Event, Action & Activity Recognition - I

Rainer Stiefelhagen
Ziad Al-Halah (ziad.al-halah@kit.edu)

24.01.2014

Overview

- Introduction
  - Motivation
  - Events, Actions, Activities
- Example Approaches
  - HMM-based Human Motion Recognition
  - Layered HMMs for Activity Recognition
  - Human movement recognition using Temporal Templates
  - Boosting-based Action detection in movies
  - Action recognition using Bag-of-Words model

What is an event?

- *"a thing that happens or takes place"* ,
  - Oxford Dictionary

- Examples:
  - Gestures
  - Actions (running, drinking, standing up, etc.)
  - Activities (preparing a meal, playing a game, etc.)
  - Nature event (fire, storm, earthquake, etc.)
  - …

Introduction

- Why event & action recognition?
  - Gain a higher level understanding of the scene
  - Not just: Person locations, movement, orientation
  - Rather:
    - What are these persons doing (walking, sitting, working, hiding)?
    - How are they doing it?
    - What is going on in the scene (meeting, party, telephone conversation, etc…)?
  - Useful for video indexing/analysis, smart-rooms, patient monitoring, surveillance, robots etc.

What are human actions?

- Definition 1:
  - Physical body motion:



KTH action Database

(http://www.nada.kth.se/cvap/actions/)

What are human actions?

- Definition 2:
  - Interaction with environment on specific purpose




  - Same physical motion – different action depending on the context

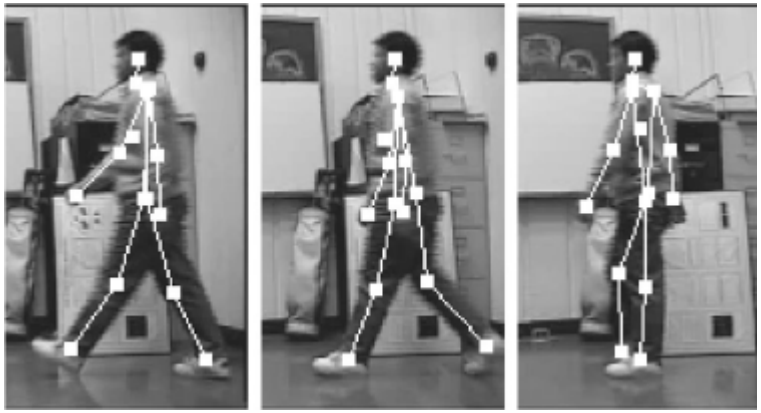
What are activities?

- *„… larger scale events that typically depend on the context of the environment, objects or interacting humans.”* [Moeslund06]
- Complex composition of actions

Types of Events

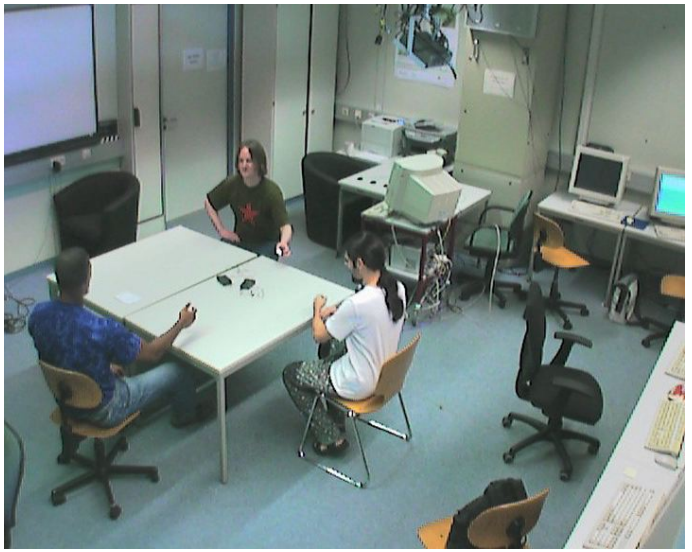
- Motion Primitives
  - Grab object
  - Move hand
  - Forehand
  - Backhand
- Actions
  - Pick up object
  - Jump
  - Walk
- Activities
  - Prepare a meal
  - Play baseball

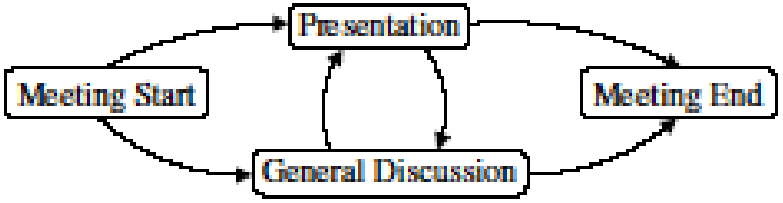
Complexity



Types of Events

- Small groups (meetings):
  - Individual actions
    - Speaking, writing, listening, walking, standing up, sitting down, “fidgeting”,…
  - Group activities
    - Meeting start, end, discussion, presentation, monologue, dialogue, white board, note-taking
  - Often audio-visual cues





Types of Events

- Rooms (office, kitchen):
  - Individual activities:
    - entering/leaving the room
    - working on the desk
    - making a phone call
    - making coffee
  - Activities, composed by individual activities (in one or more offices):
    - phone conference
    - meeting
    - short interrupt / discussion
    - fetching printouts from the printer in the lab
  - Here also, audio-visual cues, but coarser in nature.

