Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases

Svebor Karaman, Jenny Benois-Pineau - LaBRI

Rémi Megret, Vladislavs Dovgalecs – IMS

Yann Gaëstel, Jean-Francois Dartigues - INSERM U.897

University of Bordeaux

SUPPORTED BY ANR

Human Daily Activities Indexing in Videos

1. The IMMED Project
2. Wearable videos
3. Automated analysis of activities
    1. Temporal segmentation
    2. Description space
    3. Activities recognition (HMM)
4. Results
5. Conclusions and perspectives

1. The IMMED Project

- IMMED: Indexing Multimedia Data from Wearable Sensors for diagnostics and treatment of Dementia.
  - http://immed.labri.fr → Demos: Video
- Ageing society:
  - Growing impact of age-related disorders
  - Dementia, Alzheimer disease…
- Early diagnosis:
  - Bring solutions to patients and relatives in time
  - Delay the loss of autonomy and placement into nursing homes
- The IMMED project is granted by ANR - ANR-09-BLAN-0165

1. The IMMED Project

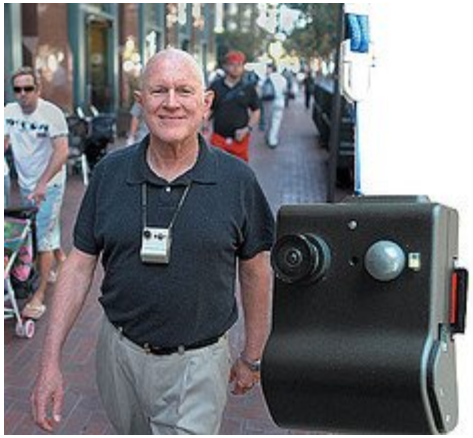
- Instrumental Activities of Daily Living (IADL)
  - Decline in IADL is correlated with future dementia

  PAQUID [Peres'2008]
- IADL analysis:
  - Survey for the patient and relatives → subjective answers
- IMMED Project:
  - Observations of IADL with the help of **video cameras** worn by the patient at home
- Objective observations of the evolution of disease
- Adjustment of the therapy for each patient

2. Wearable videos

- Related works:
- SenseCam
  - Images recorded as memory aid

  [Hodges et al.] “SenseCam: a Retrospective Memory Aid » UBICOMP’2006

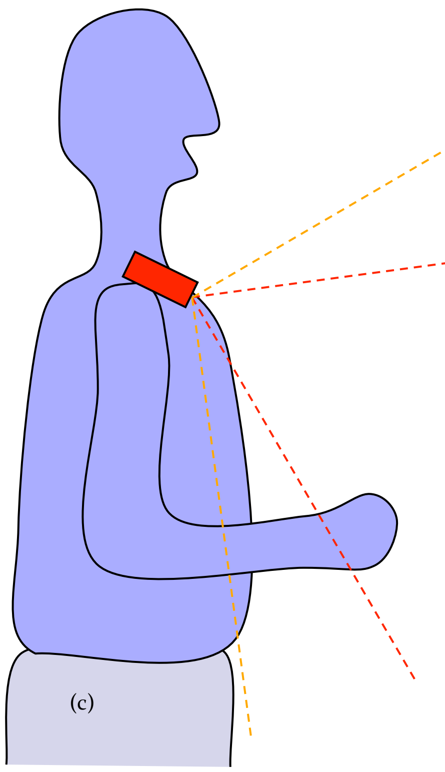


- WearCam
  - Camera strapped on the head of young children to help identifying possible deficiencies like for instance, autism

  [Picardi et al.] “WearCam: A Head Wireless Camera for Monitoring Gaze Attention and for the Diagnosis of Developmental Disorders in Young Children” International Symposium on Robot & Human Interactive Communication, 2007

2. Wearable videos

- Video acquisition setup



- Wide angle camera on shoulder
- Non intrusive and easy to use device
- IADL capture: from 40 minutes up to 2,5 hours

2. Wearable videos

- 4 examples of activities recorded with this camera: video
- Making the bed, Washing dishes, Sweeping, Hovering

3.1 Temporal Segmentation

- Pre-processing: preliminary step towards activities recognition
- Objectives:
  - Reduce the gap between the amount of data (frames) and the target number of detections (activities)
  - Associate one observation to one viewpoint
- Principle:
  - Use the global motion e.g. ego motion to segment the video in terms of viewpoints
  - One key-frame per segment: temporal center
  - Rough indexes for navigation throughout this long sequence shot
  - Automatic video summary of each new video footage

