



# Elements of behaviour recognition with ego-centric Visual Sensors. Application to AD patients assessment

**Jenny Benois-Pineau,**

LaBRI – Université de Bordeaux – CNRS UMR 5800/ University Bordeaux1

**H. Boujut, V. Buso, L. Letoupin, R. Megret, S. Karaman, V. Dovgalecs, Ivan Gonsalez-Diaz(University Bordeaux1)**

**Y. Gaestel, J.-F. Dartigues (INSERM)**

# Summary

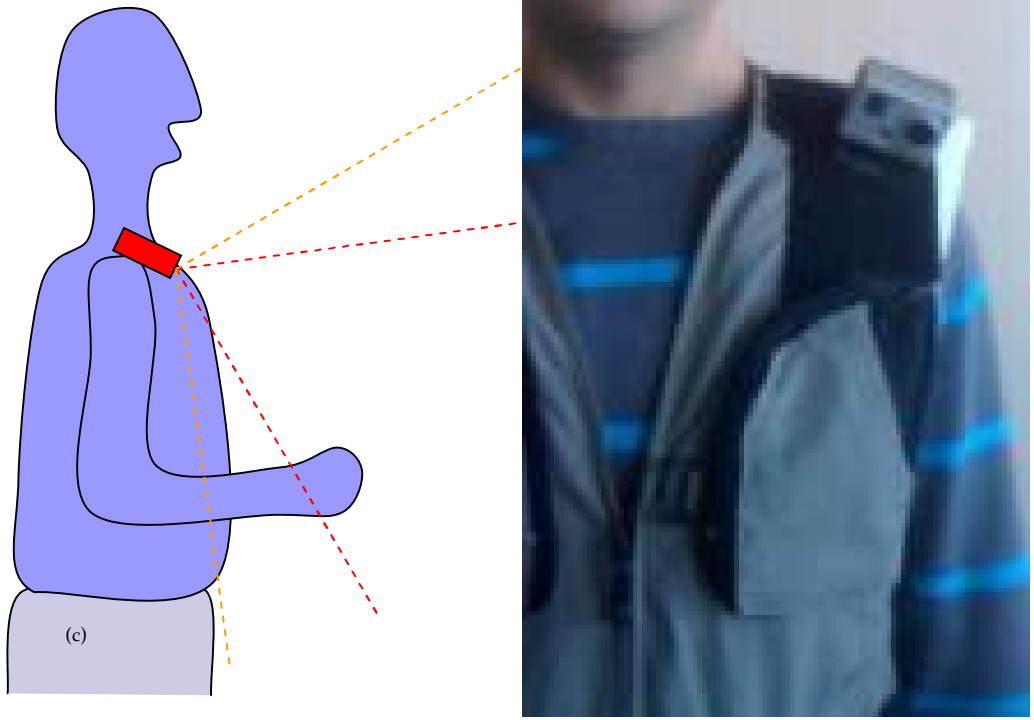
1. Introduction and motivation
2. Egocentric sensors : Wearable videos
3. Fusion of multiple cues: eraly fusion framework
4. Fusion of multiple cues: late fusion and intermediate fusion
5. Model of activities as « Active Objects&Context »
   - 5.1. 3D Localization from wearable camera
   - 5.2. Object recognition with visual saliency
6. Conclusion and perspectives

# 1. Introduction and motivation

- Recognition of Instrumental Activities of Daily Living (IADL) of patients suffering from Alzheimer Disease
  - Decline in IADL is correlated with future dementia
- IADL analysis:
  - Survey for the patient and relatives → subjective answers
  - 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: IMMED ANR, Dem@care IP FP7 EC

# 2. Egocentric sensors : Wearable video

- 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
- Natural integration into home visit by paramedical assistants protocol

Loxie – ear-weared
Looking-glaces weared with eye-tracker, iPhone….

# Wearable videos

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




# Wearable Camera analysis

Monitoring of Instrumental Activities of Daily Living (IADLs)

- Meal related activities (Preparation, meal, cleaning…)
- Instrumental activities (telephone, …)
- For directed activities in @Lab
  or undirected activities in @Home
  and @NursingHome


Advantages of wearable camera

- Provides a close-up view on IADLs
  - Objects *
  - Location *
  - Actions *
- Complements fixed sensors with more detailed information
- Follows the person in multiple places of interest (sink, table, telephone…) without full equipment of the environment

Visual positionning from wearable camera

# Wearable video datasets and scenarii

IMMED Dataset @Home

UB1, CHU Bordeaux

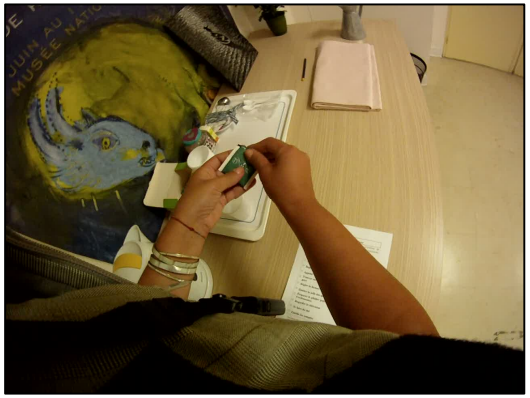
12 volunteers and 42 patients

ADL Dataset @Home

UC Irvine

20 healthy volunteers

- Dem@care Dataset
  - CHUN @Lab
  - 3 healthy, 5 MCI, 3 AD, 2 mixed dementia

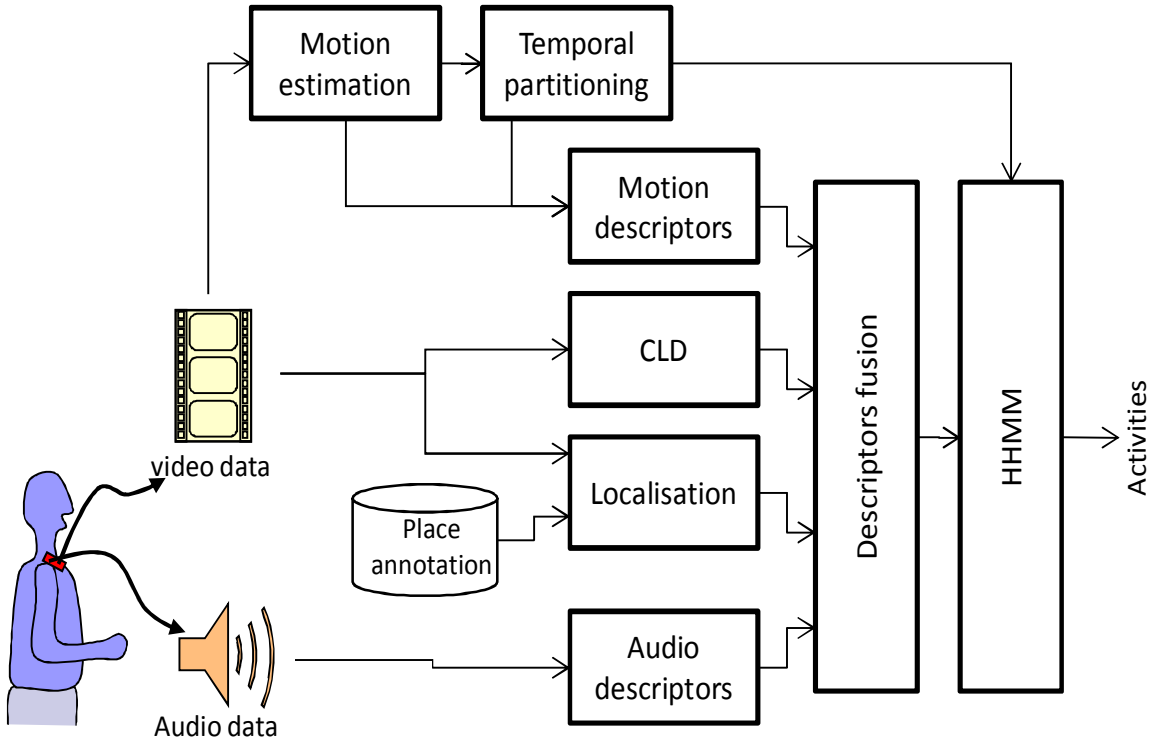


Unconstrained scenario

Constrained scenario

Visual positionning from wearable camera

# 3. Fusion of multiple cues for activities recognition Early fusion framework




The IMMED Framework

# Temporal Partitioning(1)

- 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

# Temporal Partitioning(2)

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

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

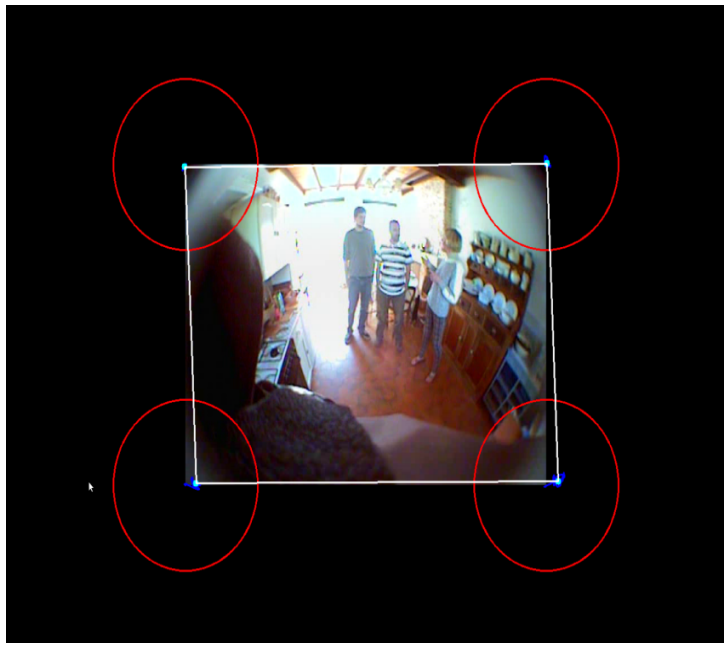
[Krämer, Benois-Pineau, Domenger] Camera Motion Detection in the Rough Indexing Paradigm, TRECVID'2005.