Types of Events

- Outdoor Actions:
  - Mostly surveillance, for ex. In parking lots, in front of stores, in train stations:
    - Car enter, car leave, person enter, pickup, drop object (bomb?), hide, follow person, etc…
  - Recently became very popular field because of the "fight against terror"

Approaches

- Time series classification problem similar to speech/gesture recognition
  - Typical classifiers:
    - HMMs and variants (e.g. Coupled HMMs, Layered HMMs)
    - Dynamic Bayesian Networks (DBN)
- Classification problem similar to object recognition/detection
  - Typical classifiers:
    - Template matching
    - Boosting
    - Bag-of-Words SVMs

Example approaches

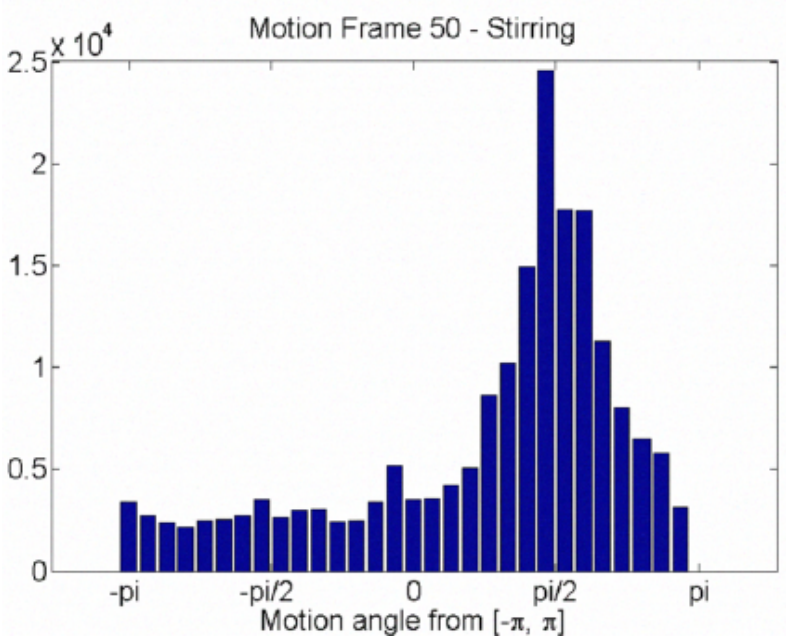
Left-to-right HMMs

HMM-based motion recognition [Gehrig09]

- Approach inspired by speech recognition systems
  - cf. pointing gesture and sign language recognition
- Motion primitives recognition
  - Calculate global motion features for each frame of a video sequence
  - HMM classification of feature streams
- Action/Activity recognition
  - Combine complex sequences of motion primitives using
    - Grammars
    - Statistical models

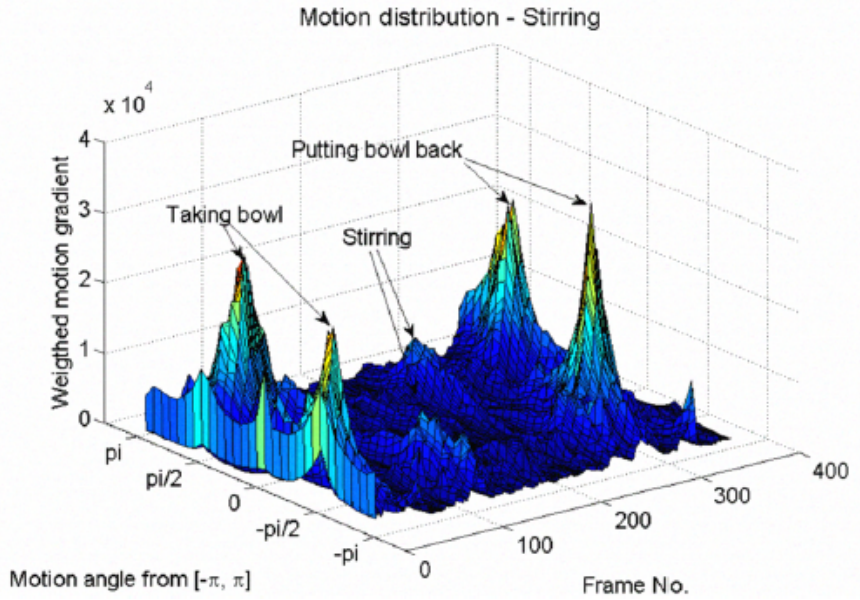
Motion Features

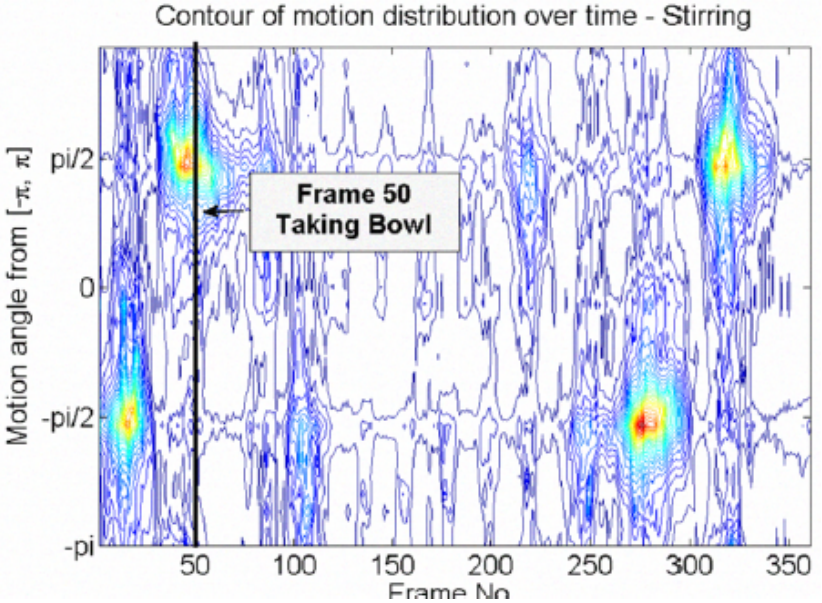
- Marker-based
  - Body-joint angles obtained from tracking system
- Video-based
  - Histogram of optical flow values for all angles



