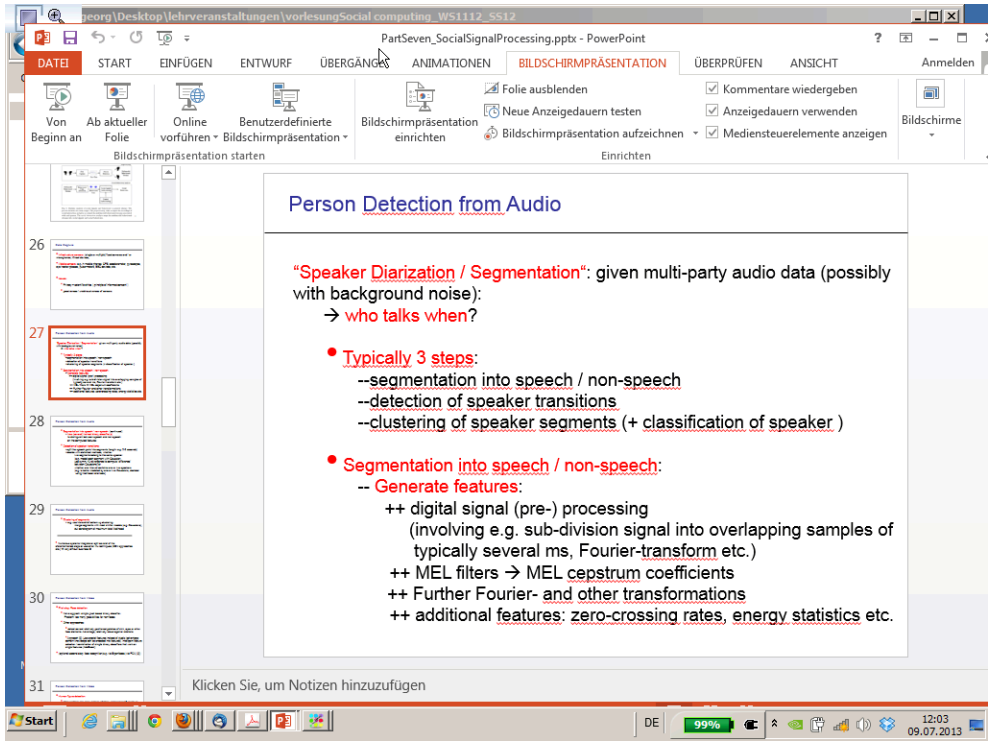
Script **generated by TTT**

Title: profile1 (09.07.2013)

Date: Tue Jul 09 12:03:17 CEST 2013

Duration: 90:34 min

Pages: 36



Person Detection from Video

- First step: Face detection
  - Naive apprach: 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 apprach: 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

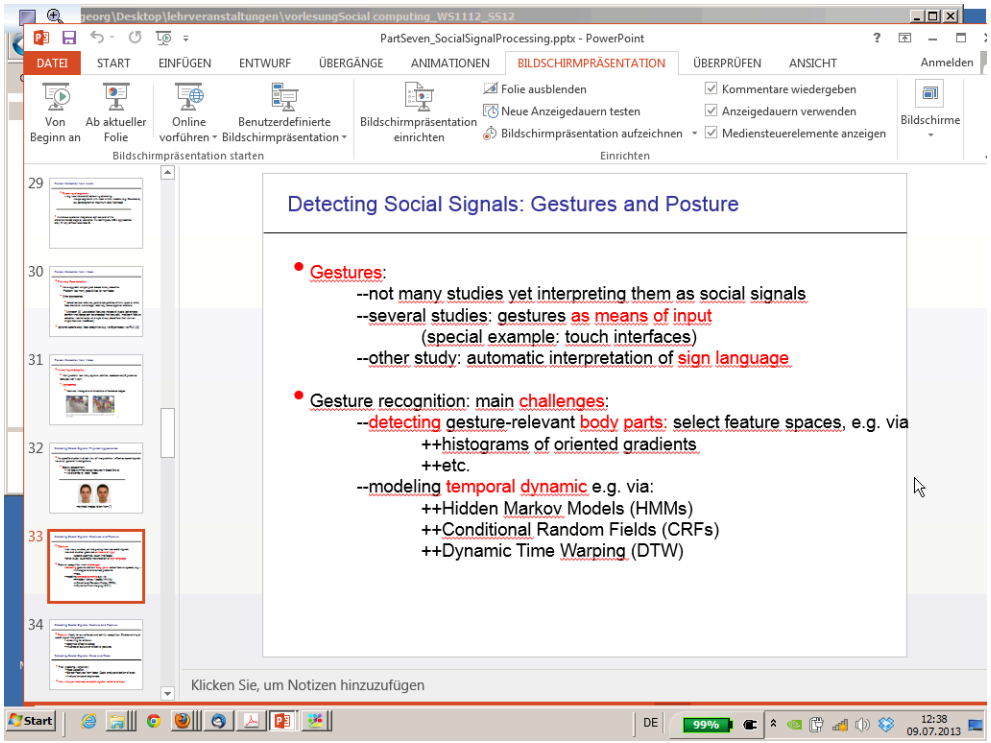
- First step: Face detection
  - Naive apprach: simple pixel based binary classifier.
    Problem: too many possibilities for non-faces
  - Other approaches:
    - detect correct relatively positioned patches of skin, eyes or other face elements. Advantage; relatively robust against rotations
    - Approach [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])