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

# Temporal Partitioning(3)

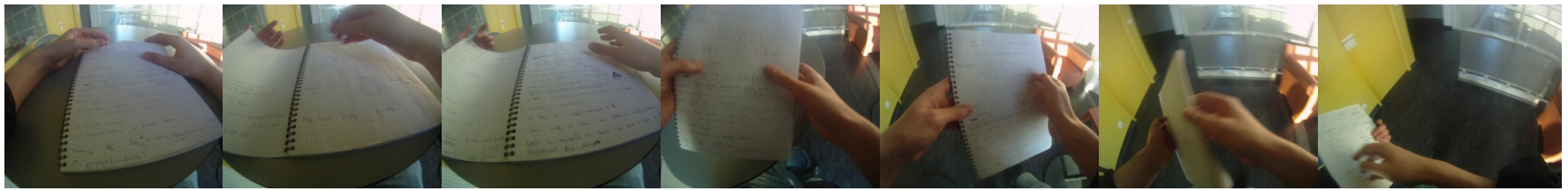
- Threshold *t* defined as a percentage *p* of image width *w*
  p=0.2 … 0.25

$$t = p \times w$$



# Temporal Partitioning(4)
# Video Summary

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



# Description space: fusion of features (1)

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

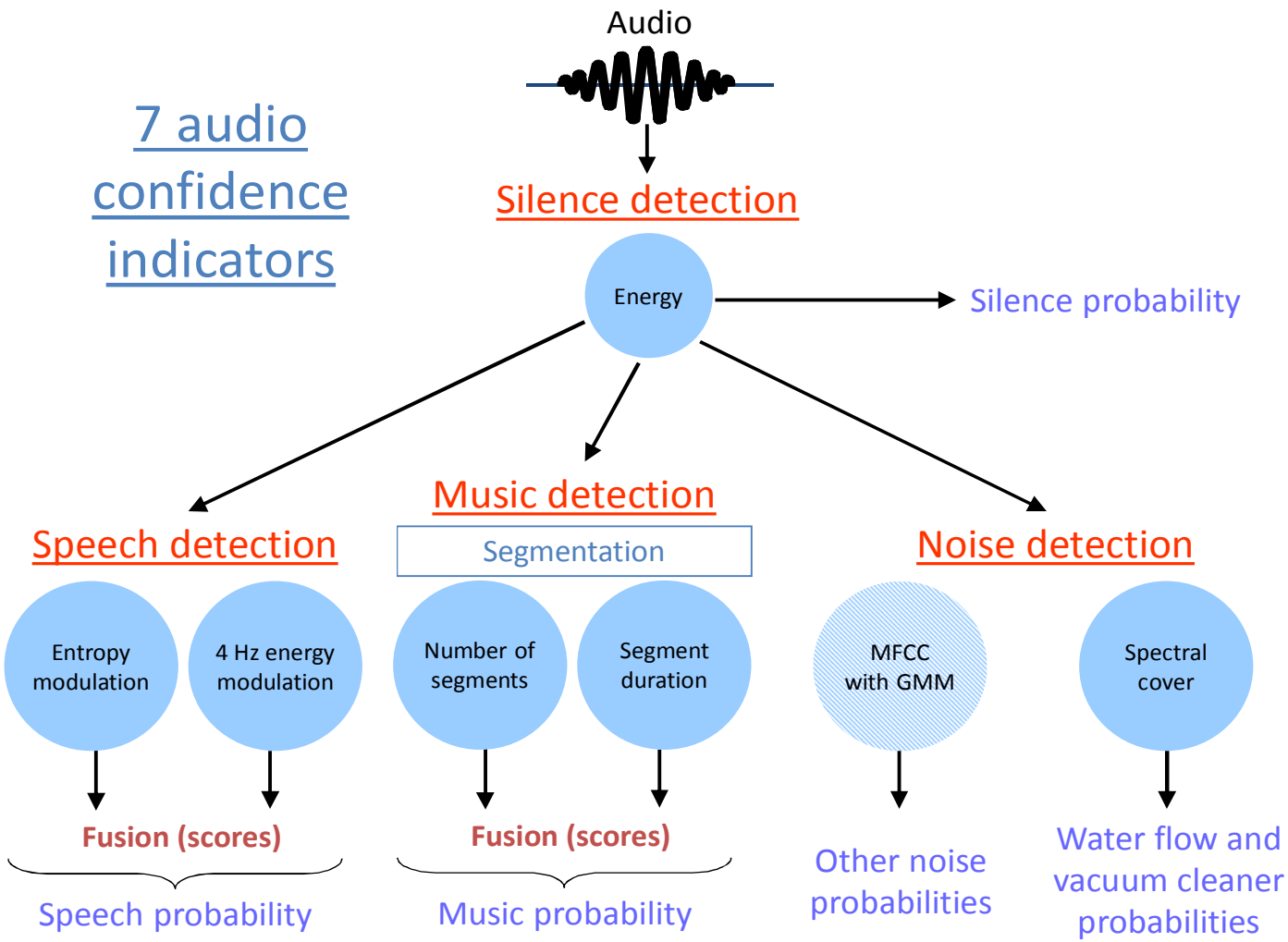
  6 coefficients for luminance, 3 for each chrominance

  - 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
- Audio: x(Audio) : probabilistic features SMN…

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

# Description space(2). Audio



# Description space(3). Motion

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

$$H_{tpe}[i] += 1 \quad if \quad \log(a^2) < i \times s_h \quad for \quad i = 1$$

$$H_{tpe}[i] += 1 \quad if \quad (i-1) \times s_h \le \log(a^2) < i \times s_h \quad for \quad i = 2..N_e - 1$$

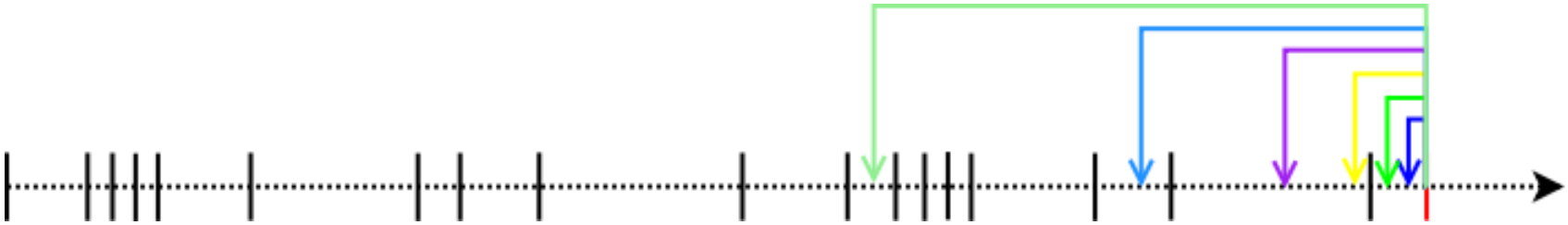
$$H_{tpe}[i] += 1 \quad if \quad \log(a^2) \ge i \times s_h \quad for \quad i = N_e$$

- 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 |

# Description space(4). Motion

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



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

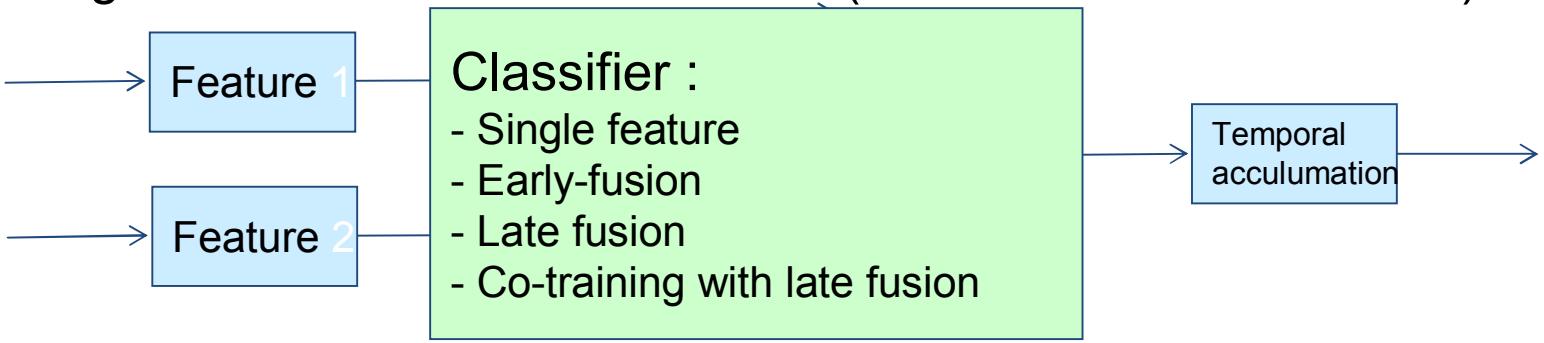
- Average histogram over all frames within the segment
- Characterizes the motion history, the strength of motion even outside the current segment

$2^6$=64 frames → 2s, $2^8$=256 frames → 8.5s x($H_c$) $\in \Re^6$ or $\Re^8$

- Residual motion $RM_b = \sqrt{\frac{\sum_{k=1,l=1}^{k=N,l=M}(\Delta \boldsymbol{dx}_{k,l}{}^2 + \Delta \boldsymbol{dy}_{k,l}{}^2)}{N * M}}$ 4x4 x(RM) $\in \Re^{16}$

# Room Recognition from Wearable Camera

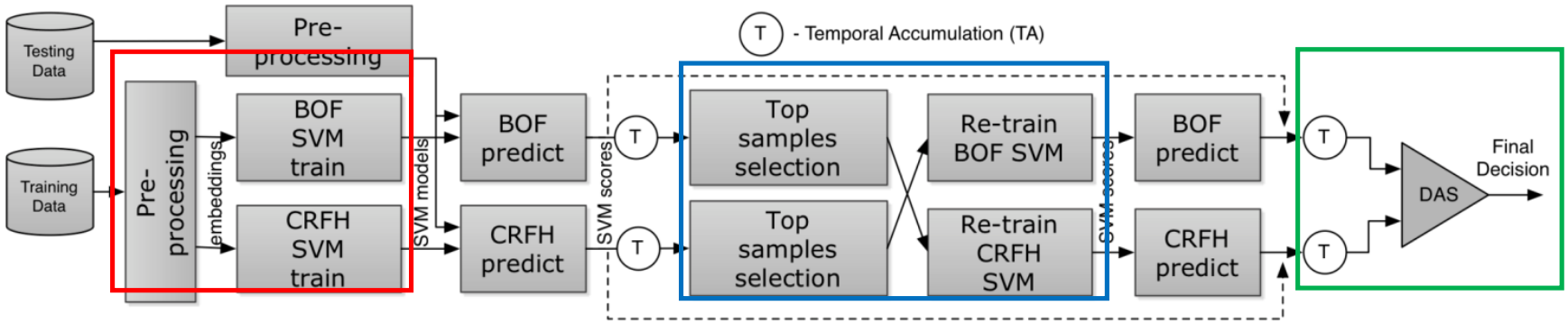
- Classification based recognition
  - **Multi-class classifiers**, trained on a bootstrap video
  - Temporal postprocessing of the estimates (temporal accumulation)
  - 4 global architectures considered (based on SVM classifiers)




  - Visual features considered :
    - Bag-of-Visual-Words (BoVW)
    - Spatial Pyramid Histograms (SPH)
    - Composed Receptive Field Histograms (CRFH)

# A semi-supervised multi-modal framework based on co-training and late fusion

**Global architecture**




Multiple features

Co-training

Late-fusion

Co-training exploits unannotated data by transfering reliable information from one feature pipeline to the other.

Time-information is used to consolidate estimates

Visual positionning from wearable camera

# Co-training: a semi-supervised multi-modal framework

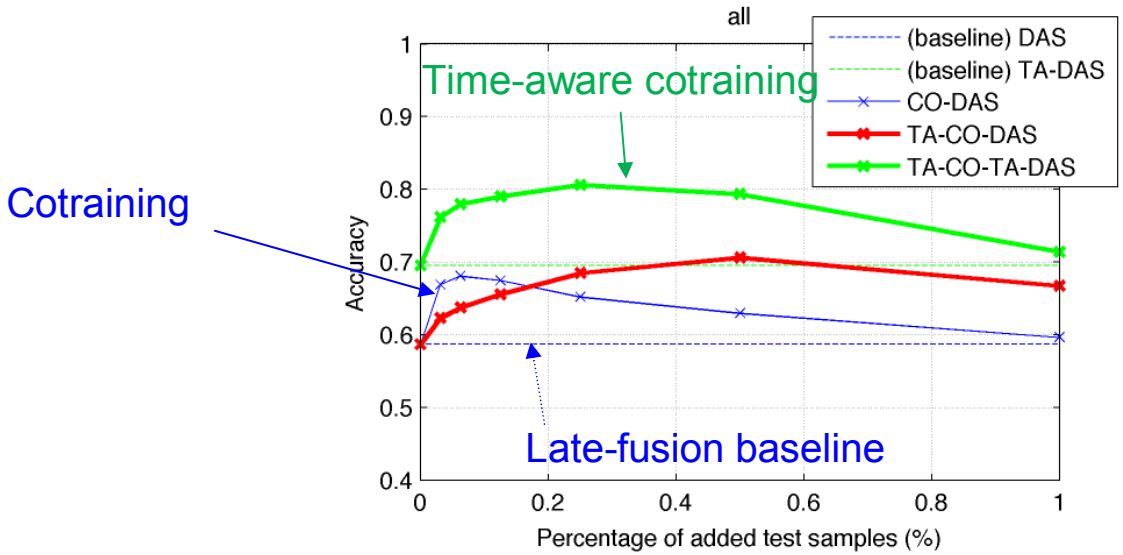



**Illustration of correct detection rate on IDOL2 database**

Combining co-training, temporal accumulation and late-fusion produce the best results, by leveraging non annotated data

Visual positionning from wearable camera

# IMMED Dataset @Home

**(subset used for experiments)**

14 sessions at patient's homes.