Motion primitives recognition

- One global feature per frame
- Continous left-to-right HMM classifier
  - 4 states per motion primitive
  - 16 Gaussians per state





Results


Marker-based Human Motion Recognition
Cognitive Systems Lab
Institute for Anthropomatics
2008

Results

- Results on 5 Motion sequences consisting of 26 different motion primitives (20 repetitions by one subject)
  - ☺ Low Word Error Rate
    - Marker-based system: 17.6 %
    - Video-based system: **13.1 %**
  - ☺ Real-time capability
  - ☺ Very fine recognition granularity

  - ☹ Problems with cyclic motions due to their short duration
  - ☹ Much annotated training data needed
  - ☹ Poor generalisation

Example approaches

Layered HMMs

Example 1: Layered Representations for Human Activity Recognition [Oliver02]

- Target:
  - Recognize complex human activities over longer period of time ("context" in an office setting).
- Types of context (situations, activities, etc):
  - Phone conversation
  - Face to face conversation
  - Presentation
  - Distant conversation
  - Others…
- Sensors:
  - Binaural microphones
  - USB camera
  - Keyboard and mouse

Hierarchical approach

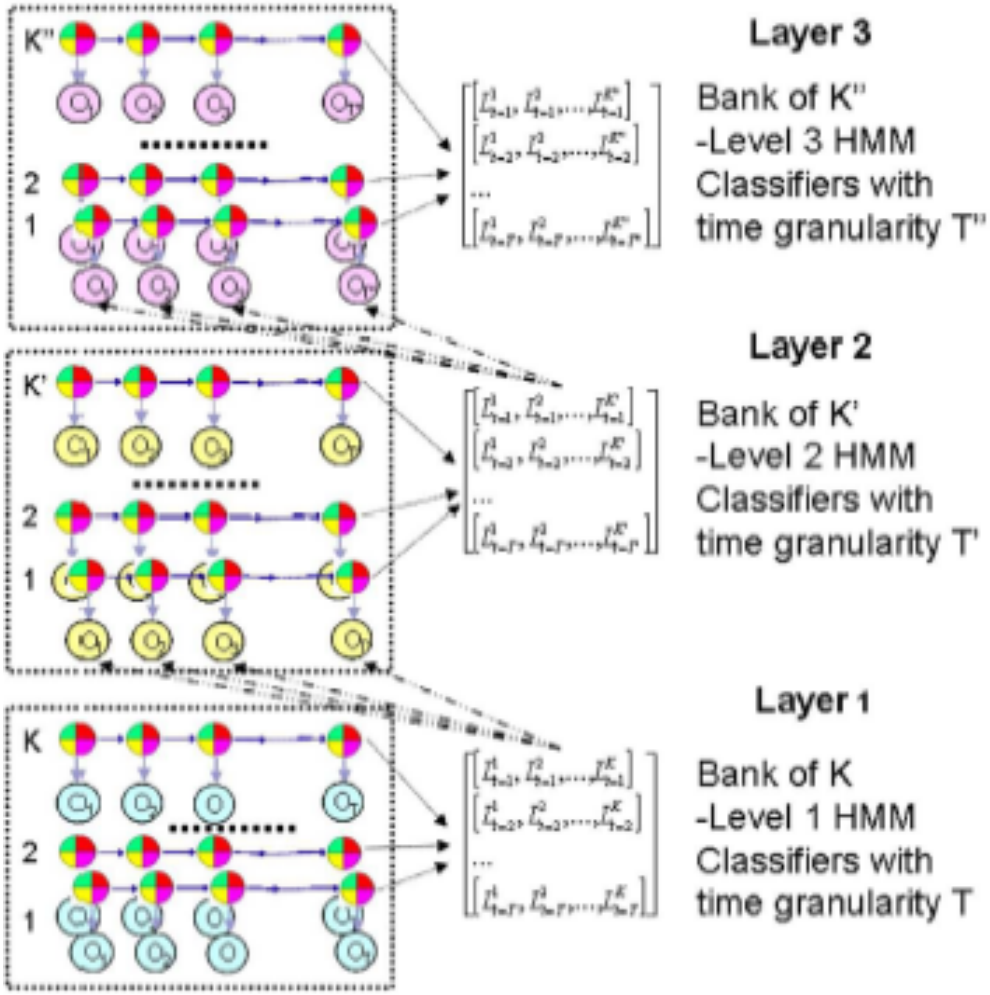
- Problem with normal HMMs:
  - Lack of structure
  - Large parameter space
  - Overfitting on long sequences with little training data
    - → Bad generalization
  - Fusion of various streams possible, but multiplies required parameters → need even more training data
- Solution:
  - Hierarchical (Layered) Hidden Markov Models (LHMMs)