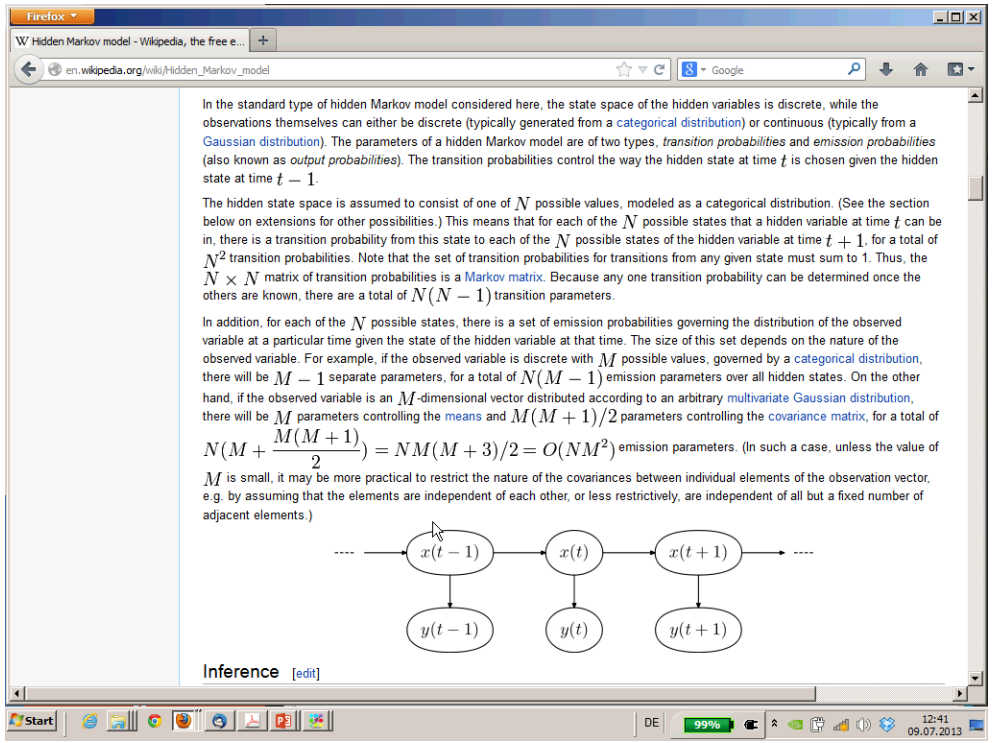
Person Detection from Video

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

Fig. 8. People detection. Examples of people detection in public spaces (pictures from [216]).

[1]





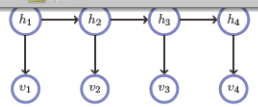


Figure 23.4: A first order hidden Markov model with 'hidden' variables $\text{dom}(h_t) = \{1, \ldots, H\}$, $t = 1 : T$. The 'visible' variables v_t can be either discrete or continuous.

23.2 Hidden Markov Models

The Hidden Markov Model (HMM) defines a Markov chain on hidden (or 'latent') variables $h_{1:T}$. The observed (or 'visible') variables are dependent on the hidden variables through an emission $p(v_t|h_t)$. This defines a joint distribution

$$p(h_{1:T}, v_{1:T}) = p(v_1|h_1)p(h_1)\prod_{t=2}^{T} p(v_t|h_t)p(h_t|h_{t-1}) \tag{23.2.1}$$