- Each session is divided in training/testing

6 classes:

- Bathroom, bedroom, kitchen, living-room, outside, other

Training

- Average 5 min / session
- Bootstrap
  - Patient asked to present the various rooms in the appartment
  - Used for location model training

Testing

- Average 25 min / session
- Guided activities



Visual positionning from wearable camera

# RRWC performance on IMMED @Home

## Discussion

Best performances obtained using **late-fusion approaches** with temporal accumulation

Global accuracy in the 50-60% range. Difficult dataset due to specific applicative constraints in IMMED leading to low quality of training data for location.

| Algorithm | Feature space | Temporal accumulation | |
|---|---|---|---|
| | | without | with |
| SVM | BOVW | 0.49 | 0.52 |
| | CRFH | 0.48 | 0.53 |
| | SPH3 | 0.47 | 0.49 |
| Early-fusion | BOVW+SPH3 | 0.48 | 0.50 |
| | BOVW+CRFH | 0.50 | 0.54 |
| | CRFH+SPH3 | 0.48 | 0.51 |
| Late-fusion | BOVW+SPH3 | 0.51 | 0.56 |
| | BOVW+CRFH | 0.51 | 0.56 |
| | CRFH+SPH3 | 0.50 | 0.54 |
| Co-training with late-fusion | BOVW+SPH3 | 0.50 | 0.53 |
| | BOVW+CRFH | 0.54 | **0.58** |
| | CRFH+SPH3 | 0.54 | 0.57 |

Average accuracy for room recognition on IMMED dataset.

# Dem@Care Dataset @Lab

Training (22 videos)

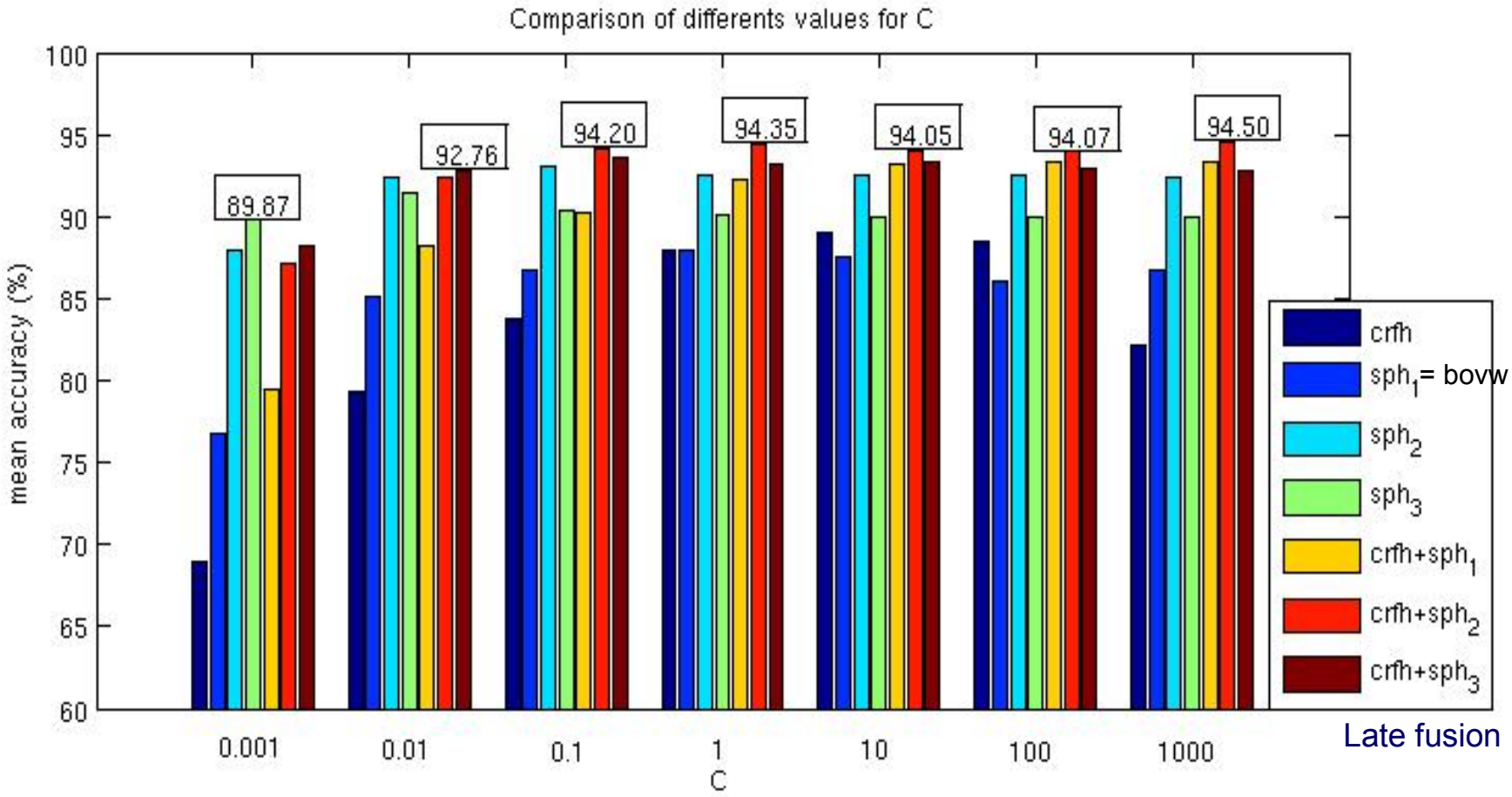
- Total time: 6h42min

Testing (22 videos).

- Total time: 6h20min
- Average duration for 1 video: 18min
- 5 classes:
  - phone_place, tea_place, tv_place, table_place, meication_place