Layered HMMs

Activities: Phone conversation, Face to face conversation, Presentation, Distant conversation

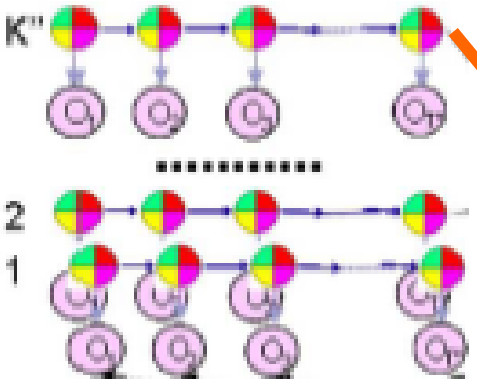
Output of lower layer is input to higher layer

Classes: Nobody present, one person, one active person, multiple people.
Music, silence, phone ring



Layered HMMs

- Maxbelief:

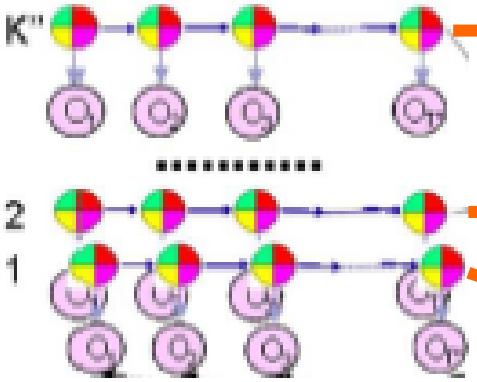


Winner(t)

$$A_t, A_{t-1}, A_{t-2}, \ldots, A_{t-T}$$

T discrete symbols in {1, …,K}

- Distributional :



$$\begin{bmatrix} L_t^K \\ \vdots \\ L_t^2 \\ L_t^1 \end{bmatrix}, \begin{bmatrix} L_{t-1}^K \\ \vdots \\ L_{t-1}^2 \\ L_{t-1}^1 \end{bmatrix}, \ldots, \begin{bmatrix} L_{t-T}^K \\ \vdots \\ L_{t-T}^2 \\ L_{t-T}^1 \end{bmatrix}$$

Vector of K likelihoods for each time step

Advantages

- Have smaller state space (parameters) than comparable conventional HMMs
  - Less prone to overfitting than HMMs
  - Need little training data at each level

- Lower level HMMs can be retrained separately

  → Adapt to new office settings

- More intuitive, structured representation
  - Encodes temporal structure of the activity modeling problem
  - Difficulty: Time granularity of each step defined manually (1sec, 5sec,…)

Visual features

- Skin color density (over whole image)

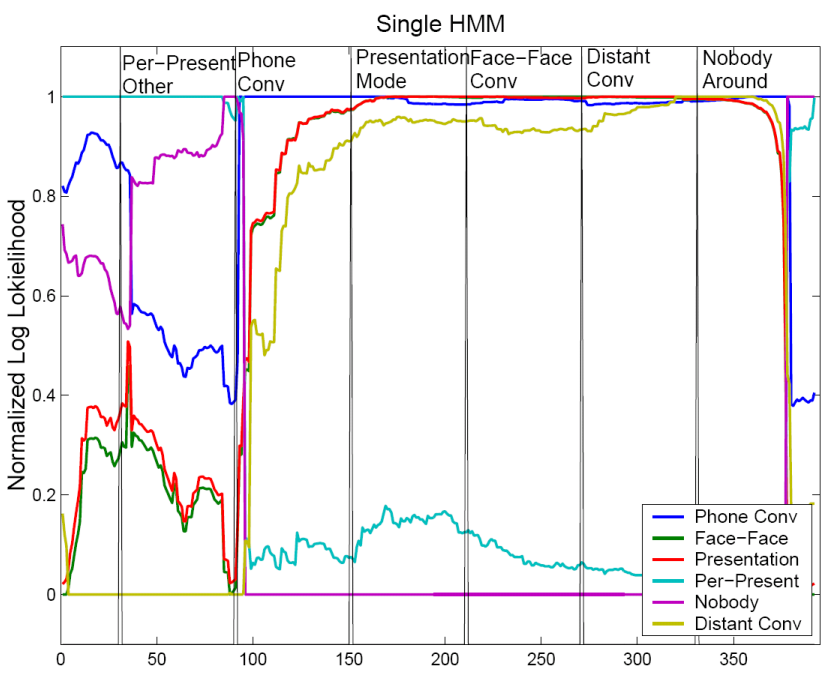
  (classification using skin/non skin color histograms in HSV space)
- Motion density