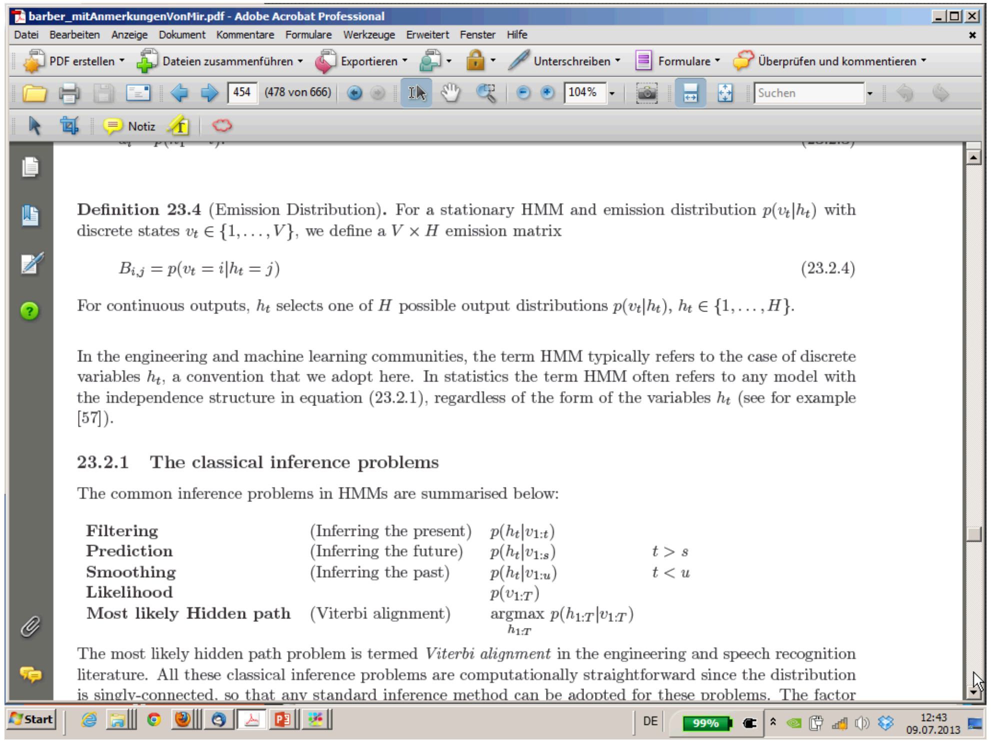
for which the graphical model is depicted in fig(23.4). For a stationary HMM the transition $p(h_t|h_{t-1})$ and emission $p(v_t|h_t)$ distributions are constant through time. The use of the HMM is widespread and a subset of the many applications of HMMs is given in section(23.5).

Definition 23.3 (Transition Distribution)**.** For a stationary HMM the transition distribution $p(h_{t+1}|h_t)$ is defined by the $H \times H$ transition matrix

$$A_{i',i} = p(h_{t+1} = i'|h_t = i) \tag{23.2.2}$$

and an initial distribution

$$a_i = p(h_1 = i). \tag{23.2.3}$$



Definition 23.4 (Emission Distribution). For a stationary HMM and emission distribution $p(v_t|h_t)$ with discrete states $v_t \in \{1, \ldots, V\}$, we define a $V \times H$ emission matrix

$$B_{i,j} = p(v_t = i|h_t = j) \qquad (23.2.4)$$

For continuous outputs, h_t selects one of H possible output distributions $p(v_t|h_t)$, $h_t \in \{1, \ldots, H\}$.

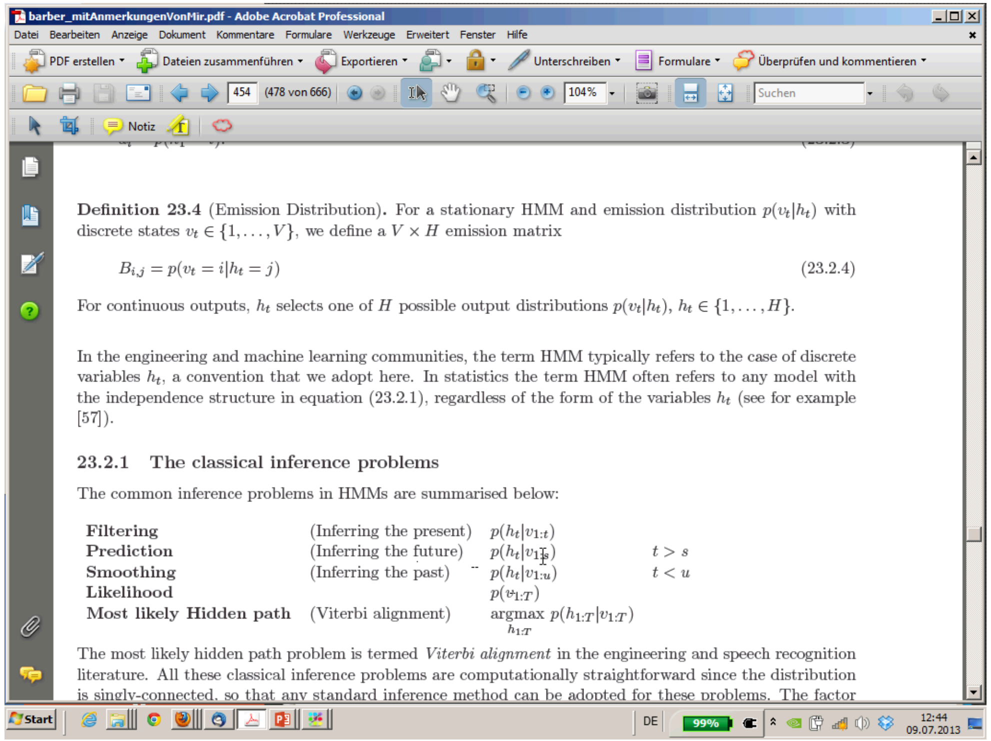
In the engineering and machine learning communities, the term HMM typically refers to the case of discrete variables h_t, a convention that we adopt here. In statistics the term HMM often refers to any model with the independence structure in equation (23.2.1), regardless of the form of the variables h_t (see for example [57]).