3.1 Temporal Segmentation

- Complete affine model of global motion (a1, a2, a3, a4, a5, a6)

$$\begin{pmatrix} dx_i \\ dy_i \end{pmatrix} = \begin{pmatrix} a_1 \\ a_4 \end{pmatrix} + \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix}$$

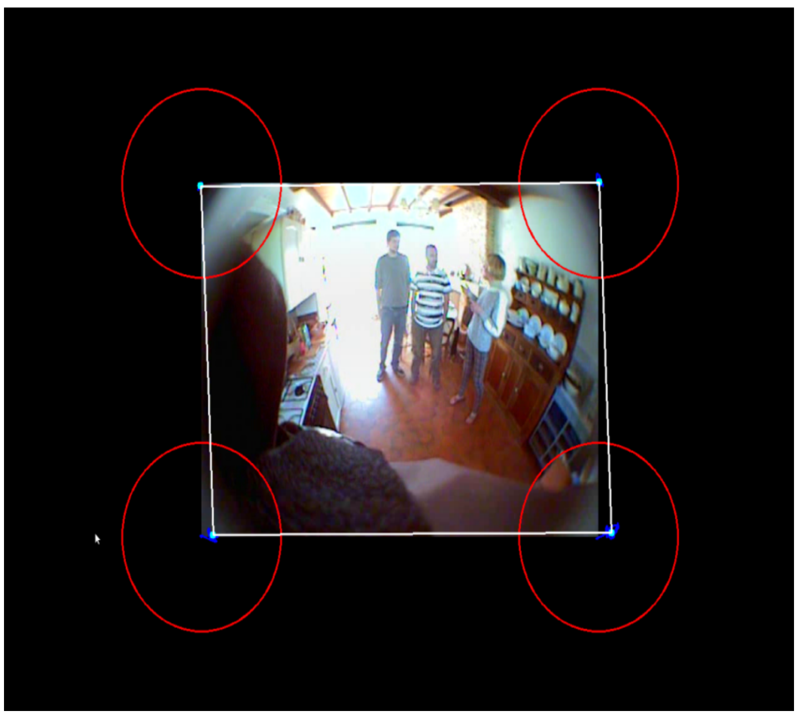
[Krämer et al.] Camera Motion Detection in the Rough Indexing Paradigm, TREC'2005.

- Principle:
  - Trajectories of corners from global motion model
  - End of segment when at least 3 corners trajectories have reached outbound positions

3.1 Temporal Segmentation

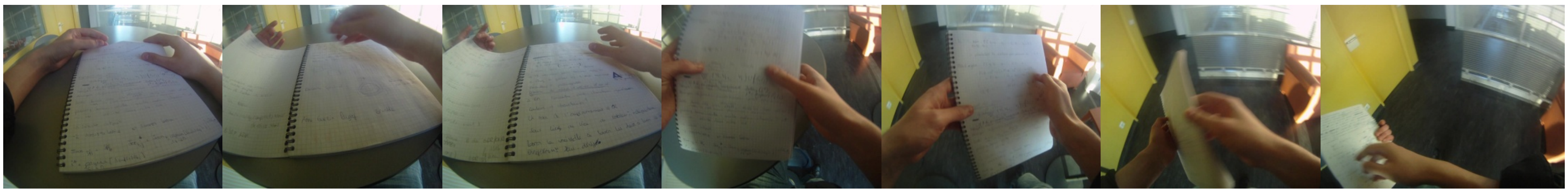
- Threshold *t* defined as a percentage *p* of image width *w*
  p=0.2 … 0.25