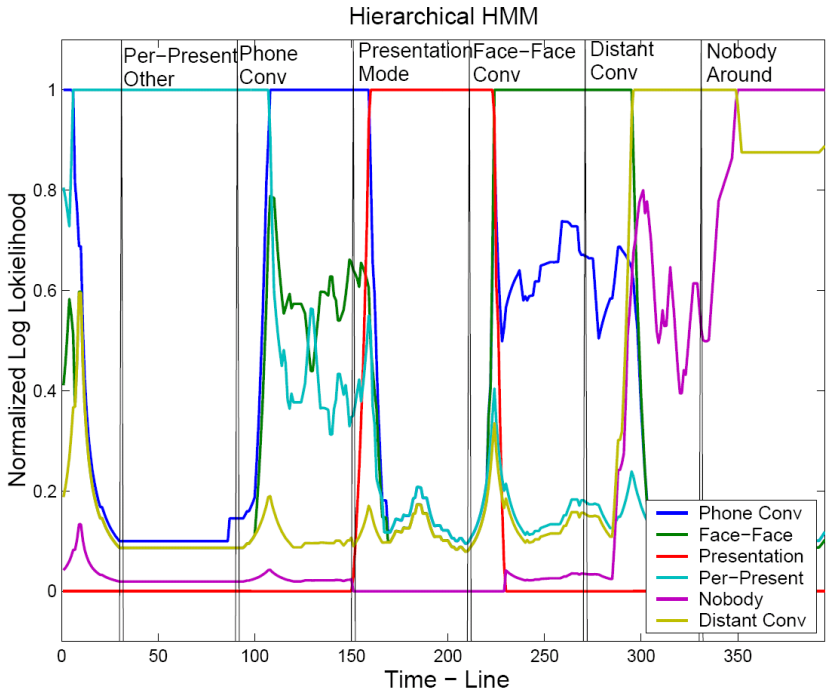
  (image differences)
- Foreground pixel density

  (background subtraction using learned background)
- Face pixel density

  (using real-time face detector)

Comparison HMM, L-HMM





Single HMM

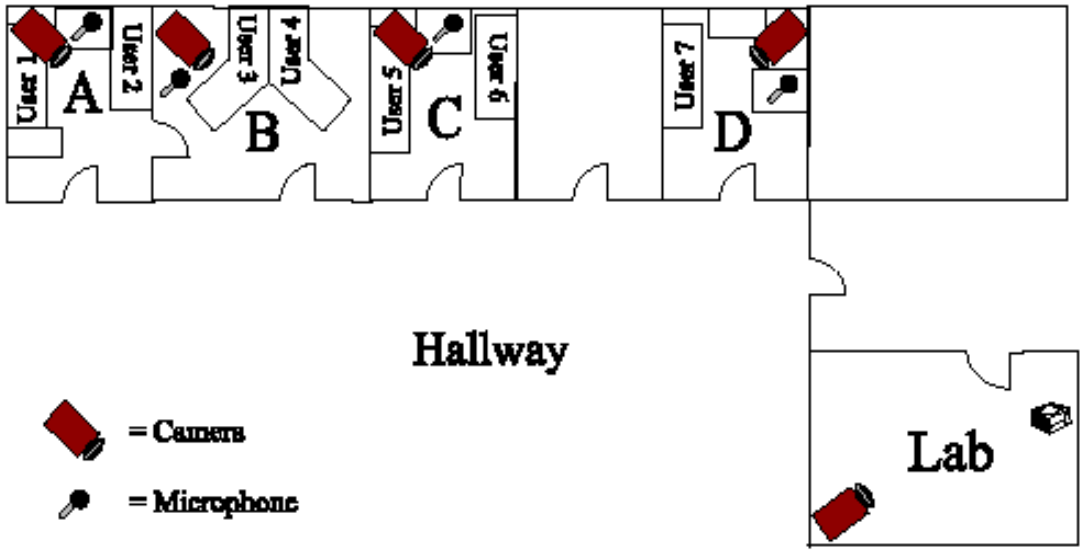
Hierarchical HMM
(Likelihoods are those of the highest level models)

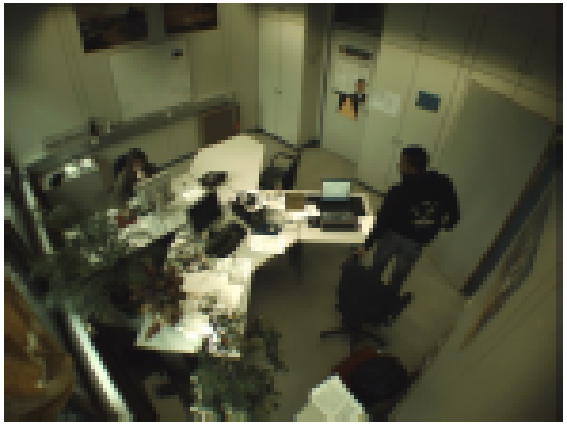
Illustration: per-frame normalized likelihoods of the models during real-time testing of different office activities

Example 2: Activity Recognition and Room-level Tracking in an Office Environment

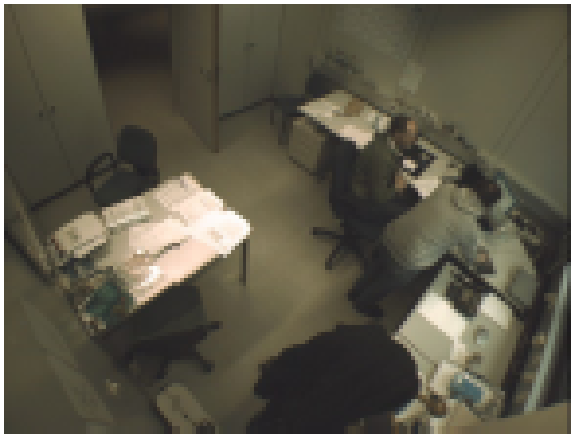
- Activity recognition allows to infer: [Wojek et al. 2006]
  - User‘s situation and availability
  - Interactions within groups
  - Can be used to produce a diary of each day
- Project goals
  - Detection of local events (e.g. somebody is entering a room, phone call, ...)
  - Fusion of those to detect global situations (e.g. meeting)
  - *(Track people‘s locations across offices)*
  - Use lightweight feature set and simple equipment that works under varying conditions

