition: main challenges:**
  - detecting gesture-relevant body parts: select feature spaces, e.g. via
    - histograms of oriented gradients
    - etc.
  - modeling temporal dynamic e.g. using
    - Hidden Markov Models (HMMs)
    - Conditional Random Fields (CRFs)
    - Dynamic Time Warping (DTW)
    - etc.

Fig. 8. People detection. Examples of people detection in public spaces (pictures from [20]).
Detecting Social Signals: Gestures and Posture

- **Gestures:**
  - not many studies yet interpreting them as social signals
  - several studies: gestures as means of input
    - special example: touch interfaces
  - other study: automatic interpretation of sign language

- **Gesture recognition: main challenges:**
  - detecting gesture-relevant body parts: select feature spaces, e.g. via
    - histograms of oriented gradients
    - etc.
  - modeling temporal dynamic, e.g. using:
    - Hidden Markov Models (HMMs)
    - Conditional Random Fields (CRFs)
    - Dynamic Time Warping (DTW)
    - etc.

Detecting Social Signals: Gaze and Face

- **AU:** smallest discernable temporal feature sequence: sequence of geometry or appearance features (modeled e.g. via Dynamic Bayesian Networks (DBN))

- **Detection:** example: basic integrative methods based on optical flow on
detected faces:
  - optical flow: motion pattern of picture elements (e.g. pixels):
    - represented by vector field of velocity $V(x,y,t)$ of intensity:
      \[
      I(x + dx, y + dy, t + dt) = I(x, y, t) + \frac{\partial I}{\partial x} dx + \frac{\partial I}{\partial y} dy + \frac{\partial I}{\partial t} dt + O(d^2)
      \]
      \[
      \Rightarrow \frac{\partial I}{\partial x} V_x + \frac{\partial I}{\partial y} V_y + \frac{\partial I}{\partial t} = 0 \quad \text{(optical flow equation)}
      \]
      - use numeric methods to compute solutions

- **Features for facial expression recognition:**
  - geometric
    - shapes of facial components, locations of focal points, etc.
  - appearance
    - skin texture in different areas

Fig. 9. AU detection. Outline of a geometric-features-based system for detection of facial AUs and their temporal phases (onset, apex, offset, neutral) proposed in [10].

- **AU:** smallest discernable temporal feature sequence: sequence of geometry or appearance features (modeled e.g. via Dynamic Bayesian Networks (DBN))

- **Detection:** example: basic integrative methods based on optical flow on
detected faces:
  - optical flow: motion pattern of picture elements (e.g. pixels):
    - represented by vector field of velocity $V(x,y,t)$ of intensity:
      \[
      I(x + dx, y + dy, t + dt) = I(x, y, t) + \frac{\partial I}{\partial x} dx + \frac{\partial I}{\partial y} dy + \frac{\partial I}{\partial t} dt + O(d^2)
      \]
      \[
      \Rightarrow \frac{\partial I}{\partial x} V_x + \frac{\partial I}{\partial y} V_y + \frac{\partial I}{\partial t} = 0 \quad \text{(optical flow equation)}
      \]
      - use numeric methods to compute solutions
Detecting Social Signals: Gaze and Face

- AU: smallest discernable temporal feature sequence: sequence of geometry or appearance features (modeled e.g. via Dynamic Bayesian Networks (DBN))
- Detection: example: basic integrative methods based on optical flow on detected faces:
  --optical flow: motion pattern of picture elements (e.g. pixels):
  represented by vector field of velocity \( \mathbf{V}(x, y, t) \) of intensity:
  \[
  I(x + dx, y + dy, t + dt) = I(x, y, t) + \frac{\partial I}{\partial x} dx + \frac{\partial I}{\partial y} dy + \frac{\partial I}{\partial t} dt + O(d^2)
  \]
  \[
  \rightarrow \frac{\partial I}{\partial x} V_x + \frac{\partial I}{\partial y} V_y + \frac{\partial I}{\partial t} = 0 \quad \text{(optical flow equation)}
  \]
  use numeric methods to compute solutions
Detecting Social Signals: From Audio

- Vocal features: up to now: mostly investigated for speech detection
- Prosody: pitch, tempo, energy
  - pitch: first fundamental frequency (1st maximum in Fourier transform (e.g. 30ms frames))
  - tempo: vowels / sec.; vowel: phonetically relevant unit
  - energy E of signal s(t): $E = \sum_i s(t_i)^2$
- Few efforts so far in analysis of non-linguistic vocalizations
  - example: laughter detection (e.g. via SVMs)
  - and linguistic vocalizations
- silence detection: e.g. via energy as feature (often as by-product of speaker diarization)
**Context**

- Important issue: behavioral cues can have different meaning if happening in different outer contexts
- Example: temporal dynamics of behavioral cues / social signals (e.g. relative person–person timing, person-environment timing etc.)
- Other important issue: multi-modal combination / fusion of social signals (e.g. audio and interaction geometry)

**Applications**

- Predict outcome of dyadic interaction (selling, dating etc.) via audio and derived via social signals such as
  --activity (via energy),
  --influence (via stat. analysis of influence of A’s speaking patterns on B’s speaking patterns)
  --consistency (stability of person’s speaking patterns)
  --mimicry (mirroring)
- Eigenbehaviors (via PCA on features such as location, co-presence etc.)
- Analyzing interactions in small groups (e.g. meetings), role structures and detection of user interest via audio and video:
Applications