23.2.1 The classical inference problems

The common inference problems in HMMs are summarised below:

Filtering	(Inferring the present)	$p(h_t\|v_{1:t})$	
Prediction	(Inferring the future)	$p(h_t\|v_{1:s})$	$t > s$
Smoothing	(Inferring the past)	$p(h_t\|v_{1:u})$	$t < u$
Likelihood		$p(v_{1:T})$	
Most likely Hidden path	(Viterbi alignment)	$\underset{h_{1:T}}{\text{argmax}}\ p(h_{1:T}\|v_{1:T})$	

The most likely hidden path problem is termed *Viterbi alignment* in the engineering and speech recognition literature. All these classical inference problems are computationally straightforward since the distribution is singly-connected, so that any standard inference method can be adopted for these problems. The factor



Definition 23.4 (Emission Distribution). For a stationary HMM and emission distribution $p(v_t|h_t)$ with discrete states $v_t \in \{1, \ldots, V\}$, we define a $V \times H$ emission matrix

$$B_{i,j} = p(v_t = i|h_t = j) \qquad (23.2.4)$$

For continuous outputs, h_t selects one of H possible output distributions $p(v_t|h_t)$, $h_t \in \{1, \ldots, H\}$.

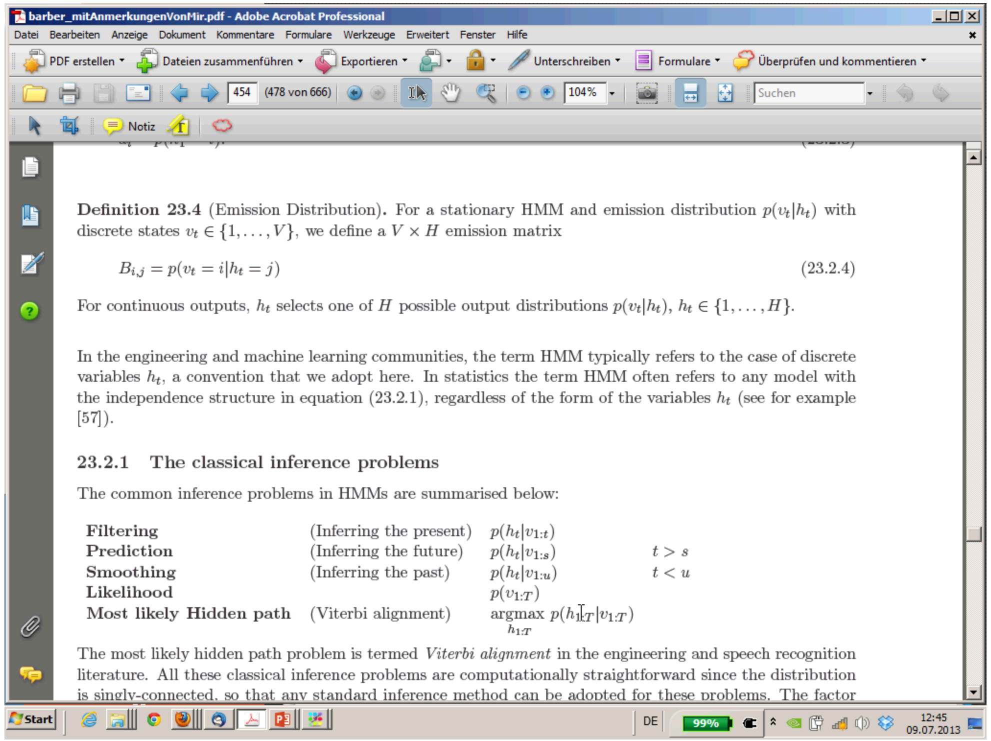
In the engineering and machine learning communities, the term HMM typically refers to the case of discrete variables h_t, a convention that we adopt here. In statistics the term HMM often refers to any model with the independence structure in equation (23.2.1), regardless of the form of the variables h_t (see for example [57]).

23.2.1 The classical inference problems

The common inference problems in HMMs are summarised below:

Filtering (Inferring the present) $p(h_t|v_{1:t})$
Prediction (Inferring the future) $p(h_t|v_{1:s})$ $t > s$
Smoothing (Inferring the past) $p(h_t|v_{1:u})$ $t < u$
Likelihood $p(v_{1:T})$
Most likely Hidden path (Viterbi alignment) $\underset{h_{1:T}}{\text{argmax}}\ p(h_{1:T}|v_{1:T})$

The most likely hidden path problem is termed *Viterbi alignment* in the engineering and speech recognition literature. All these classical inference problems are computationally straightforward since the distribution is singly-connected, so that any standard inference method can be adopted for these problems. The factor



Definition 23.4 (Emission Distribution). For a stationary HMM and emission distribution $p(v_t|h_t)$ with discrete states $v_t \in \{1, \ldots, V\}$, we define a $V \times H$ emission matrix

$$B_{i,j} = p(v_t = i|h_t = j) \qquad (23.2.4)$$

For continuous outputs, h_t selects one of H possible output distributions $p(v_t|h_t)$, $h_t \in \{1, \ldots, H\}$.