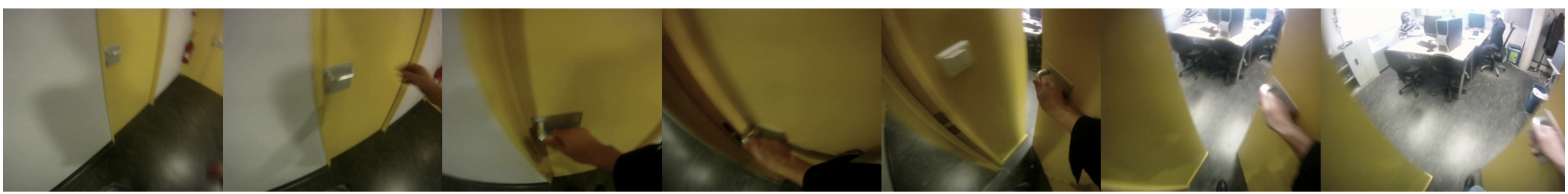
$$t = p \times w$$



3.1 Temporal Segmentation
Video Summary

- 332 key-frames, 17772 frames initially
- Video summary (6 fps)





3.2 Description space

- Color: MPEG-7 Color Layout Descriptor (CLD)

  6 coefficients for luminance, 3 for each chrominance
  - For a segment: CLD of the key-frame, $x(CLD) \in \Re^{12}$

- Localization: feature vector adaptable to individual home environment.
- N_{home} localizations. $x(Loc) \in \Re^{N_{home}}$
- Localization estimated for each frame
- For a segment: mean vector over the frames within the segment

V. Dovgalecs, R. Mégret, H. Wannous, Y. Berthoumieu. "Semi-Supervised Learning for Location Recognition from Wearable Video". CBMI'2010, France.

3.2 Description space

- H_{tpe} log-scale histogram of the translation parameters energy

  Characterizes the global motion strength and aims to distinguish activities with strong or low motion

- $N_e = 5$, $s_h = 0.2$. Feature vectors $x(H_{tpe}, a_1)$ and $x(H_{tpe}, a_4) \in \Re^5$

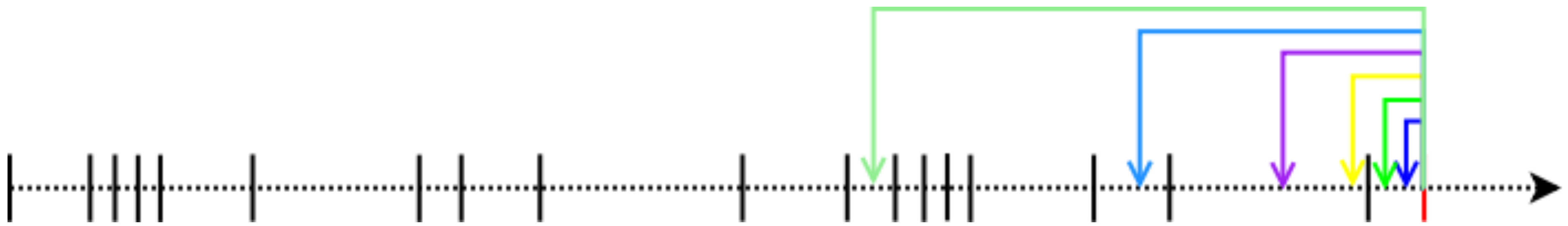
$$
\begin{array}{llll}
H_{tpe}[i] += 1 & if \quad \log(a^2) < i \times s_h & for & i = 1 \\
H_{tpe}[i] += 1 & if \quad (i-1) \times s_h \leq \log(a^2) < i \times s_h & for & i = 2..N_e - 1 \\
H_{tpe}[i] += 1 & if \quad \log(a^2) \geq i \times s_h & for & i = N_e
\end{array}
$$

- Histograms are averaged over all frames within the segment

	$x(H_{tpe}, a_1)$	$x(H_{tpe}, a_4)$
Low motion segment	0,87 0,03 0,02 0 0,08	0,93 0,01 0,01 0 0,05
Strong motion segment	0,05 0 0,01 0,11 0,83	0 0 0 0,06 0,94

3.2 Description space

- H_c: cut histogram. The i^{th} bin of the histogram contains the number of temporal segmentation cuts in the 2^i last frames



$H_c[1]=0$, $H_c[2]=0$, $H_c[3]=1$, $H_c[4]=1$, $H_c[5]=2$, $H_c[6]=7$

- Average histogram over all frames within the segment
- Characterizes the motion history, the strength of motion even outside the current segment