Visual
wearable camera

# RRWC performance on Dem@Care @Lab

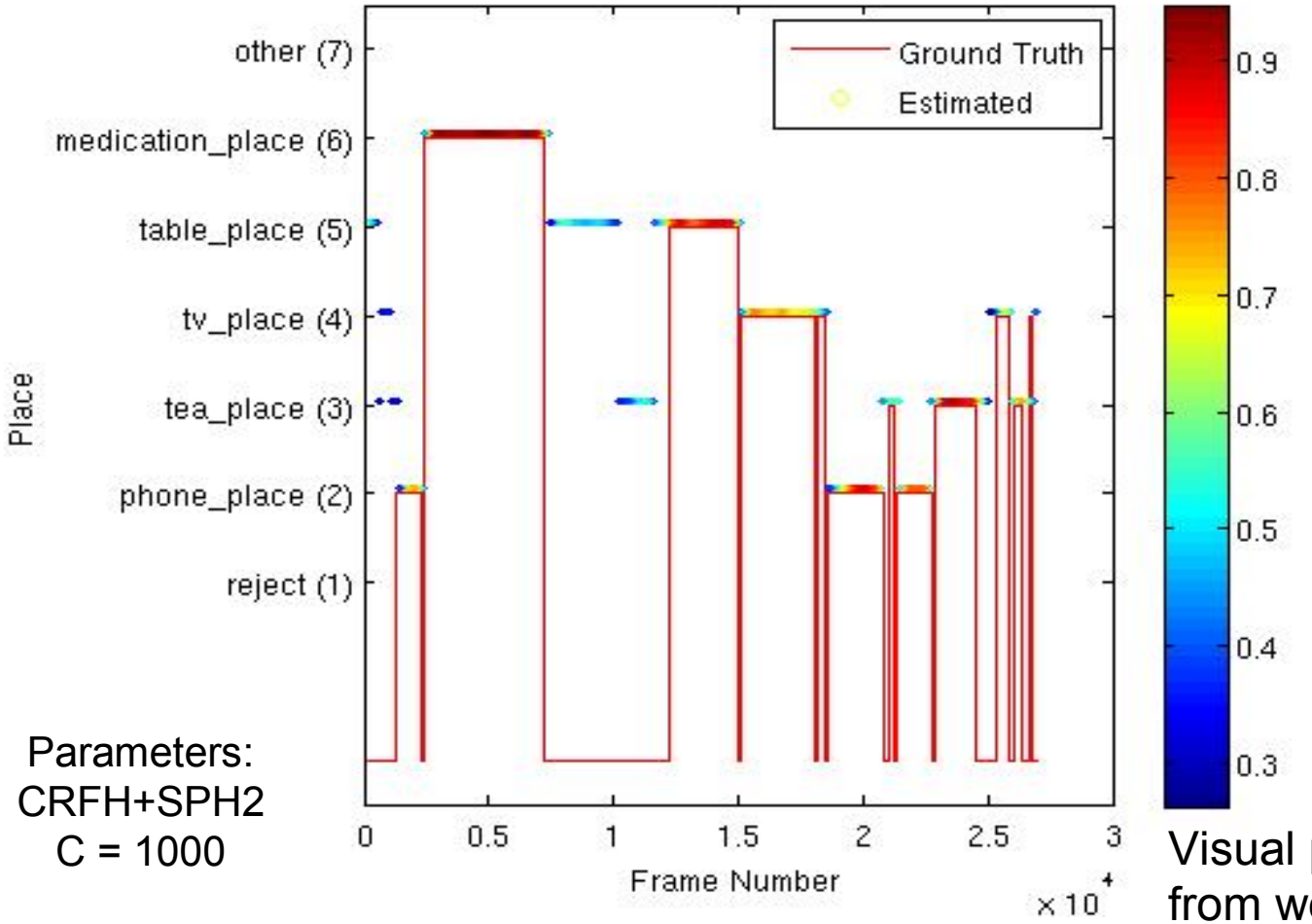



Late fusion

**Larger homogeneous training set yield: 94% accuracy**

Visual positionning from wearable camera

# Details on one test video




Normalized score for best class

Parameters:
CRFH+SPH2
C = 1000

Visual positionning from wearable camera

# Discussion on the results

Feature fusion (late fusion) works best for place classification

Semi-supervised training slightly improve performances on IMMED dataset

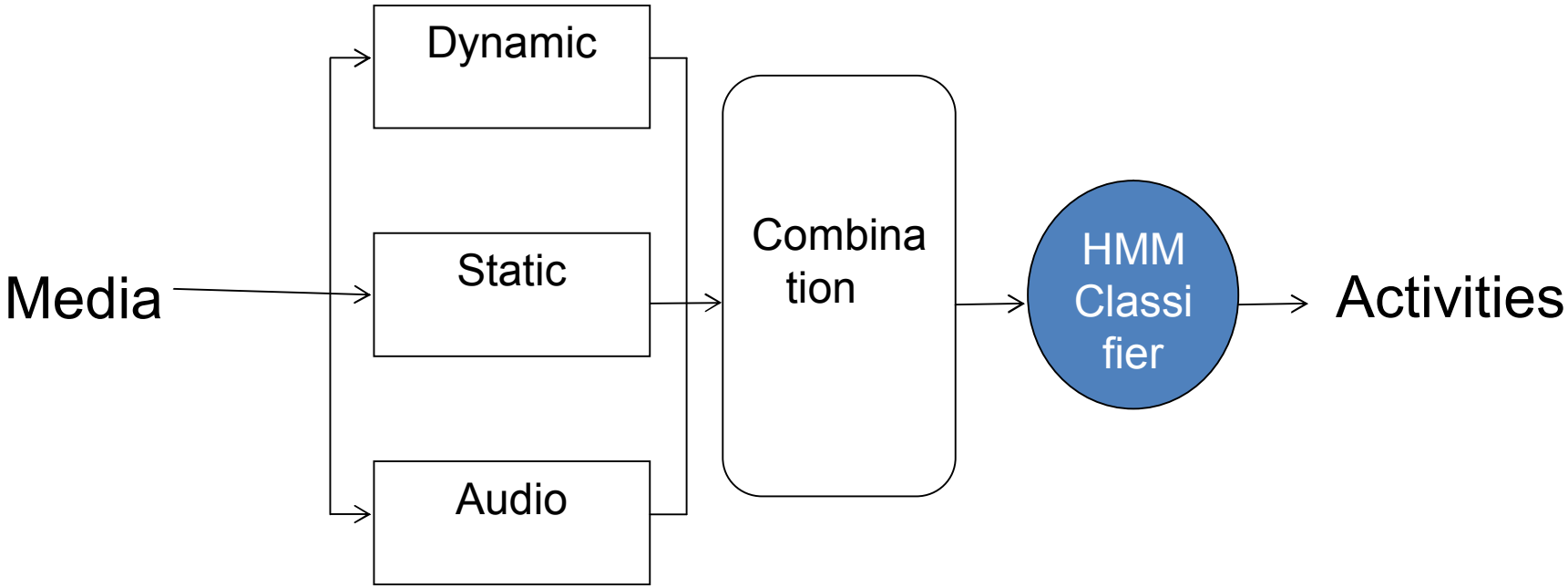
But performance highly dependent on good training:

- Sufficient amount of data
- Homogeneous content

Open issues

- More selective criteria for recognition/rejection to improve reliability of positive decision

# Early fusion of all features



# Description space(5)

- 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 $\Re^{8}$
    - $x(RM) \in \Re^{16}$
  - Localization: $N_{home}$ between 5 and 10.
    - $x(Loc) \in \Re^{Nhome}$
  - Audio :
    - $x(Audio) \in \Re^{7}$

# Description space(6)

- ♯Possible combinations of descriptors : $2^6 - 1 = 63$

| **Descriptors** | Audio | Loc | RM | Htpe | Hc | CLD | config min | config max |
|---|---|---|---|---|---|---|---|---|
| **Dimensions** | 7 | 7 | 16 | 10 | 8 | 12 | 7 | 60 |

# Model of the content: activities recognition

A two level hierarchical HMM:

- Higher level:

transition between activities

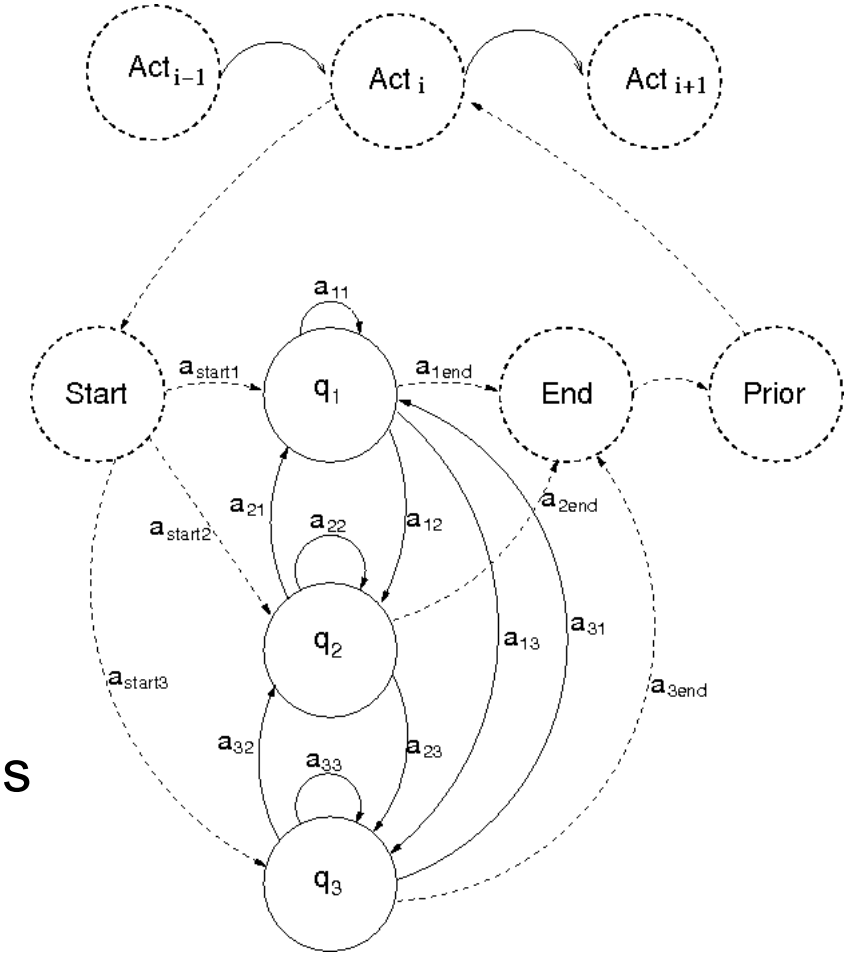
  - Example activities:

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

- Bottom level:

activity description

  - Activity: HMM with 3/5/7/8 states
  - Observations model: GMM
  - Prior probability of activity



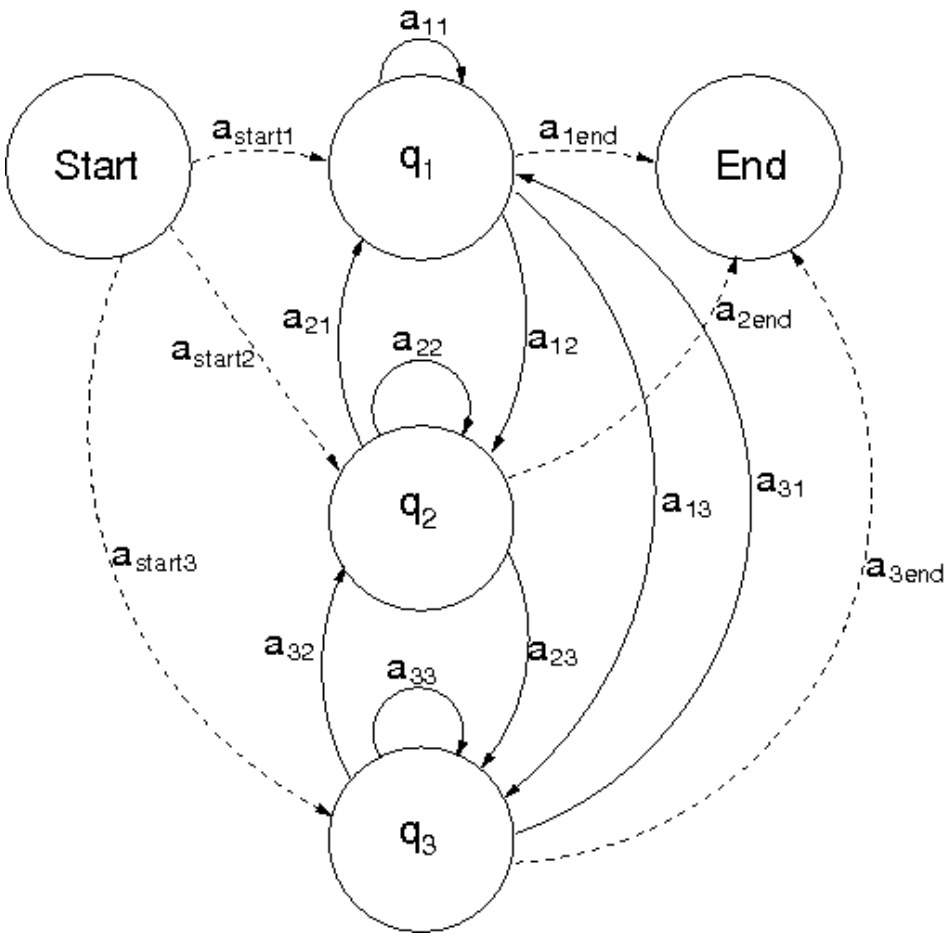
# Activities recognition

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 $a_{iend}$



# Video Corpus

| Corpus | Healthy volonteers/ Patients | Number of videos | Duration |
|---|---|---|---|
| IMMED | 12 healthy volonteers | 15 | 7H16 |
| IMMED | 42 patients | 46 | 17H04 |
| TOTAL | 12 healthy volonteers + 42 patients | 61 | 24H20 |

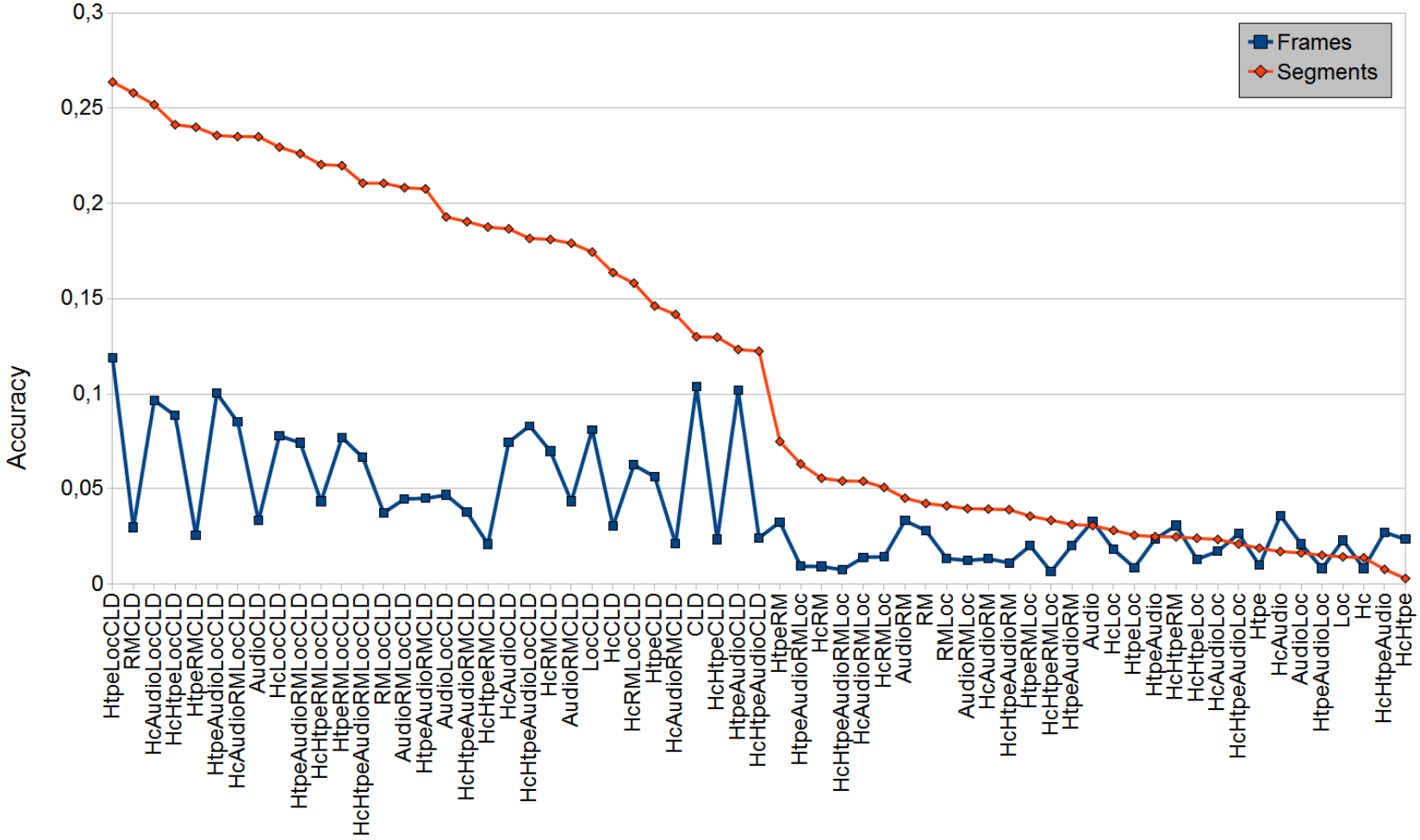
# Evaluation protocol

| | |
|---|---|
| precision= TP/(TP+FP) | recall= TP/(TP+FN) |
| accuracy= (TP+TN)/(TP+FP+TN+FN) | F-score= 2/(1/precision+1/recall) |