In the engineering and machine learning communities, the term HMM typically refers to the case of discrete variables h_t, a convention that we adopt here. In statistics the term HMM often refers to any model with the independence structure in equation (23.2.1), regardless of the form of the variables h_t (see for example [57]).

23.2.1 The classical inference problems

The common inference problems in HMMs are summarised below:

Filtering (Inferring the present) $p(h_t|v_{1:t})$
Prediction (Inferring the future) $p(h_t|v_{1:s})$ $t > s$
Smoothing (Inferring the past) $p(h_t|v_{1:u})$ $t < u$
Likelihood $p(v_{1:T})$
Most likely Hidden path (Viterbi alignment) $\underset{h_{1:T}}{\text{argmax}}\ p(h_{1:T}|v_{1:T})$

The most likely hidden path problem is termed *Viterbi alignment* in the engineering and speech recognition literature. All these classical inference problems are computationally straightforward since the distribution is singly-connected, so that any standard inference method can be adopted for these problems. The factor

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. via:
    - ++Hidden Markov Models (HMMs)
    - ++Conditional Random Fields (CRFs)
    - ++Dynamic Time Warping (DTW)

Detecting Social Signals: Gesture and Posture

- Posture: Mostly for surveillance and activity recognition; Studies aiming at social signal interpretation:
  - -- e-learning for children
  - --recognize affective states
  - --influence of culture on affective postures

Detecting Social Signals: Gaze and Face

- First: problems: (repetition)
  - --Face Detection
  - --Extract Features from faces; Gaze: analyze direction of eyes
  - --Analyze temporal sequences
- then: interpret instances as social signals / behavioral cues

Detecting Social Signals: Gaze and Face

- AU: smallest discernable temporal feature sequence: sequence of geometry or appearance features (modelled 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:

$$I(x+dx, y+dy, t+dt) = I(x,y,t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + O(d^2)$$

$$\longrightarrow \quad \frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \qquad \text{(optical flow equation)}$$

  - -- estimate algorithmic approximation e.g. with Lucas–Kanade method

Detecting Social Signals: Gaze and Face

- AU: smallest discernable temporal feature sequence: sequence of geometry or appearance features (modelled 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:

$$I(x+dx, y+dy, t+dt) = I(x,y,t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + O(d^2)$$

$$\longrightarrow \quad \frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \qquad \text{(optical flow equation)}$$

  - -- estimate algorithmic approximation e.g. with Lucas–Kanade method

Detecting Social Signals: Gaze and Face

- AU: smallest discernable temporal feature sequence: sequence of geometry or appearance features (modelled 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:

$$I(x+dx, y+dy, t+dt) = I(x,y,t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + O(d^2)$$

$$\longrightarrow \frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \quad \text{(optical flow equation)}$$

  - -- estimate algorithmic approximation e.g.
    with Lucas–Kanade method

Detecting Social Signals: Gaze and Face

- AU: smallest discernable temporal feature sequence: sequence of geometry or appearance features (modelled 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:

$$I(x+dx, y+dy, t+dt) = I(x,y,t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + O(d^2)$$

$$\longrightarrow \frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \quad \text{(optical flow equation)}$$

  - -- estimate algorithmic approximation e.g.
    with Lucas–Kanade method

Detecting Social Signals: Gaze and Face

- AU: smallest discernable temporal feature sequence: sequence of geometry or appearance features (modelled 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:

$$I(x+dx, y+dy, t+dt) = I(x,y,t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + O(d^2)$$

$$\longrightarrow \frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \quad \text{(optical flow equation)}$$

  - -- estimate algorithmic approximation e.g.
    with Lucas–Kanade method

Detecting Social Signals: Gaze and Face

- AU: smallest discernable temporal feature sequence: sequence of geometry or appearance features (modelled 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:

$$I(x+dx, y+dy, t+dt) = I(x,y,t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + O(d^2)$$

$$\longrightarrow \frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \quad \text{(optical flow equation)}$$

  - -- estimate algorithmic approximation e.g.
    with Lucas–Kanade method

Detecting Social Signals: Gaze and Face

- AU: smallest discernable temporal feature sequence: sequence of geometry or appearance features (modelled 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:

$$I(x+dx, y+dy, t+dt) = I(x,y,t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + O(d^2)$$

$$\longrightarrow \quad \frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0 \qquad \text{(optical flow equation)}$$