$2^6=64$ frames → 2s, $2^8=256$ frames → 8.5s

$x(H_c) \in \Re^6$ or $\Re^8$

3.2 Description space

- Feature vector fusion: early fusion
    - CLD $\rightarrow$ $x(CLD) \in \Re^{12}$
    - Motion
        - $x(H_{tpe}) \in \Re^{10}$
        - $x(H_c) \in \Re^{6}$ or $\Re^{8}$
    - Localization: N_{home} between 5 and 10.
        - $x(Loc) \in \Re^{Nhome}$
- Final feature vector size: between 33 and 40 if all descriptors are used
- Our example:
    - $x \in \Re^{33} = (x(CLD), x(H_{tpe},a_1), x(H_{tpe},a_4), x(H_c), x(Loc))$

3.3 Activities recognition

HMMs: efficient for classification with temporal causality

An activity is complex, it can hardly be modeled by one single state

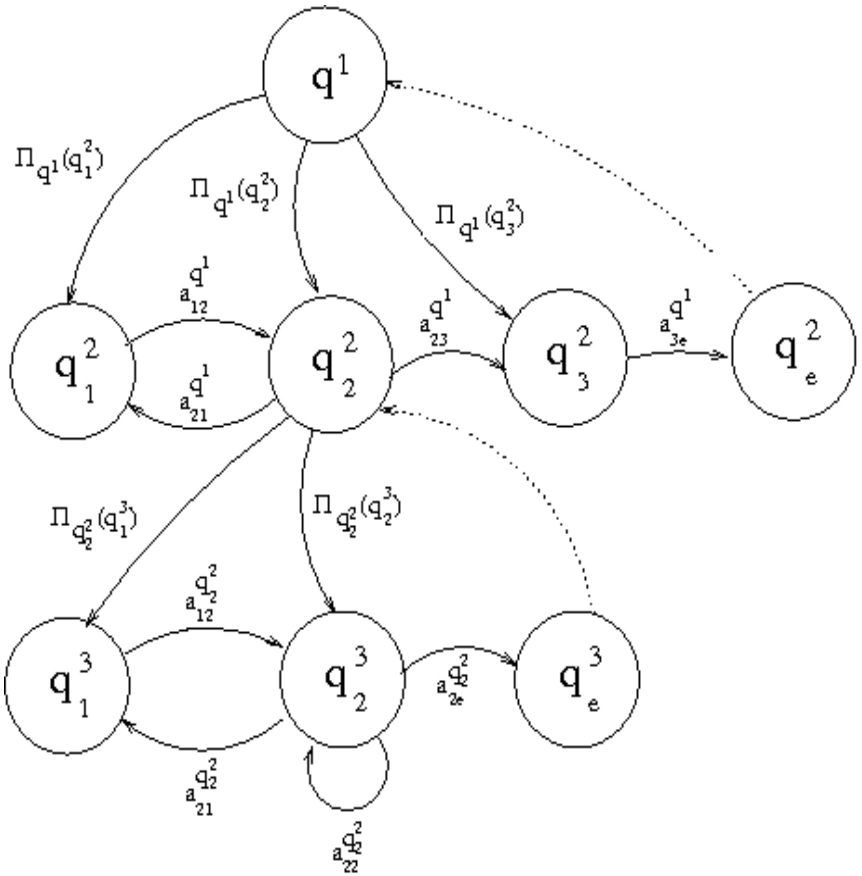
Hierarchical HMM? [Fine98], [Bui04]

- Multiple levels
- Computational cost/Learning

- $Q^D=\{q_i^d\}$ states set
- $\Pi_{q_i^d}(q_j^{d+1})$ = initial probability

of child q_j^{d+1} of state q_i^d

- $A_{ij}^{q^d}$ = transition probabilities between children of q^d



3.3 Activities recognition

A two level hierarchical HMM:

- Higher level:

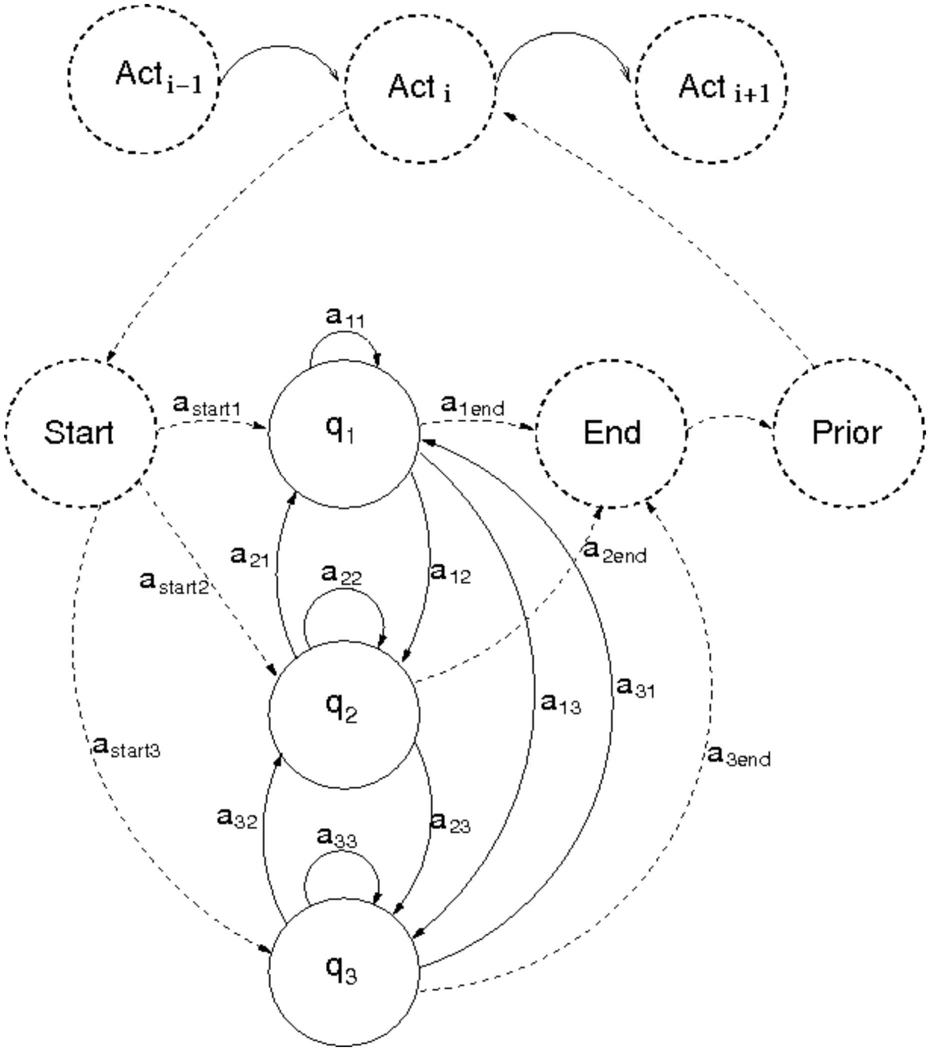
transition between activities

  - Example activities:
  Washing the dishes, Hovering, Making coffee, Making tea...

- Bottom level:

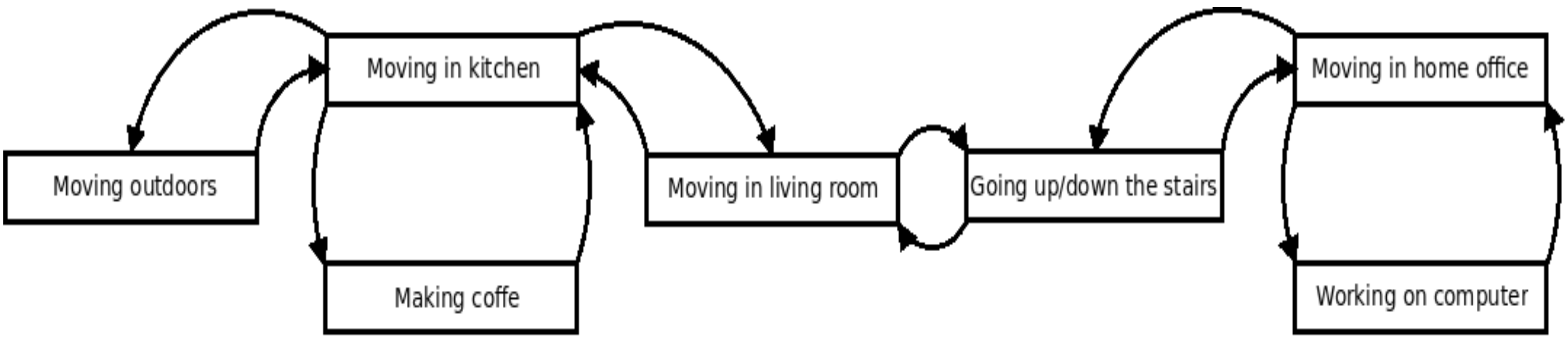
activity description

  - Activity: HMM with 3/5/7 states
  - Observations model: GMM
  - Prior probability of activity



