Activities of Daily Living Indexing by Hierarchical HMM for Dementia Diagnostics

Svebor Karaman, Jenny Benois-Pineau – LaBRI,

Rémi Mégret – IMS,

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

University of Bordeaux

Julien Pinquier – IRIT, University of Toulouse

SUPPORTED BY ANR

Activities of Daily Living Indexing

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

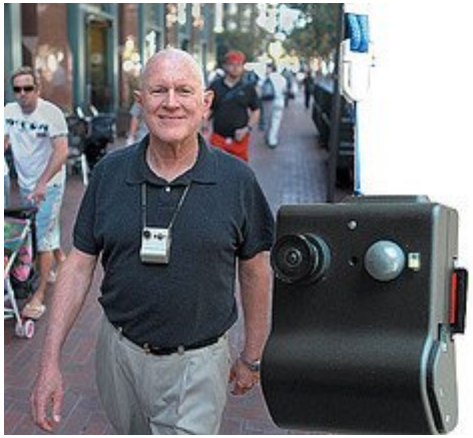
- Instrumental Activities of Daily Living (IADL)
  - Decline in IADL is correlated with future dementia

  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
  - Recording by paramedical staff when visiting the patient
- 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

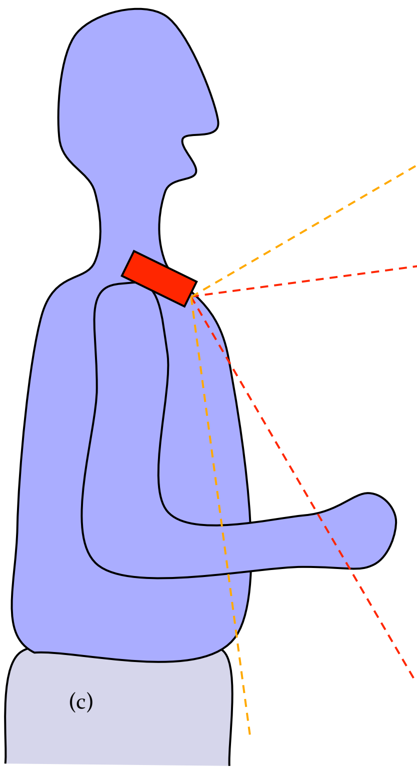


- WearCam
    - Camera strapped on the head of young children to help identifying possible deficiencies like for instance, autism

    [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

Contributions

- Framework introduced in *Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases*, ICPR'2010.
- In present work, definition of a cross-media feature space: motion, visual and audio features
- Learning of optimal parameter for temporal segmentation
- Experiments to find the optimal feature space
- Experiments on new real-world data

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)

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

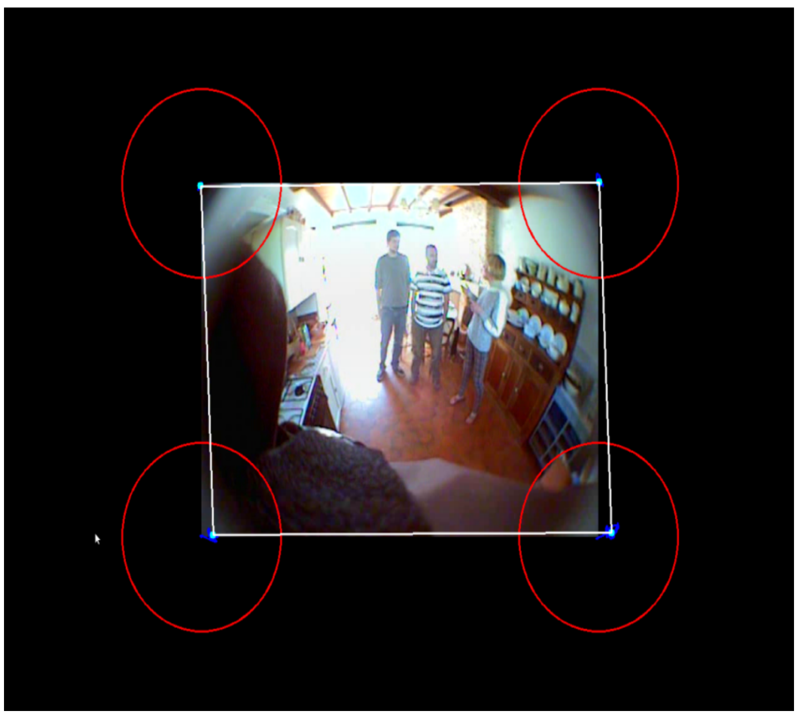
[Krämer et al.] Camera Motion Detection in the Rough Indexing Paradigm, TREC'2005.