Floor Layout / Sensor Setup





Office B



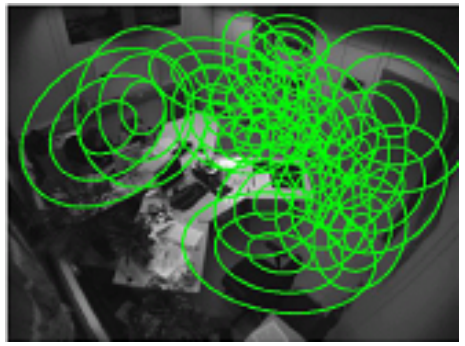
Office D

- Seven people in four offices (plus smart room)
- Sensors:
  - one camera per office
  - one omnidirectional microphone per office

Features

- Video features
  - Foreground
  - Optical flow
- Audio
  - Signal Energy
  - Zero Crossing Rate
  - Pitch
- Uses data driven local feature model
  - Foreground is modeled as GMM
  - Video features are calculated for each Gaussian
  - Data driven way to find meaningful areas
  - Reduces dimensionality !

Foreground



Learned FG model

Optical Flow

Foreground Detection

- Alpha-weighted difference images to detect foreground regions
  - Simple background model:
  - Pixels classified as foreground with distance > m to background:

$$fg_t[i] = \begin{cases} 0 & \text{if } |p_t[i] - bg_{t-1}[i]| \leq m \\ |p_t[i] - bg_{t-1}[i]| & \text{if } |p_t[i] - bg_{t-1}[i]| > m \end{cases}$$

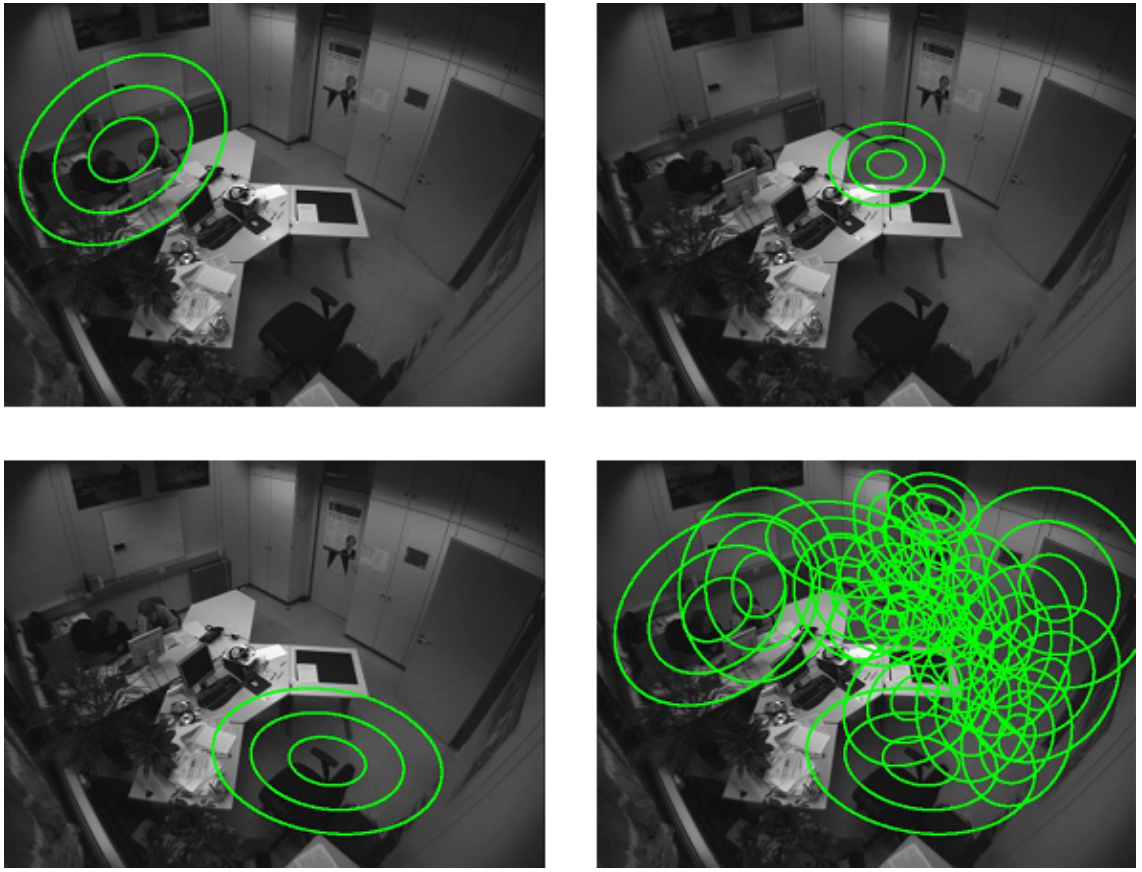
  - Adaptation speed set via alpha

$$bg_t[i] = \alpha * bg_{t-1}[i] + (1 - \alpha) * p_t[i]$$

  - Fast and robust

Example Foreground Segmentation

Activity Recognition – Local feature model



Resulting Gaussians for significant image areas and their first three standard deviations

Activity Recognition with Layered HMMs

- Idea
  - First layer consists of two groups of HMMs (Audio HMMs and Video HMMs) to detect *events*
  - Higher level HMMs are fed with the output probabilities of lower level HMMs in order to detect *situations*