- Leave-one-out cross validation scheme (one video left)
- Resultas are averaged
- Training is performed over a sub sampling of smoothed (10 frames) data.
- Label of a segment is derived by majority vote of frames results

# Recognition of Activities 5 videos. Descriptors?

3states LL HMM,
1 state None

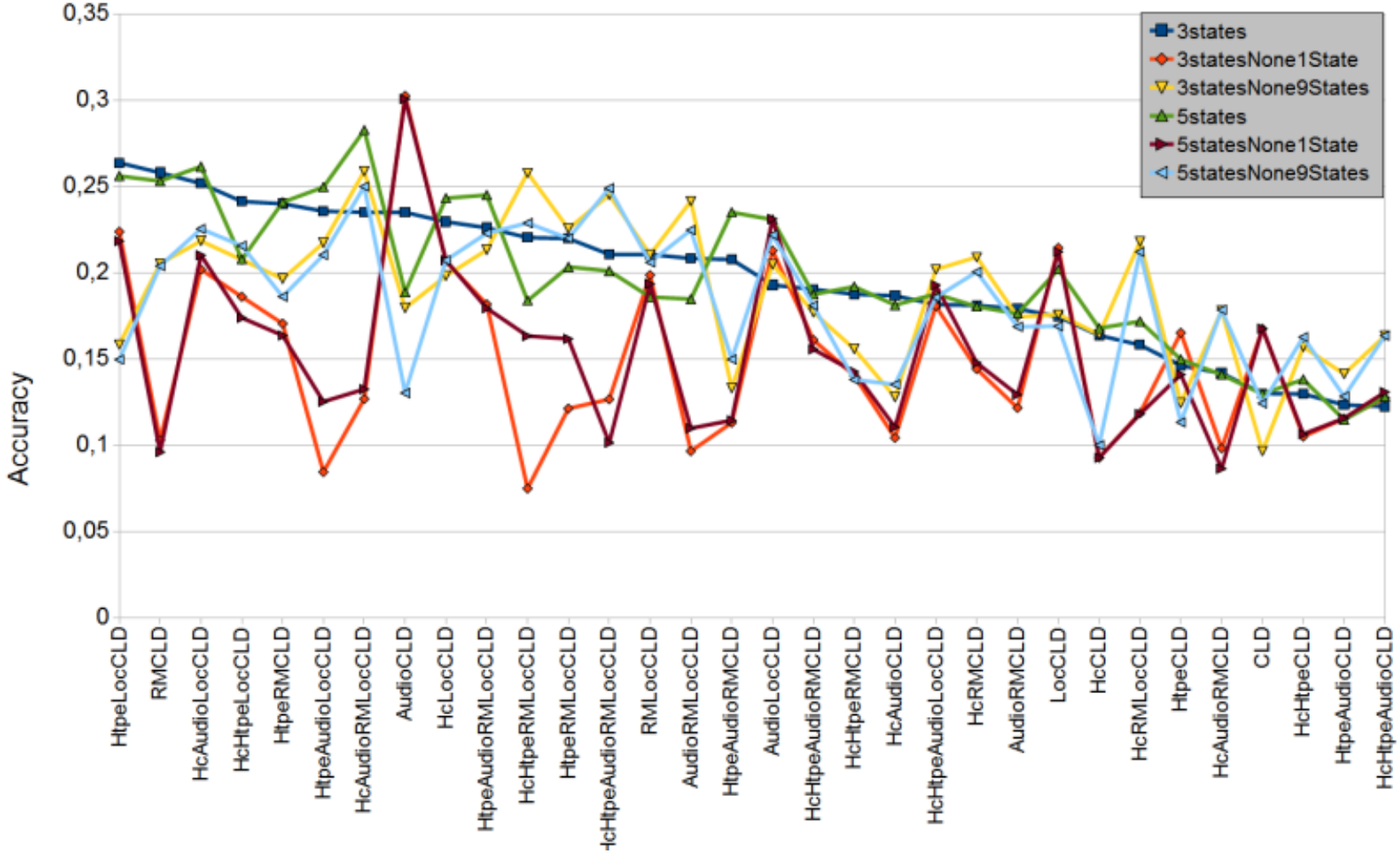


# Activities

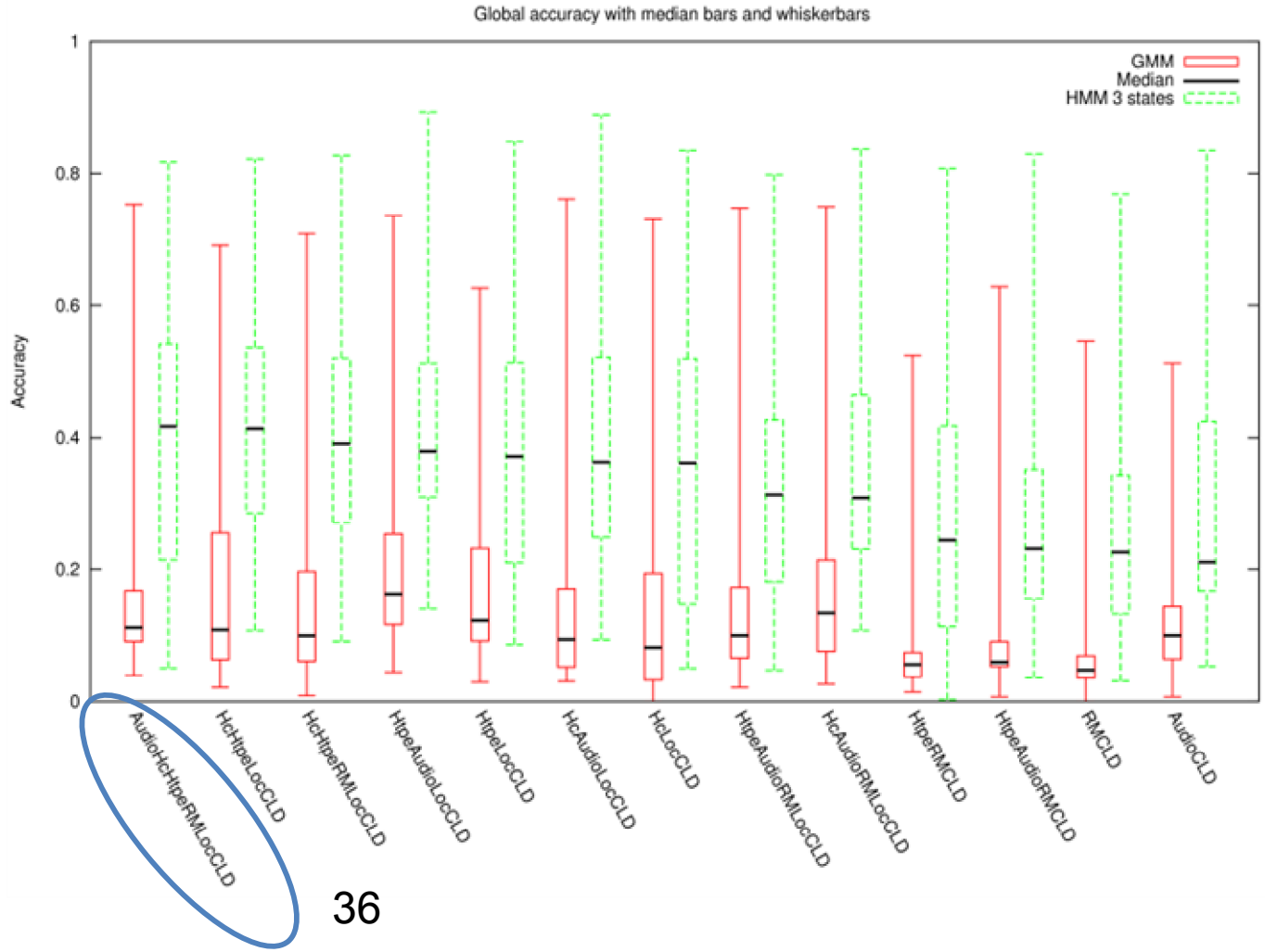
“Plant Spraying”, “Remove Dishes”,
“Wipe Dishes”, “Meds Management”,
“Hand Cleaning”, “Brushing Teeth”,
“Washing Dishes”, “Sweeping”, “Making Coffee”,
“Making Snack”, “Picking Up Dust”,
“Put Sweep Back”, “Hair Brushing”, “Serving”,
“Phone” and “TV”. (16)