- 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

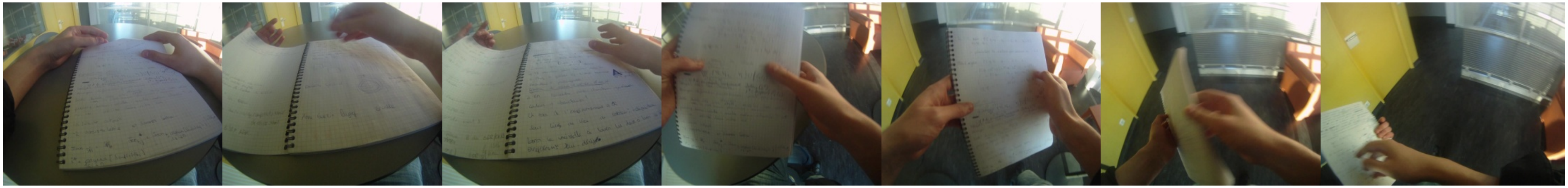
- Threshold *t* defined as a percentage *p* of image width *w*
  p=0.2 … 0.5

$$t = p \times w$$



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, $x(CLD) \in \Re^{12}$
- Audio (J. Pinquier and R. André-Obrecht, IRIT)
  - 5 audio classes: speech, music, noise, silence and percussion and periodic sounds
  - 4Hz energy modulation and entropy modulation for speech
  - Number of segments and segment duration from Forward-Backward divergence algorithm for music
  - Energy for silence detection
  - Spectral coefficients for percussion and periodic sounds

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 $x(H_{tpe}, a_4) \in \Re^5$

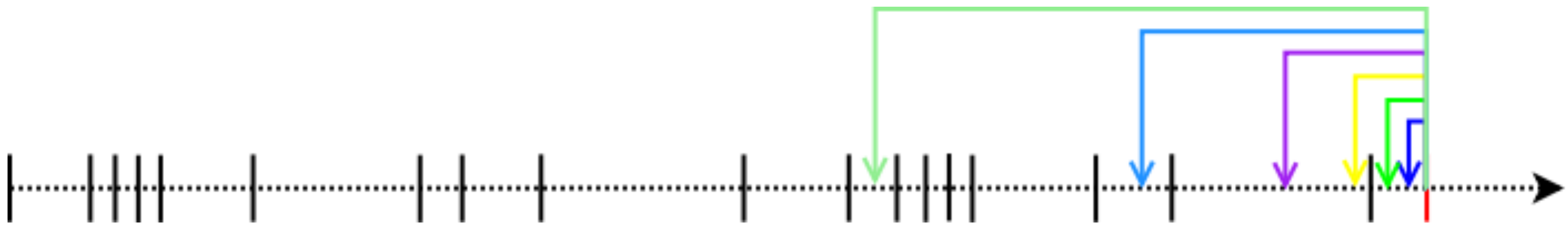
$$
\begin{array}{llll}
H_{tpe}[i] += 1 & if \quad \log(a^2) < i \times s_h & for & i = 1 \\
H_{tpe}[i] += 1 & if \quad (i-1) \times s_h \leq \log(a^2) < i \times s_h & for & i = 2..N_e - 1 \\
H_{tpe}[i] += 1 & if \quad \log(a^2) \geq i \times s_h & for & i = N_e
\end{array}
$$

- 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

$2^8=256$ frames $\rightarrow$ 8.5s

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

3.2 Description space

- Feature vector fusion: early fusion
  - CLD $\rightarrow$ $x(CLD) \in \Re^{12}$
  - Motion
    - $x(H_{tpe}) \in \Re^{10}$
    - $x(H_c) \in \Re^{8}$
  - Audio
    - $x(Audio) \in \Re^{5}$
- Final feature vector size: 35 if all descriptors are used

$$x \in \Re^{35} = (\ x(CLD),\ x(H_{tpe},a_1),\ x(H_{tpe},a_4),\ x(H_c),\ x(Audio)\ )$$