- Interactions in small groups: dominance of persons, recognition of collective actions
- Recognition of roles and extraction of small social networks (e.g., via analyzing meetings or broadcast TV shows)
- Analyze reaction of users to embodied conversational agents (ECAs)
Social Situation Models as Models of Social Context

**Social Situation:**

Co-located social interaction with full mutual awareness

**Simplified Social Situation Model:**

- Participating persons: \( P \): set of IDs
- Spatio-temporal reference: \( X \): sub-set of \( \mathbb{R} \times \mathbb{R}^3 \)
- \( S = (P, X) \)

---

Detecting Social Situations: Mobile Social Signal Processing

**Social Situation detection**

- **Example:** microphone \( \rightarrow \) audio-signals \( \rightarrow \) speaker diarization \( \rightarrow \) set of interacting persons
- **Example:** gyroscope, accelerometer, ultrasound-s. \( \rightarrow \) relative body distance & orientation \( \rightarrow \) set of interacting persons

**Social Situation understanding**

- **Example:** microphone \( \rightarrow \) audio-signals \( \rightarrow \) analysis of prosody \( \rightarrow \) emotion detection \( \rightarrow \) model of state of mind of person(s)
Detecting Social Situations: Mobile Social Signal Processing

Social Situation detection

- Example: microphone → audio-signals → speaker diarization → set of interacting persons
- Example: gyroscope, accelerometer, ultrasound-s. → relative body distance & orientation → set of interacting persons

Social Situation understanding

- Example: microphone → audio-signals → analysis of prosody → emotion detection → model of state of mind of person(s)

Research Questions

- Method for measuring human social interaction geometry (mobile → live; experiment → sociology model)
- General, quantitative, algorithmically processable model for human social interaction geometry
- Use of model to detect Social Situations (e.g. from mobile device measurements)
- Use as social context for applications maintaining privacy

Geometry of Social Interaction

Interpersonal distances

- Hall: “general quality” of social relation → 4 personal zones

Other influences (?): social context:
  - architectural environment (socio-petal, socio-fugal forces (Watson)), density, gender, etc.
  - individual context:
    - culture, age, self-esteem, disabilities,

Body angles

- Kendon: F-Formations [8]

Experiment

Camera
Model for Human Social Int. Geometry: Function $p(\delta \theta, \delta d)$

- **Idea**: Reduce n-ary social interaction to binary; infer n-ary by graph clustering

- **Binary**: $p(\delta \theta, \delta d)$
  
  - $\delta d(t) = \pm |P_{x,y} s_1(t) - P_{x,y} s_2(t)|$
  
  - $\delta \theta(t) = \theta_z (R_{12}(t)) = \theta_z \left( (R_1(t) (R_2(t))^T \right)$

- **Optional**: $p(\delta \theta, \delta d, \overline{\delta d})$

---

Experiment data: Manual annotation

| $\mathcal{S}^\oplus$ | $321307$ (\(\delta \theta, \delta d\)) pairs corresponding to “in a social situation”
| $\mathcal{S}^\ominus$ | $398335$ (\(\delta \theta, \delta d\)) pairs corresponding to “not in a social situation”

Example:

| $\mathcal{S}^\oplus$ | $4$
| $\mathcal{S}^\ominus$ | $6$
**Results**

**Experiment data: Manual annotation**

| \( S^{\Phi} \) | = 321307 \((\delta \theta, \delta d)\) pairs corresponding to "in a social situation"
| \( S^{\nabla} \) | = 398335 \((\delta \theta, \delta d)\) pairs corresponding to "not in a social situation"

**Example:**

\( | S^{\Phi} | = 4 \)
\( | S^{\nabla} | = 6 \)

---

**Experiment data: Manual annotation**

| \( S^{\Phi} \) | = 321307 \((\delta \theta, \delta d)\) pairs corresponding to "in a social situation"
| \( S^{\nabla} \) | = 398335 \((\delta \theta, \delta d)\) pairs corresponding to "not in a social situation"

**Example:**

\( | S^{\Phi} | = 4 \)
\( | S^{\nabla} | = 6 \)
Classification Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Mixture Model (3 Gaussians)</td>
<td>74.34 %</td>
</tr>
<tr>
<td>Gaussian Mixture Model (5 Gaussians)</td>
<td>74.67 %</td>
</tr>
<tr>
<td>Gaussian Mixture Model (7 Gaussians)</td>
<td>74.59 %</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>65.45 %</td>
</tr>
<tr>
<td>Support Vector Machine (Polyn. Kernel)</td>
<td>77.81 %</td>
</tr>
</tbody>
</table>

(*) w. 10-fold cross validation

Reconstructing Social Situations

- For each $i$ : complete weighted Graph $G(V,E,w,t)$ with $V$=set of persons,

$$w((s_i,s_j)) = \frac{p^{\Theta}(\delta s_i \cdot \delta s_j)}{p^{\Theta}(\delta s_i \cdot \delta s_j)}$$

- Average Link Clustering of $G(V,E,w,t)$ + Maximum Modularity
  Dendrogram Cut $\rightarrow$ Partition $X$ of $V$

- Compare $X$ with annotation $X'$ via $RAND(X,X')$ $\rightarrow$ Accuracy of Social Situation Detection for each $i$

- Average over all $t$ : RAND $\sim$ 0.76  Adj.Rand $\sim$ 0.529