- Feature vector structure on lowest level:
  - Video features:

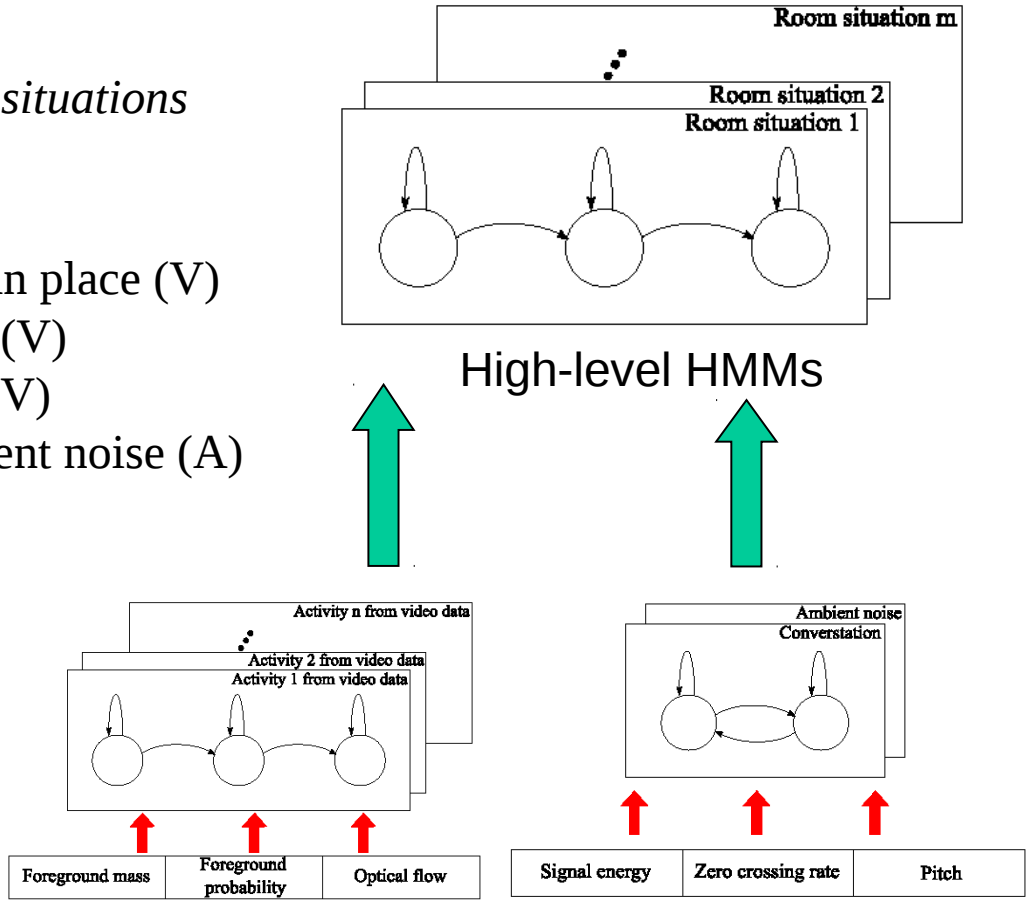
Foreground mass	Foreground probability from MoG model	Optical flow in X- and Y-direction

  - Audio features:

Mean and variance of signal energy	Mean and variance of zero crossing rate	Mean and variance of pitch

Multi-Layer HMMs

- Multi-Layered HMM Approach
  - First layer to detect *events*
  - Higher level HMMs to detect *situations*
- Examples for events
  - Somebody is sitting at a certain place (V)
  - Somebody is entering a room (V)
  - Somebody is leaving a room (V)
  - Somebody is talking vs. ambient noise (A)
- Examples for situations
  - Meeting with a visitor
  - Desk work
  - Discussion in an office
  - Nobody in office



Results (Office B, two persons)

- Training data: 4 full days
- Test data: 2 full days
- Both included:
  - day light
  - artificial light (evening)
  - cloudy skies
  - Sunny light
  - …

Results (Office B, two persons)

- First Level:

Description	*Recognition rate*	*False positive rate*	*Percentage of data*
Somebody at User 4's desk	92.2 %	4.0 %	75.3 %
Somebody at User 5's desk	98.4 %	1.8 %	65.6 %
Visitor behind User 4's desk	63.2 %	17.2 %	3.8 %
Visitor behind User 5's desk	78.1 %	13.9 %	2.4 %
Somebody around visitor's chair	98.3 %	11.5 %	1.2 %
Somebody enters	100.0 %	3.8 %	0.2 %
Somebody leaves	98.2 %	4.4 %	0.2 %
Somebody entering through side door	94.7 %	3.2 %	0.2 %
Somebody leaving through side door	91.0 %	2.5 %	0.3 %