3.3 Activities recognition

A two level hierarchical HMM:

- Higher level:

transition between activities

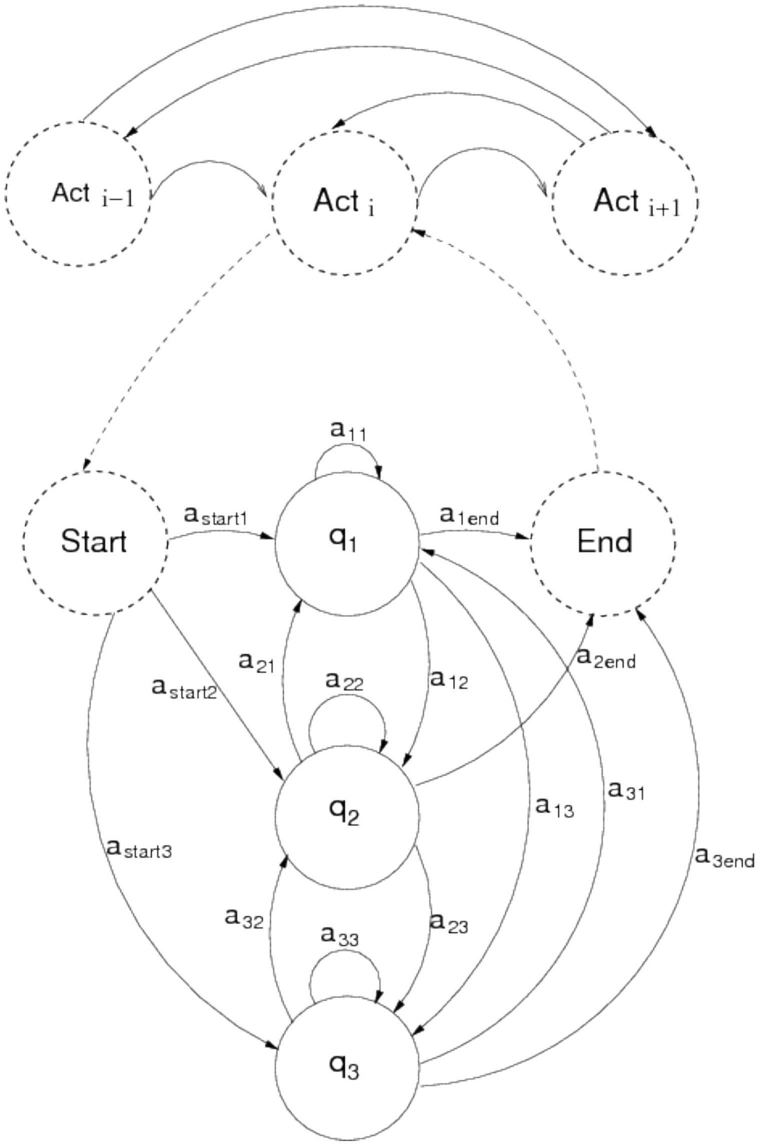
  - Example activities:

  Washing the dishes, Hovering, Making coffee, Making tea...

- Bottom level:

activity description

  - Activity: HMM with 3/5/7 states
  - Observations model: GMM



3.3 Activities recognition

- Higher level HMM
  - Connectivity of HMM can be 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
  - In this study, the activities are:
    - “making coffee”, “making tea”, “washing the dishes”, “discussing”, “reading”
    - and a reject class for all other not relevant events “NR”