3.3 Activities recognition

- Higher level HMM
  - Connectivity of HMM is defined by personal environment constraints



  - Transitions between activities can be penalized according to an a priori knowledge of most frequent transitions
  - No re-learning of transitions probabilities at this level

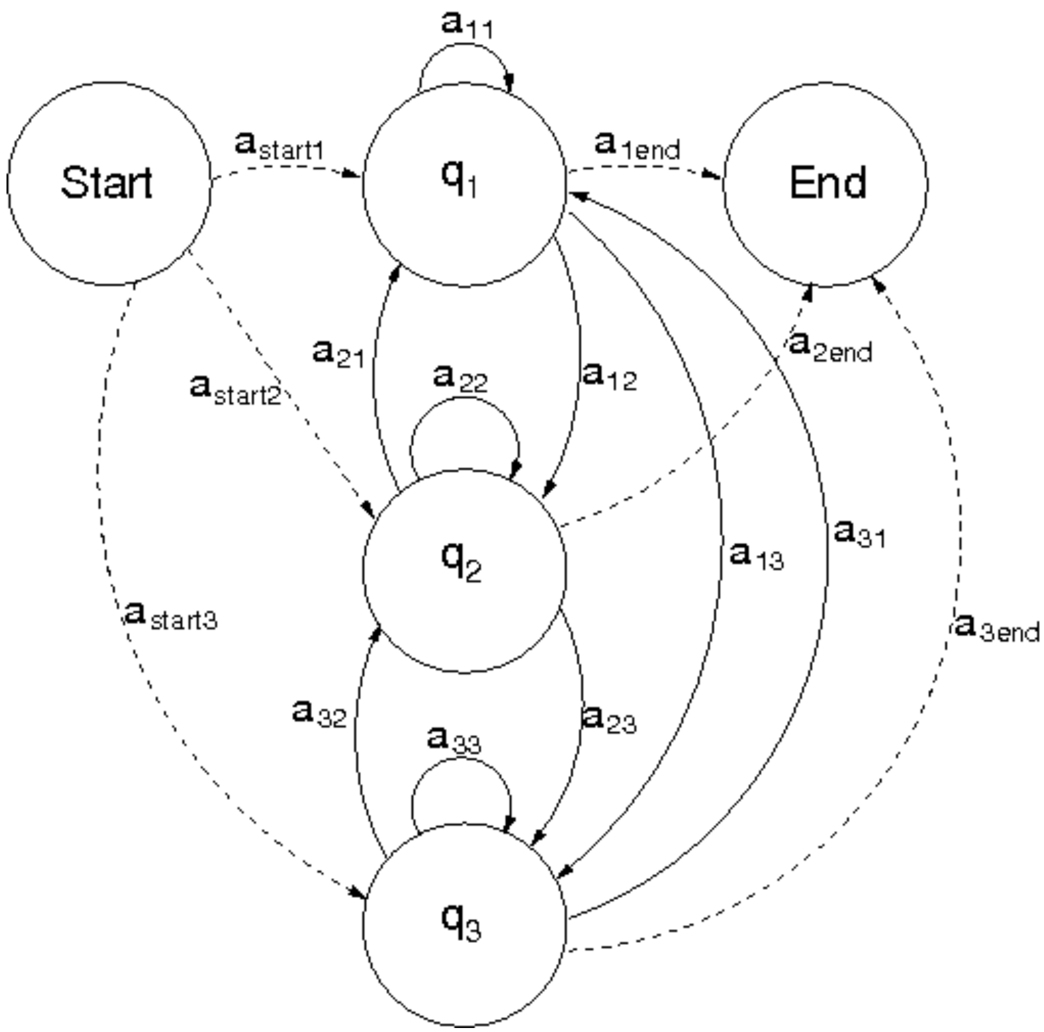
3.3 Activities recognition

Bottom level HMM

- Start/End

→ Non emitting state

- Observation x only for emitting states q_i
- Transitions probabilities and GMM parameters are learnt by Baum-Welsh algorithm
- A priori fixed number of states
- HMM initialization:
  - Strong loop probability a_{ii}
  - Weak out probability a_{iend}