Result : 5 times better than chance

# Results using segments as observations



# Comparison with a GMM baseline (23 activities on 26 videos)



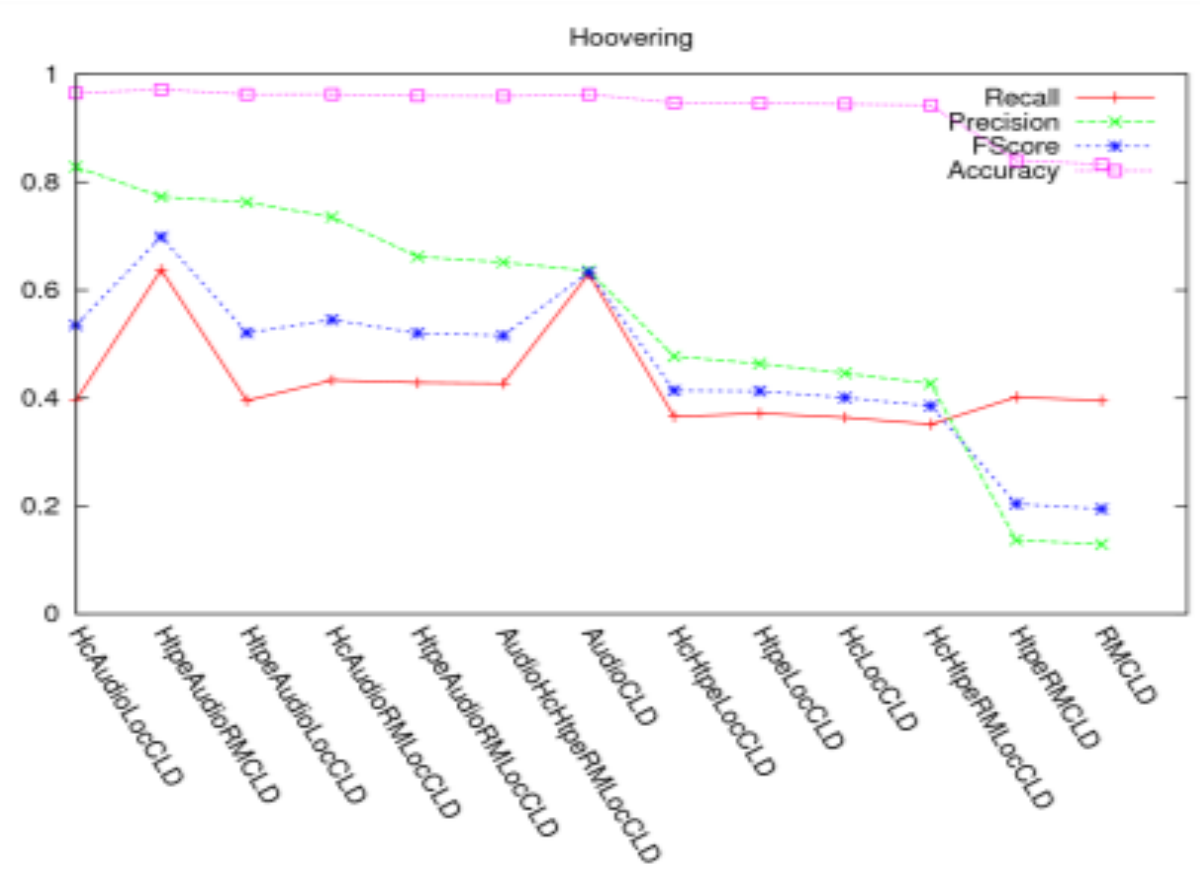
## Slide 36

**2** Jenny Benois-Pineau; 20/1/2013

# 23 activities

“Food manual preparation”, “Displacement free”, “Hoovering”, “Sweeping”, “Cleaning”, “Making a bed”, “Using dustpan”, “Throwing into the dustbin”, “Cleaning dishes by hand”, “Body hygiene”, “Hygiene beauty”, “Getting dressed”, “Gardening”, “Reading”, “Watching TV”, “Working on computer”, “Making coffee”, “Cooking”, “Using washing machine”, “Using microwave”, “Taking medicines”, “Using phone”, “Making home visit”.

# « Hovering »


Hoovering
Recall
Precision
FScore
Accuracy
1
0.8
0.6
0.4
0.2
0
HcAudioLocCLD
HtpeAudioRMCLD
HtpeAudioLocCLD
HcAudioRMLocCLD
HtpeAudioRMLocCLD
AudioHcHtpeRMLocCLD
AudioCLD
HcHtpeLocCLD
HtpeLocCLD
HcLocCLD
HcHtpeRMLocCLD
HtpeRMCLD
RMCLD

# Conclusion on early fusion

Early fusion requires tests of numerous combinations of features.

The best results were achieved for the complete description space

For specific activities optimal combinations of descriptors vary and correspond to « common sense » approach.

S. Karaman, J. Benois-Pineau , V. Dovgalecs, R. Megret, J. Pinquier, R. André-Obrecht, Y. Gaestel, J.-F. Dartigues, « Hierarchical Hidden Markov Model in detecting activities of daily living in wearable videos for studies of dementia », Multimedia Tools and Applications, Springer, 2012, pp. 1-29, DOI 10.1007/s11042-012-117-x Article in Press

# 4. Fusion of multiple cues: intermediate fusion and late fusion

## Intermediate fusion

Treat the different modalities (Dynamic, Static, Audio) separately.

We represent each modality by a stream, that is a set of measures along the time.

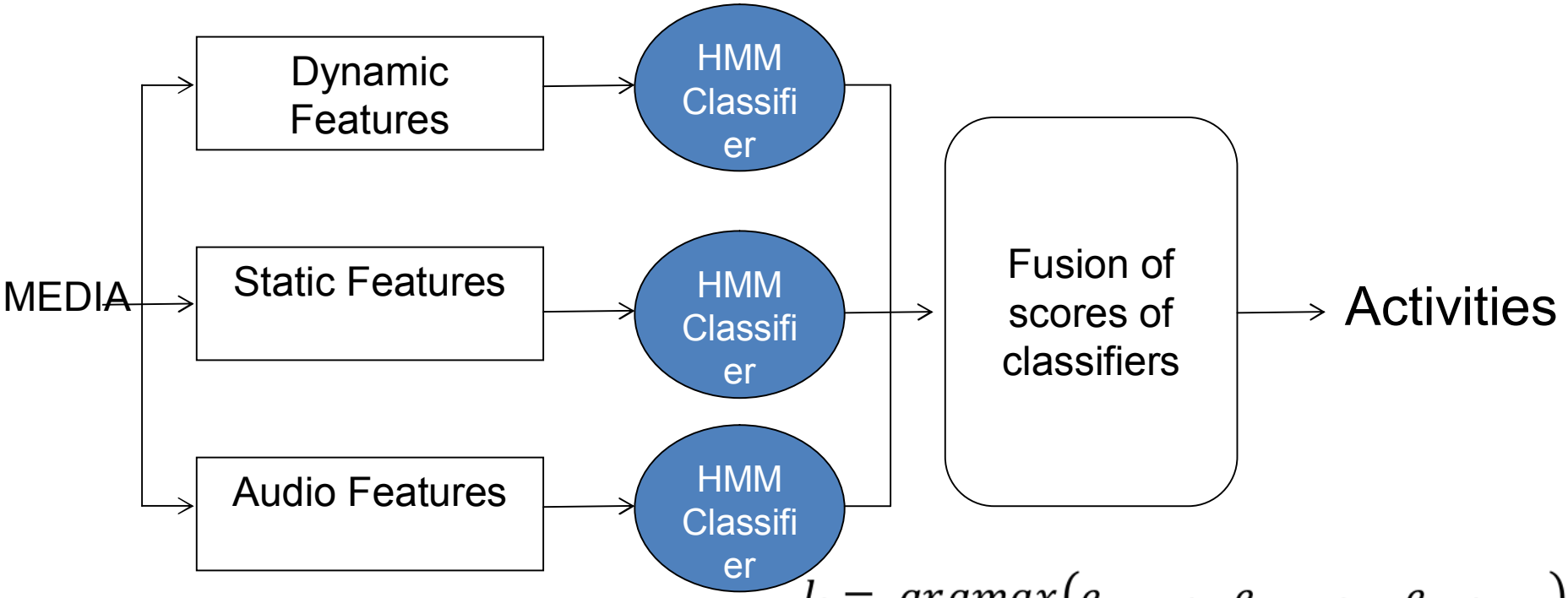
Each state of the HMM models the observations of each stream separately by a Gaussian mixture.

*K* streams of observations $o_{i,1}, \ldots, o_{i,k}$ $o_{i,k} \in \mathbb{R}^{N_k}$

$$\sum N_k = N$$

$$p(o_i, q_j) = \prod_{k=1}^{K} p_k(o_{i,k}, q_j)^{w_{lk}}$$

# Late Fusion framework




$$l_i = argmax\left(e_{l_i Audio}, e_{l_i Motion}, e_{l_i Visual}\right)$$

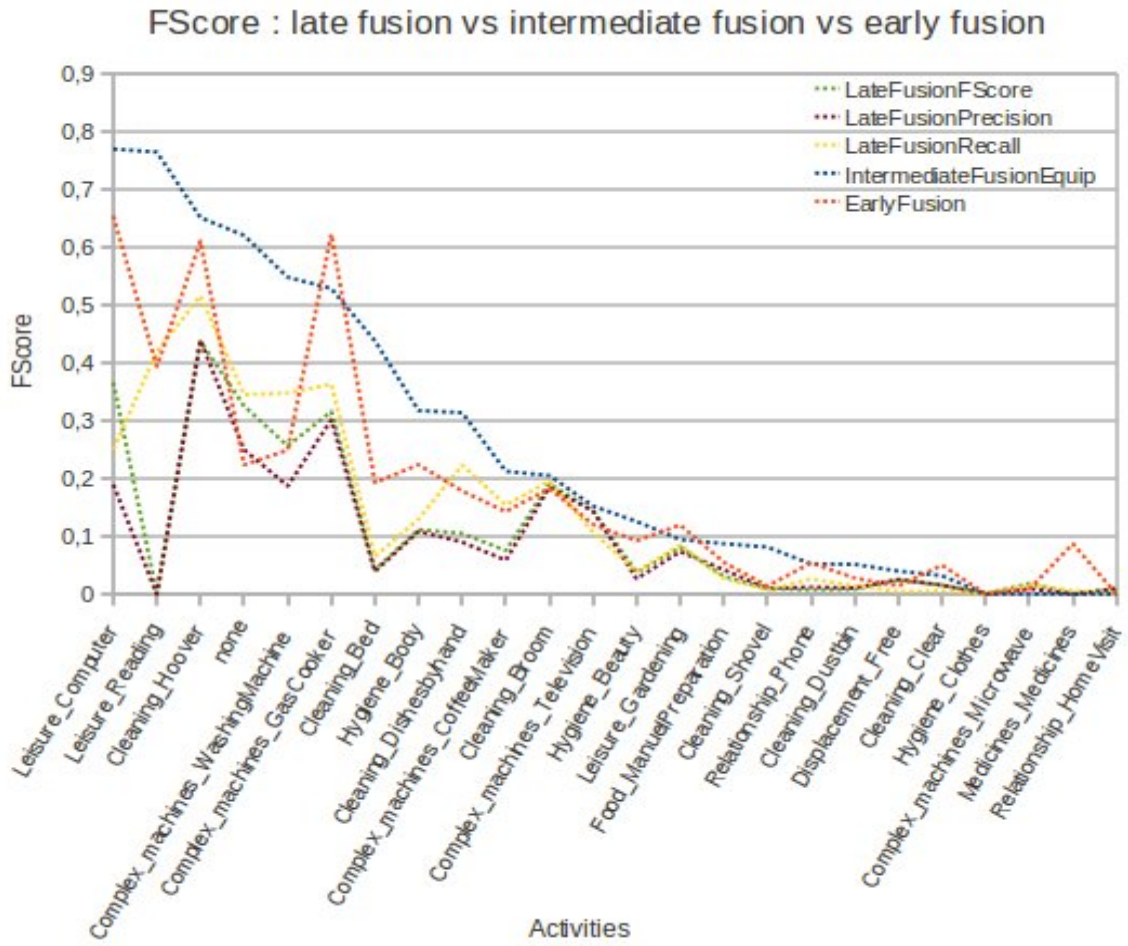
Performance measure of classifiers :
modality $k$, activity $l$

$$e_{lk} = \frac{perf_{lk}}{\sum_k perf_{lk}}$$

# Experimental video corpus in 3-fusion experiment

37 videos recorded by 34 persons (healthy volunteers and patients) for a total of 14 hours of content.