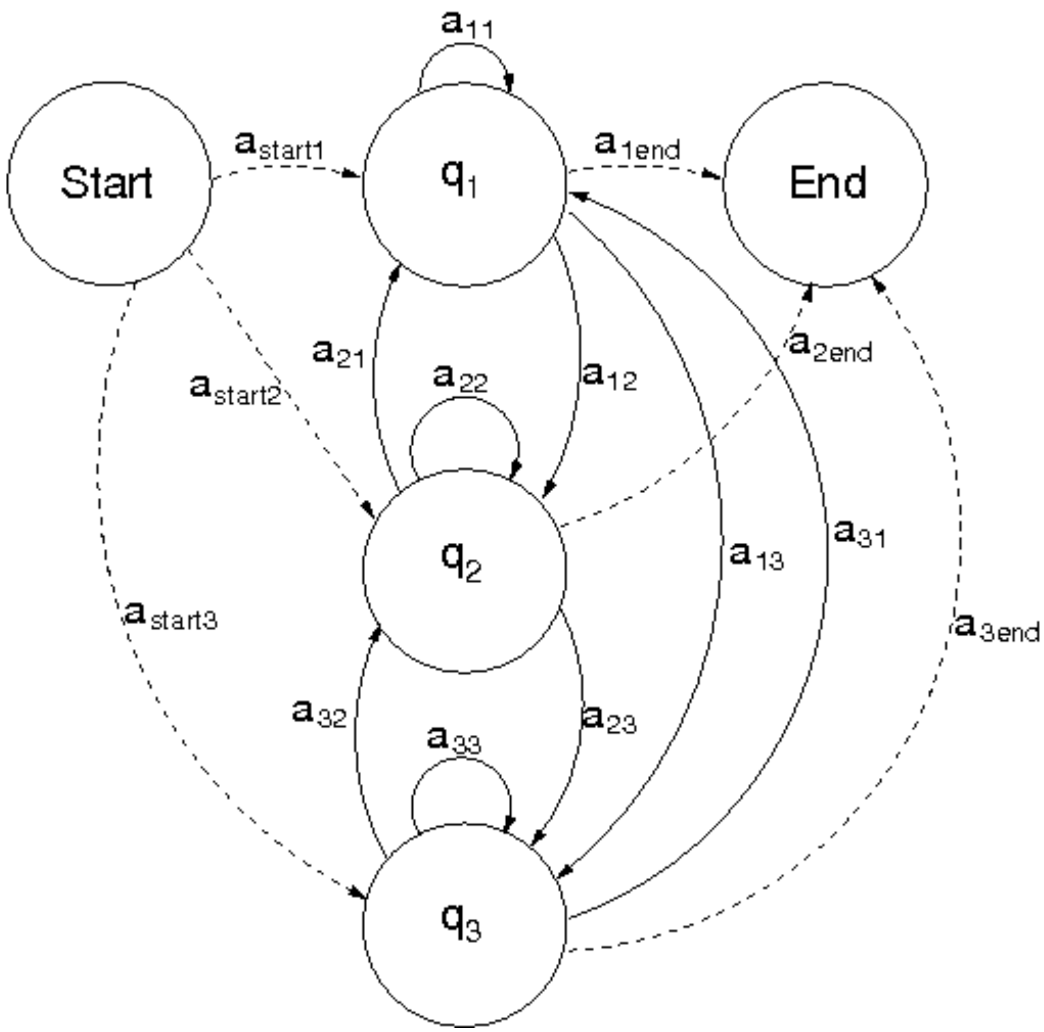
3.3 Activities recognition

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}

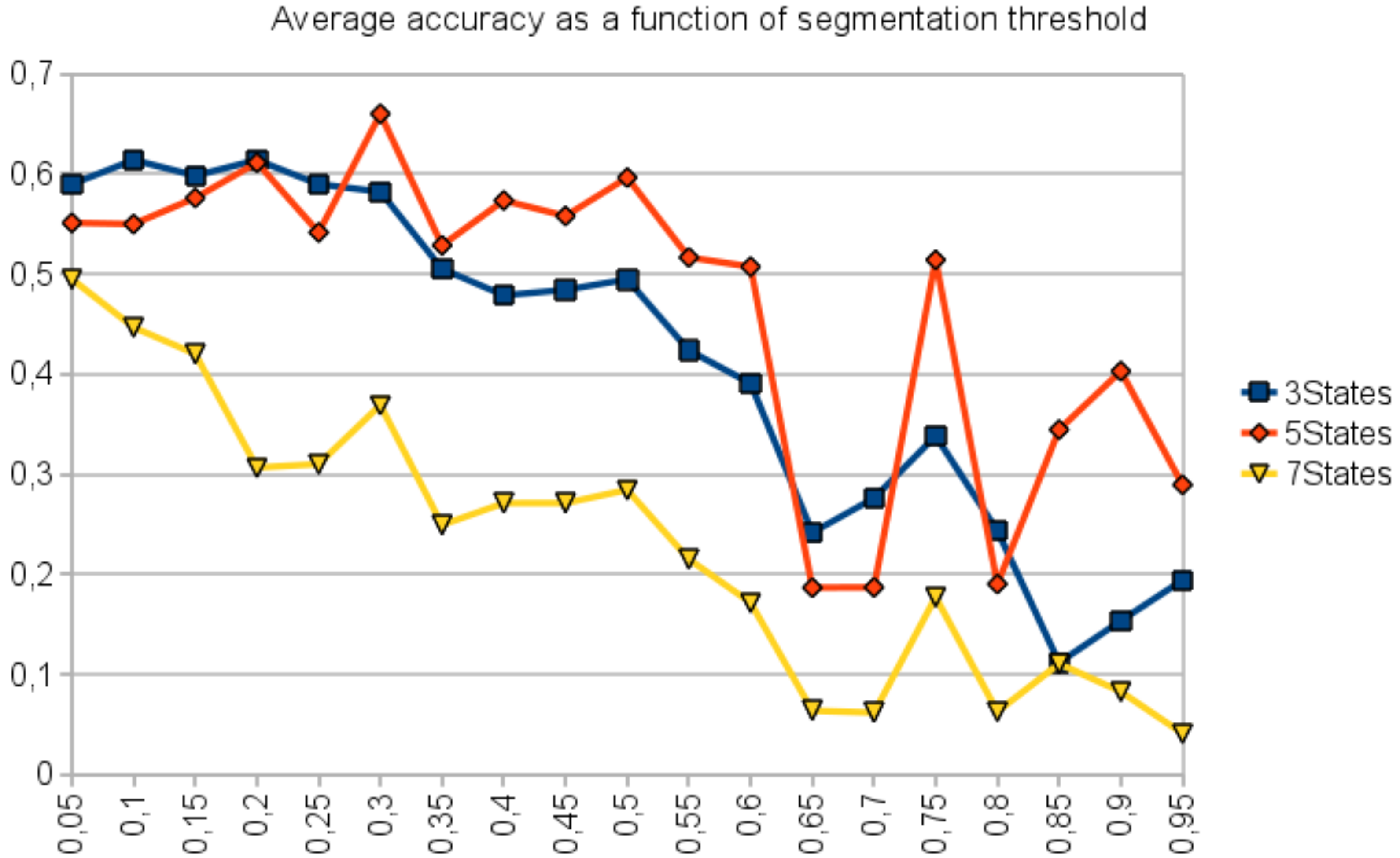


4. Results

- No public database available.
- In this experiments, videos are recorded at the LaBRI:
  - 3 volunteers carrying out some of the activities “making coffee”, “making tea”, “washing the dishes”, “discussing”, “reading”. Not all activities are present in a video
- 6 videos, 81435 frames, 45 minutes
- Cross validation: learning on all videos but one, remaining one for testing purpose
- Parameters studied:
  - Temporal segmentation threshold
  - Number of states in the activity HMM
  - Description space