  -- estimate algorithmic approximation e.g. with Lucas–Kanade method

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)

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)

Social Signal Processing Chain

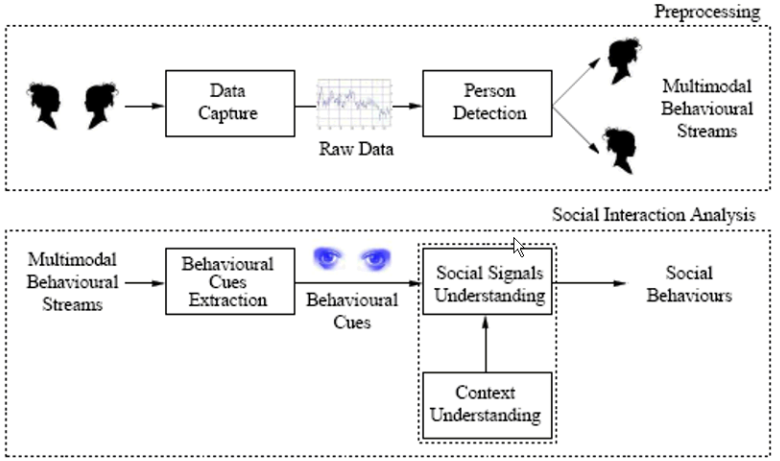


Fig. 6. Machine analysis of social signals and behaviours: a general scheme. The process includes two main stages: The *preprocessing*, takes as input the recordings of social interaction and gives as output the multimodal behavioural streams associated with each person. The *social interaction analysis* maps the multimodal behavioural streams into social signals and social behaviours.

[1]

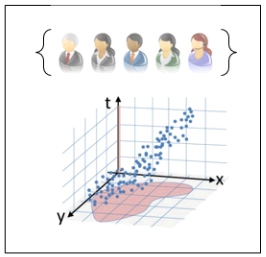
Social Situation Models as Models of Social Context

Social Situation:

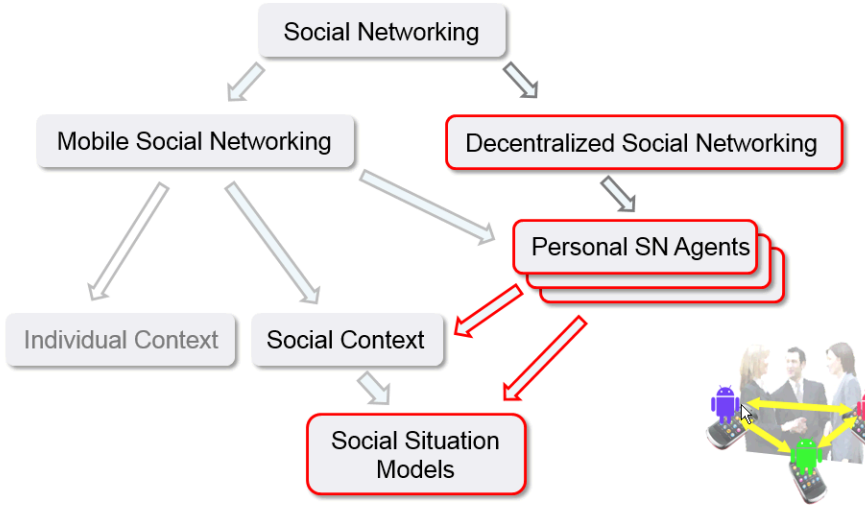
Co-located social interaction
with full mutual awareness

Simplified Social Situation Model:

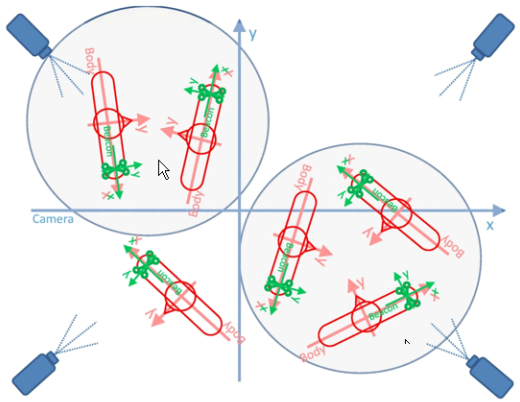
- Participating persons: P: set of IDs
- Spatio-temporal reference: X: sub-set of $\mathbb{R} \times \mathbb{R}^3$
- → S = (P, X)



Social Situation Models and Agents



Experiment



Results

Experiment data: Manual annotation

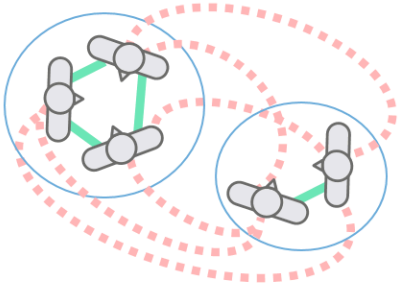
$|S^{\oplus}| = 321307$ $(\delta\theta, \delta d)$ pairs corresponding to „in a social situation“