4. Results

- No database available. One video. Total: 47489 frames.
- Learning on 10% of frames for each activity: 3974 frames. Recognition over 310 segments
- Tests: number of states of the HMM and space description changed. Prior probabilities were set equal.
- Best results:

Configuration	Nb States	F-Score	Recall	Precision
H_c + Localization	5	**0.64**	0.66	**0.67**
H_c + CLD + Localization	3	0.62	**0.7**	0.66

4. Results

- 7 activities:

  Moving in home office, Moving in kitchen, Going up/down the stairs, Moving outdoors, Moving in living room, Making coffee, Working on computer

- Confusion between Moving in home office and Going up/down the stairs (1 and 3)

  → proximity

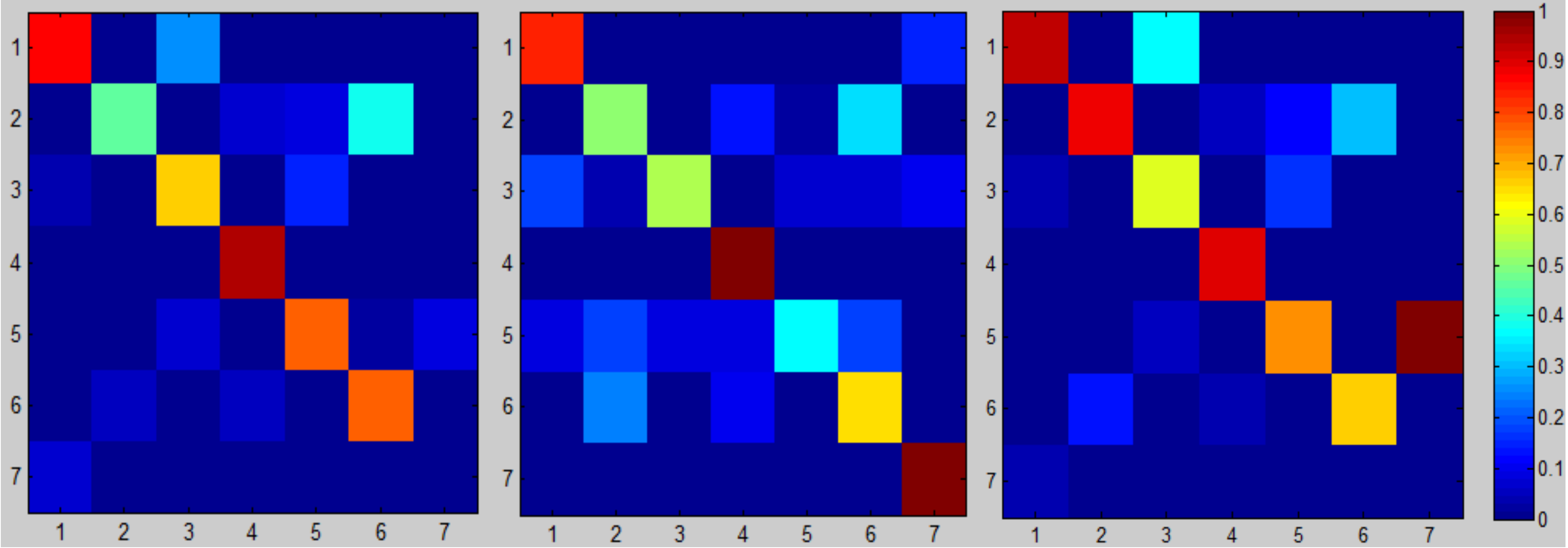
- Confusion between Moving in kitchen and Making coffee (2 and 6)

  → same localization/environment

4. Results

- 7 activities: Moving in home office, Moving in kitchen, Going up/ down the stairs, Moving outdoors, Moving in living room, Making coffee, Working on computer

Confusion matrixes:



F-Score Recall Precision

5. Conclusions and perspectives

- Human Activities Indexing and Motion Based Temporal Segmentation methods have been presented
- Encouraging results
- Difficulty to obtain videos (no such database available) and cost of annotation
- Tests on a larger corpus: 6h of videos available (work in progress)
- Audio integration (work in progress)
- Mid-level and local descriptors
    - Hand detection/tracking
    - Object detection
    - Local motion analysis

Thank you for your attention.

Questions?