4. Results

- Segmentation threshold influence when varying number of states in HMM



4. Results

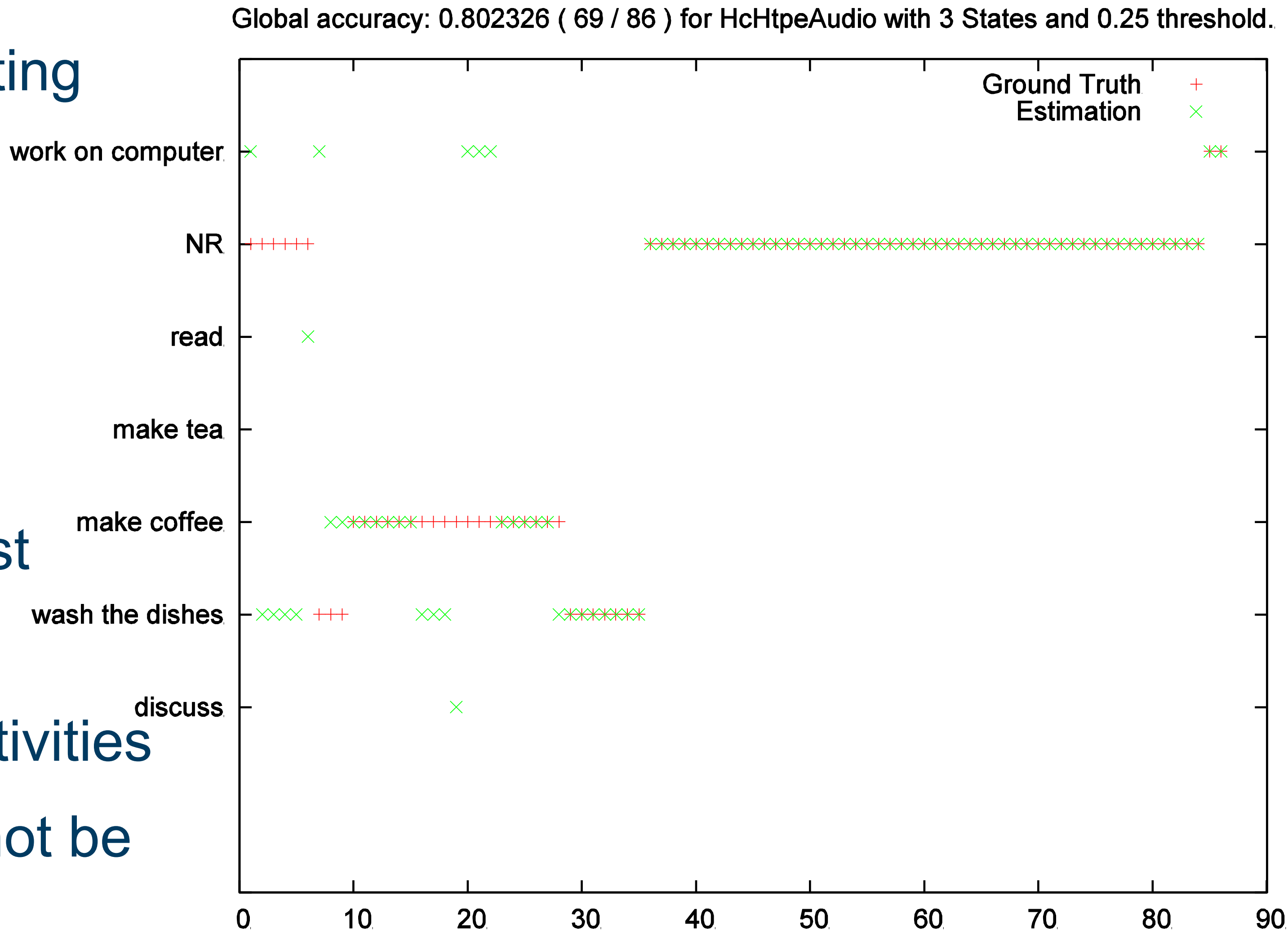
- Selection of best results after cross-validation:

Description Space	Number of States	Threshold	Accuracy
H_{tpe} Audio	3	0.35	0.75
H_{tpe} CLD	5	0.35	0.75
H_{tpe} CLD Audio	3	0.40	0.74
H_c CLD Audio	7	0.25	0.73
H_c H_{tpe} CLD Audio	3	0.15	0.73

- Top 10:
  - Descriptors: 7 HtpeAudio, 2 HtpeCLD, 1 HtpeCLDAudio
  - States: 3 “3StatesHMM”, 5 “5StatesHMM”, 2 “7StatesHMM”
  - Threshold: Between 0.2 and 0.5

4. Results

- NR/Interest: Max: 0.85
- Most interesting events are detected
- Some confusion between interest activities
- Semantic activities start/end may not be really clear



5. Conclusions and perspectives

- Activities of Daily Living Indexing and Motion Based Temporal Segmentation methods have been presented
- Encouraging results. Good discriminative power between interest and not relevant activities. Difficulty of modeling activities which may seems similar in current description space
- Difficulty to obtain videos (no such public databases available)
- Tests on a larger corpus recorded in different patients' home: 10h of videos available (work in progress)
- Mid-level and local descriptors: Object detection
- Activity dependent number of states via Entropy Minimization
- Late fusion with Coupled HMMs

Thank you for your attention.

Questions?