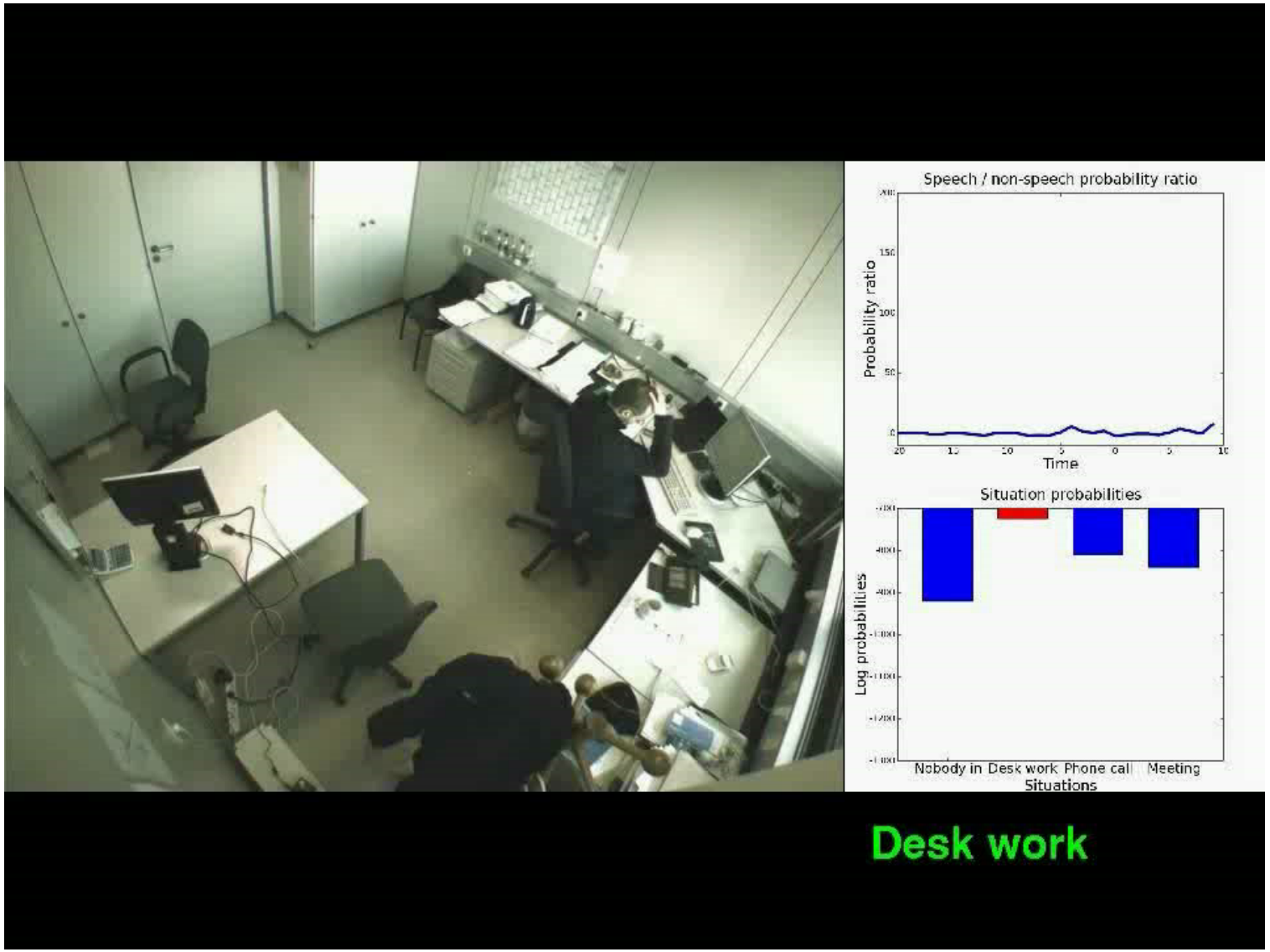
- Second Level:

Description	*Recognition rate*	*False positive rate*	*Percentage of data*
Nobody in the office	95.5%	0.5%	10.1%
Paperwork	90.7%	5.8%	62.4%
Discussion	73.9%	4.8%	18.4%
Meeting	69.6%	2.8%	9.1%

Confusion matrix (in seconds):

Description		[1]	[2]	[3]	[4]
Nobody in the office	[1]	**3462**	10	144	11
Paperwork	[2]	695	**20341**	723	663
Discussion	[3]	76	123	**4890**	1524
Meeting	[4]	0	793	203	**2278**

Video


Speech / non-speech probability ratio
Probability ratio
Time
Situation probabilities
Log probabilities
Nobody in
Desk work
Phone call
Meeting
Situations
Desk work

Summary [Wojek et al.]

- System allows
  - for detecting events and situations in several offices
  - for tracking colleagues on the floor *(not explained here, see paper)*
- Real-world data used
  - recorded seven days during working hours (tested on two days)
  - data includes all kinds of illumination (sunlight, cloudy sky, artificial illumination at night, etc.)
- Useful
  - to provide a semantic description of what is going on (and where)
    - for example as a diary
  - to determine availability of people

Summary

- *Event*: a thing that happens or takes place (contains actions)
- *Human action*: Physical body motion / Interaction with environment for a specific purpose
- *Activity:* temporal sequence of actions

Approaches (today):

- Left-to-right HMMs
  - Accurate recognition of motion primitives based on motion features
  - Video-based features seem to work better than model-based ones
- Layered HMMs
  - Deduce high-level activities from low-level events
  - Reduces state-space, amount of needed training data, helpful to model temporal granularities

References

T. Moeslund, A. Hilton, & V. Krüger. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 2006

D. Gehrig, D, H. Kuehne, A.Woerner & T. Schultz. HMM-based human motion recognition with optical flow data. Humanoids, 2009

N. Oliver, E. Horvitz, A. Garg. Layered Representations for Human Activity Recognition. Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI)

C. Wojek, K. Nickel, R. Stiefelhagen, Activity Recognition and Room Level Tracking in an Office Environment , IEEE Int. Conference on Multisensor Fusion and Integration for Intelligent Systems - MFI06, September 2006