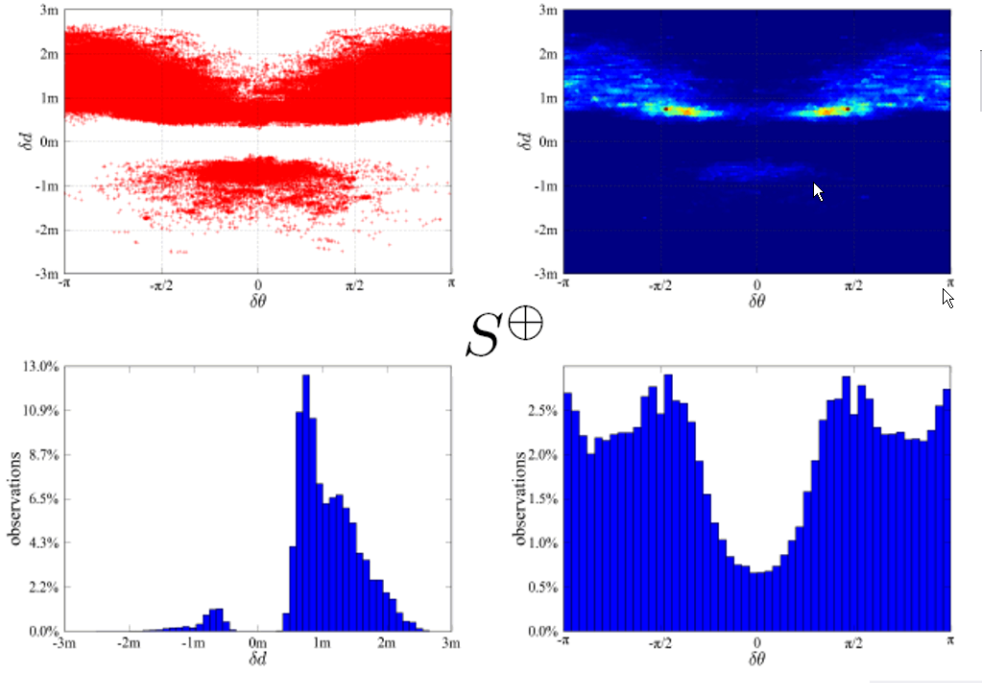
$|S^{\ominus}| = 398335$ $(\delta\theta, \delta d)$ pairs corresponding to „not in a social situation“

Example:

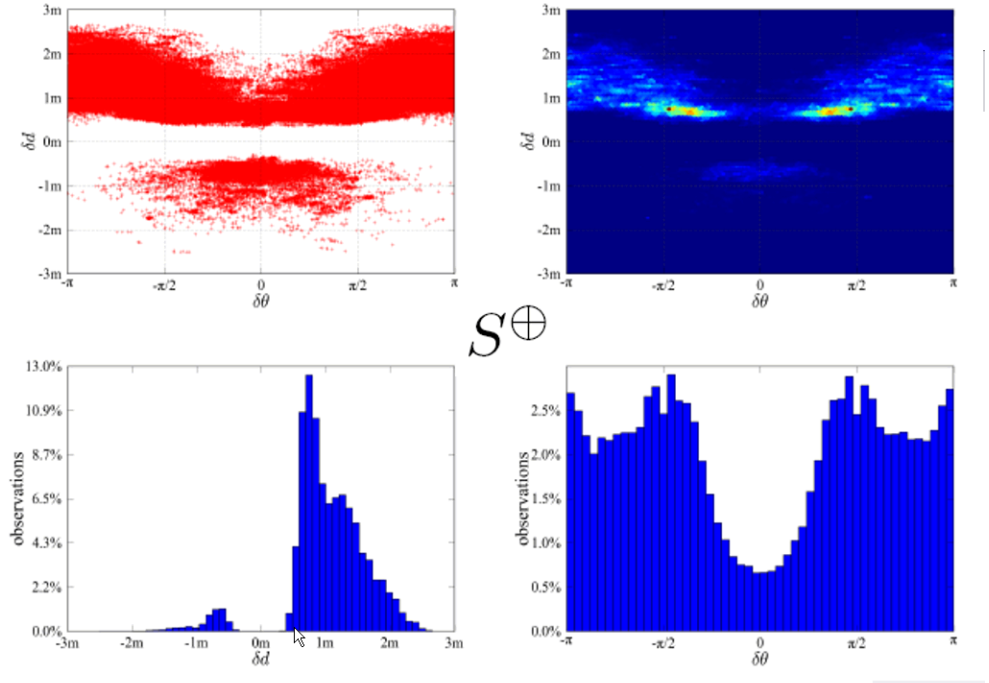
$|S^{\oplus}| = 4$

$|S^{\ominus}| = 6$

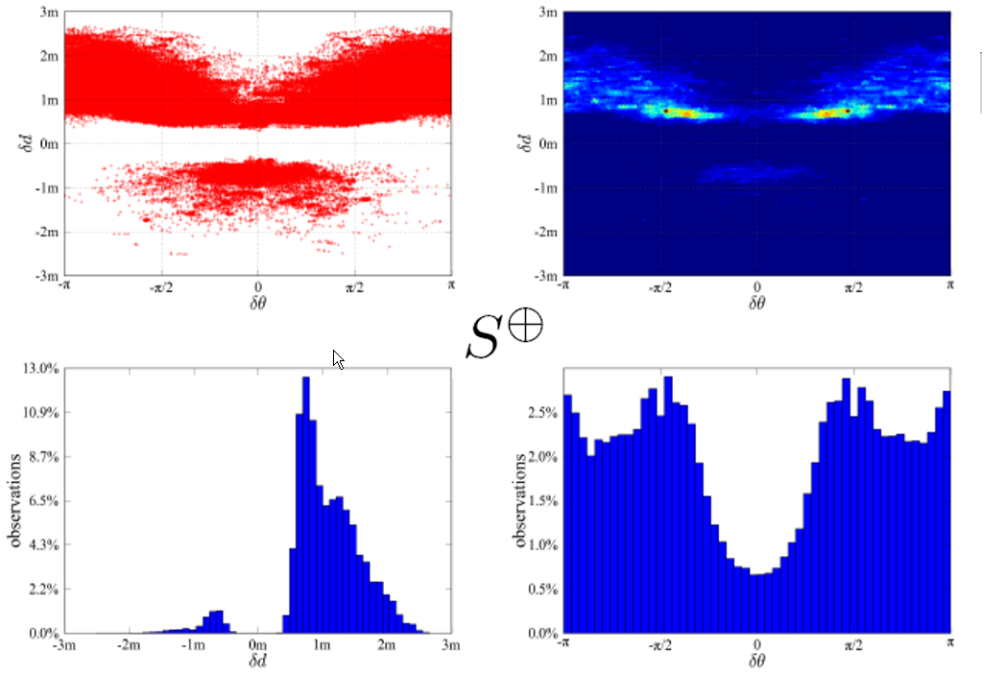




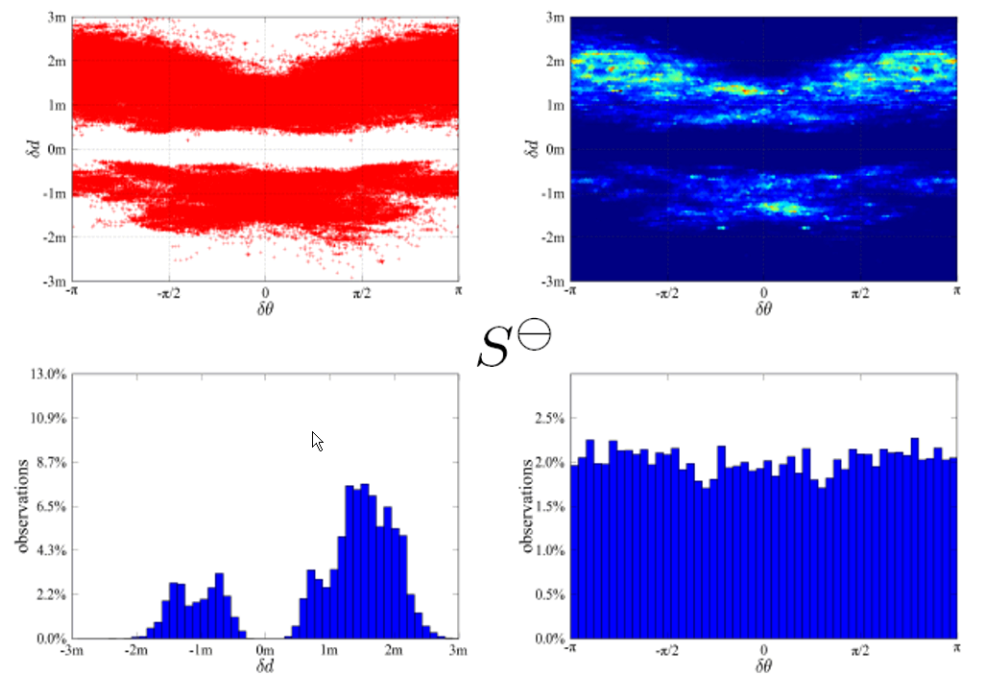
11 / 18



11 / 18



11 / 18



12 / 18

Classification Results

Classifier	Accuracy*
Gaussian Mixture Model (3 Gaussians)	74,34 %
Gaussian Mixture Model (5 Gaussians)	74,67 %
Gaussian Mixture Model (7 Gaussians)	74,59 %
Naive Bayes	65,45 %
Support Vector Machine (Polyn. Kernel)	77,81 %

(*) w. 10-fold cross validation