# Results of 3- fusion experiment(1)



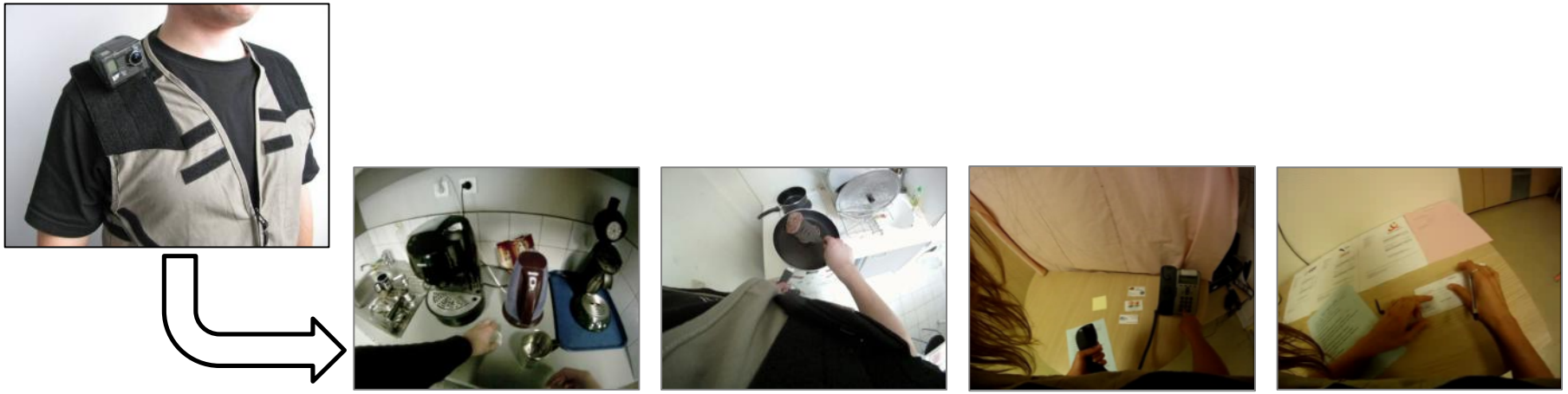
# Results of 3- fusion experiment(2)

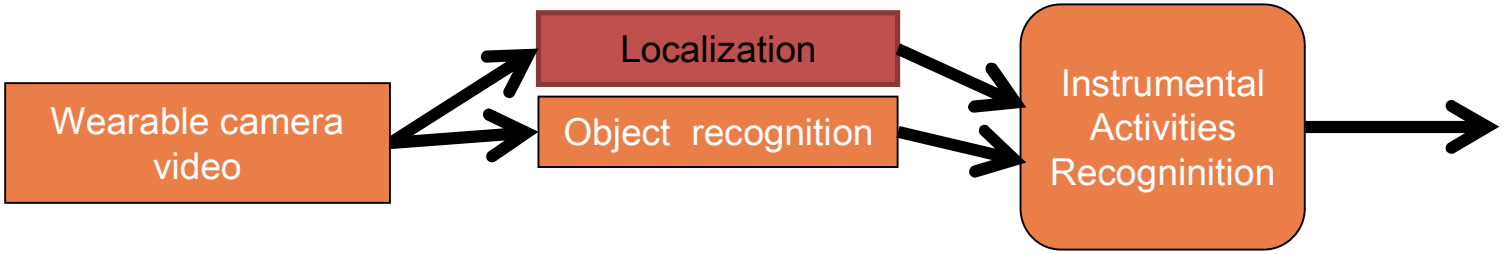
| Metrics (averaged) | Early fusion | Interm. fusion | Late Fusion | | |
|---|---|---|---|---|---|
| | | | F-score trust | Prec. trust | Recall trust |
| Accuracy | 0.207 | **0.442** | 0.215 | 0.188 | 0.210 |
| Precision | 0.174 | **0.267** | 0.106 | 0.097 | 0.131 |
| Recall | 0.284 | **0.288** | 0.171 | 0.155 | 0.221 |
| F-Score | 0.180 | **0.253** | 0.109 | 0.092 | 0.139 |

# Conclusion on 3-fusion experiment

Overall, the experiments have shown that the intermediate fusion has provided consistently better results than the other fusion approaches, on such complex data, supporting its use and expansion in future work.

# 5. Model of activities as « Active Objects & Context »








Visual positionning from wearable camera

# 5.1. 3D localization from wearable camera

Problematics

- How to generate precise position events from wearable video ?
- How to improve the precision of 3D estimation ?

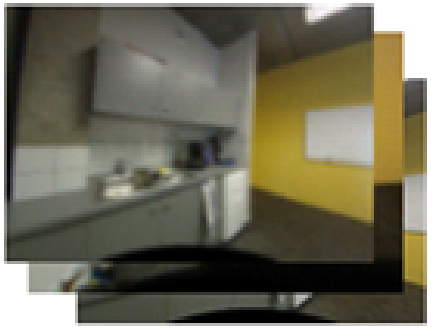
# 2 step positionning

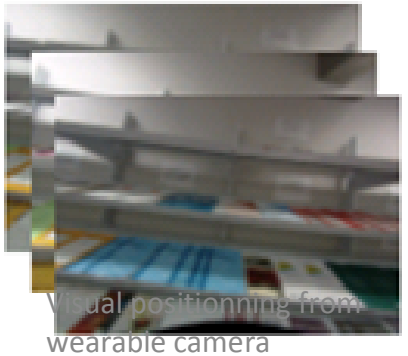
Train: Reconstruct 3D models of places

- Bootstrap frames
- Partial models of environement focused on places of interest
- Structure from Motion (SfM) techniques

Test: Match test images to 3D models

- Extraction of visual features
- Robust 2D-3D matching

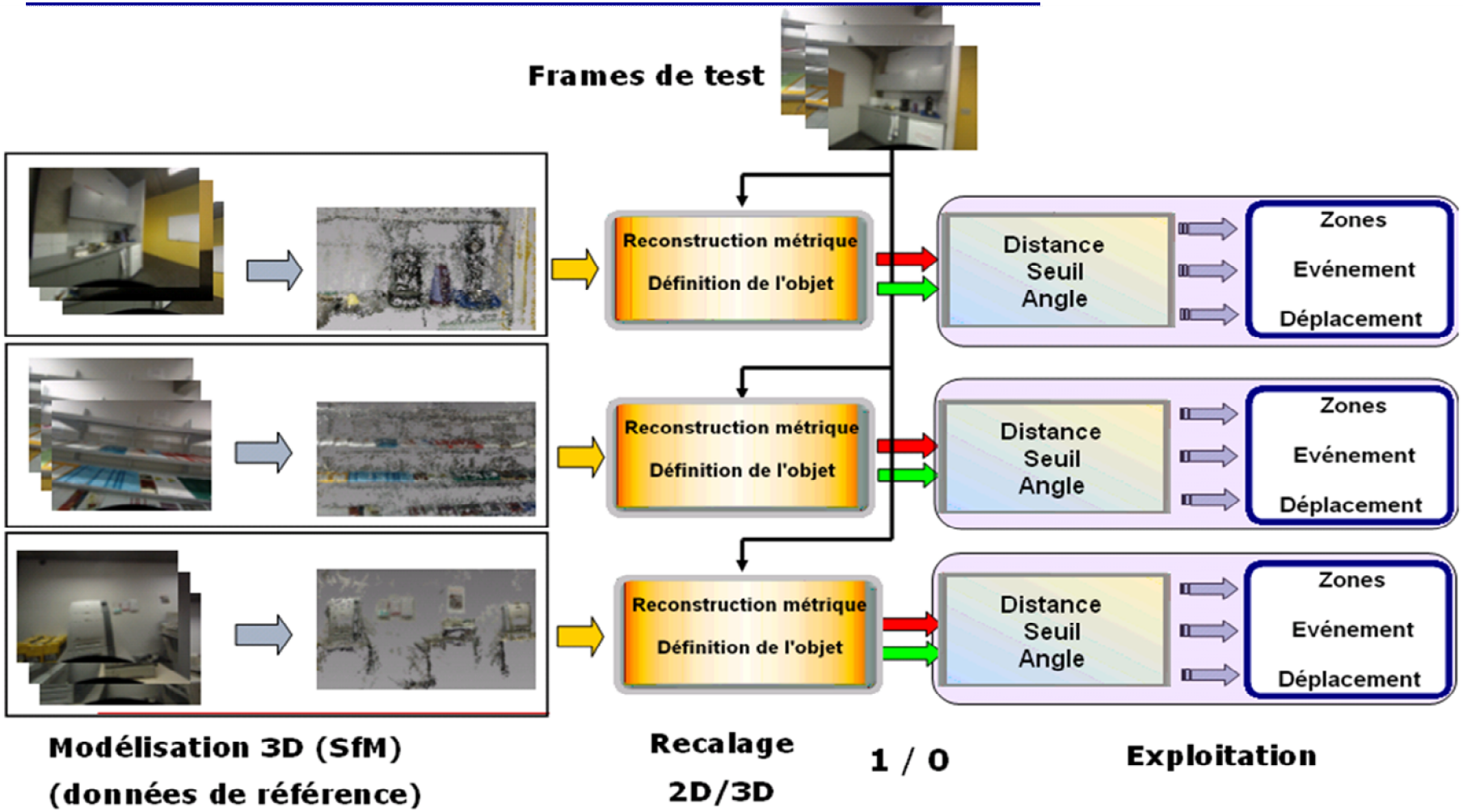






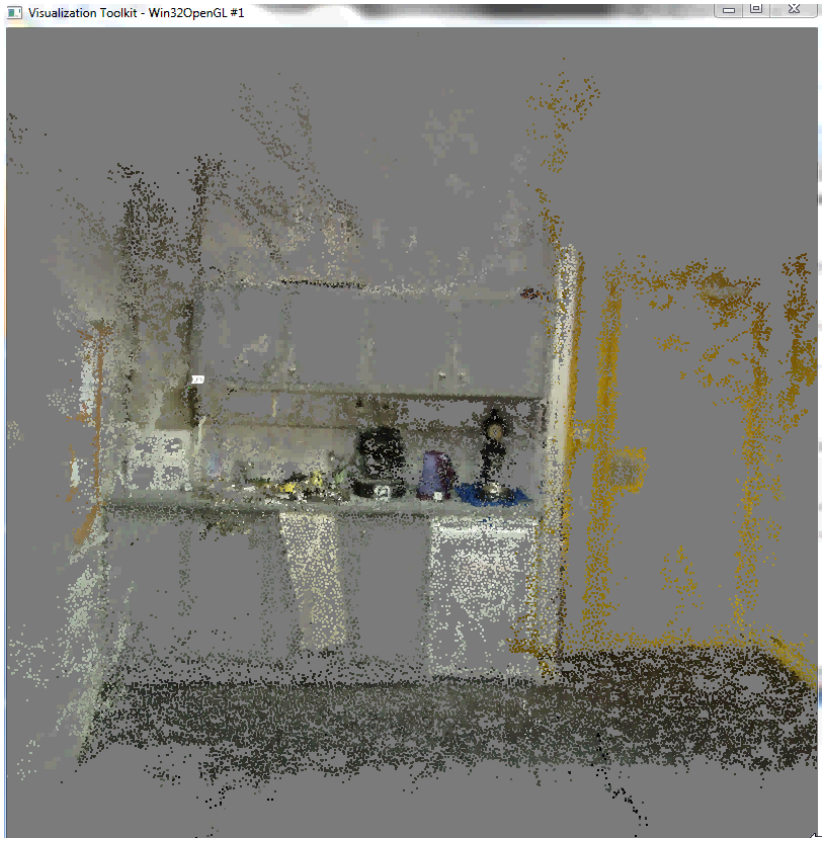
Visual positionning from wearable camera

# GENERAL ARCHITECTURE



# 3D model finalization

- Densification
  - for visualization
- Map global geolocalization
  - manual, using natural markers on the scene
- Definition of specific 3D places of interest
  - Coffee machine, sink…
  - Directly on 3D map

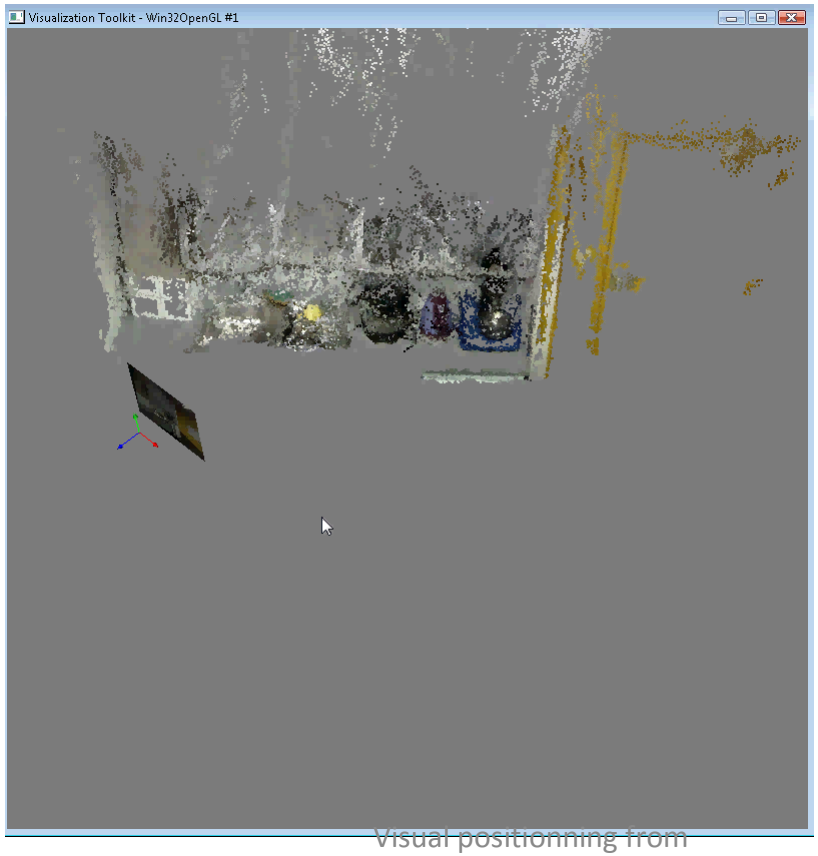



Visual positionning from wearable camera

# 3D positioning

## Using Perspective-n-Point approach

Match 2D points in frame with 3D points in model and apply PnP techniques




Visual positionning from wearable camera

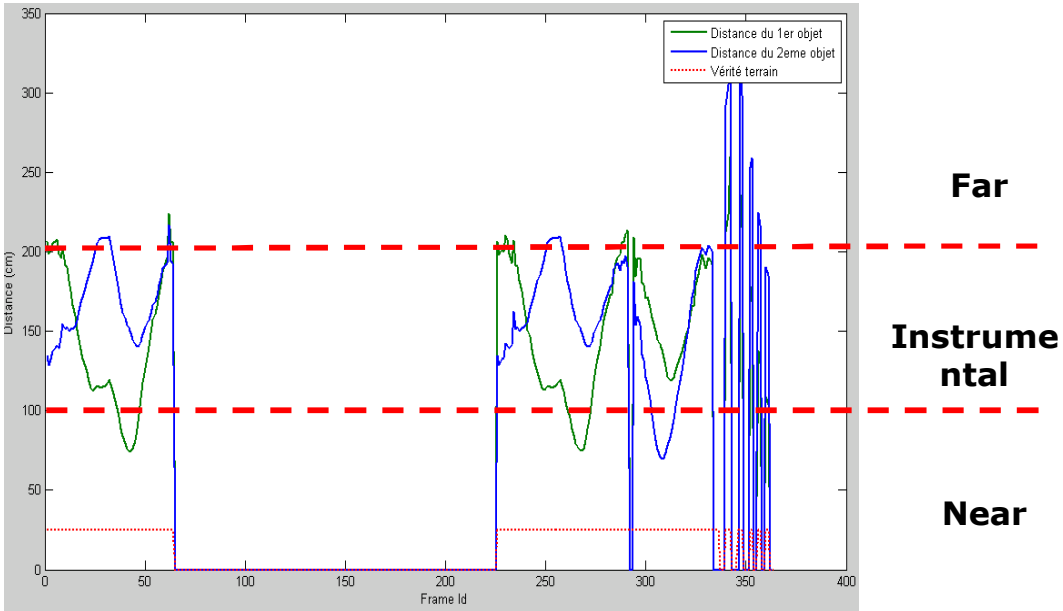
# Exploitation: zone detection

Distance to objects of interest

Zones: Near/Instrumental/Far

Events: zone entrance/exit

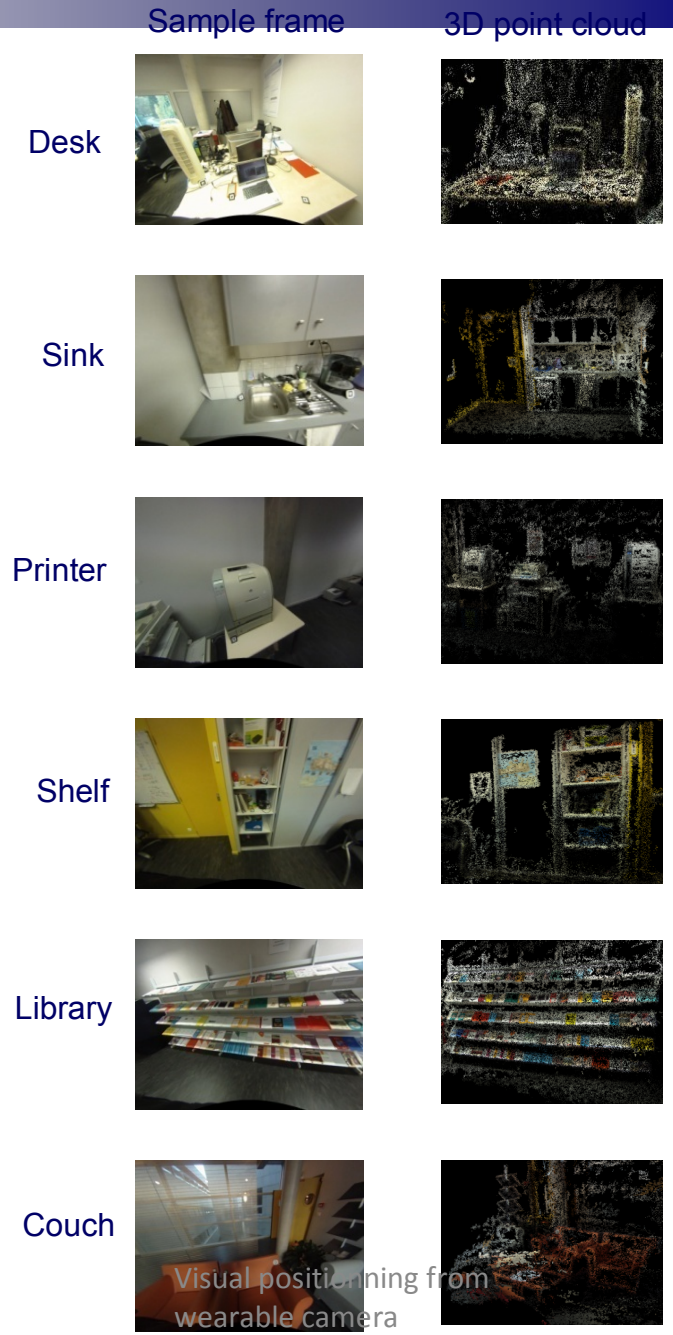
Produce discrete events suitable for higher level analysis



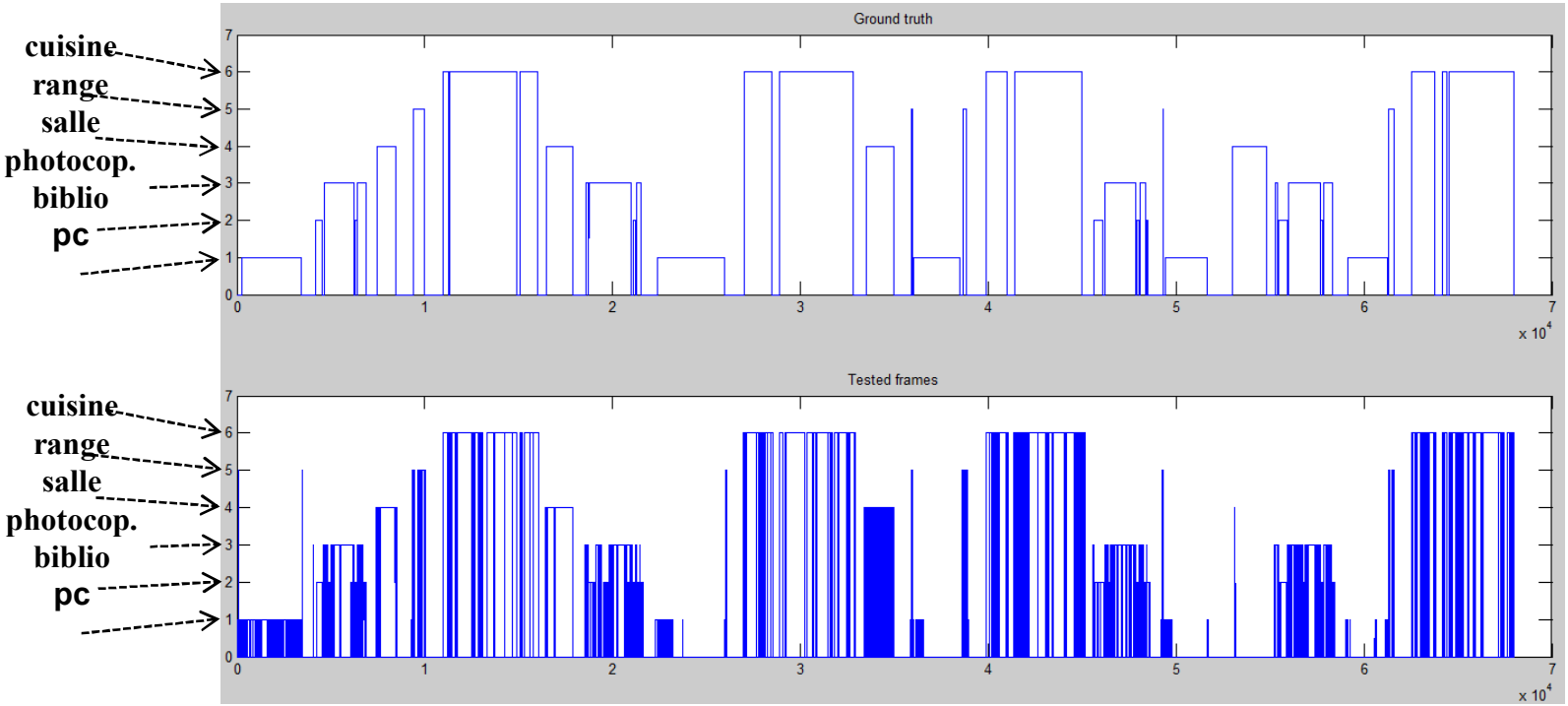

**Object-camera distance**
**(green: cofee machine, blue: sink)**

# Experiments

- 2 test sequences
  - 68000 and 63000 frames
  - 6 classes of interest (desk, sink, printer, shelf, library, couch)
  - Repeated events
- One bootstrap video
  - Through each place
  - 6 reference 3D models
  - Reconstruction requires less than 100 frames / model



# Recognition performances

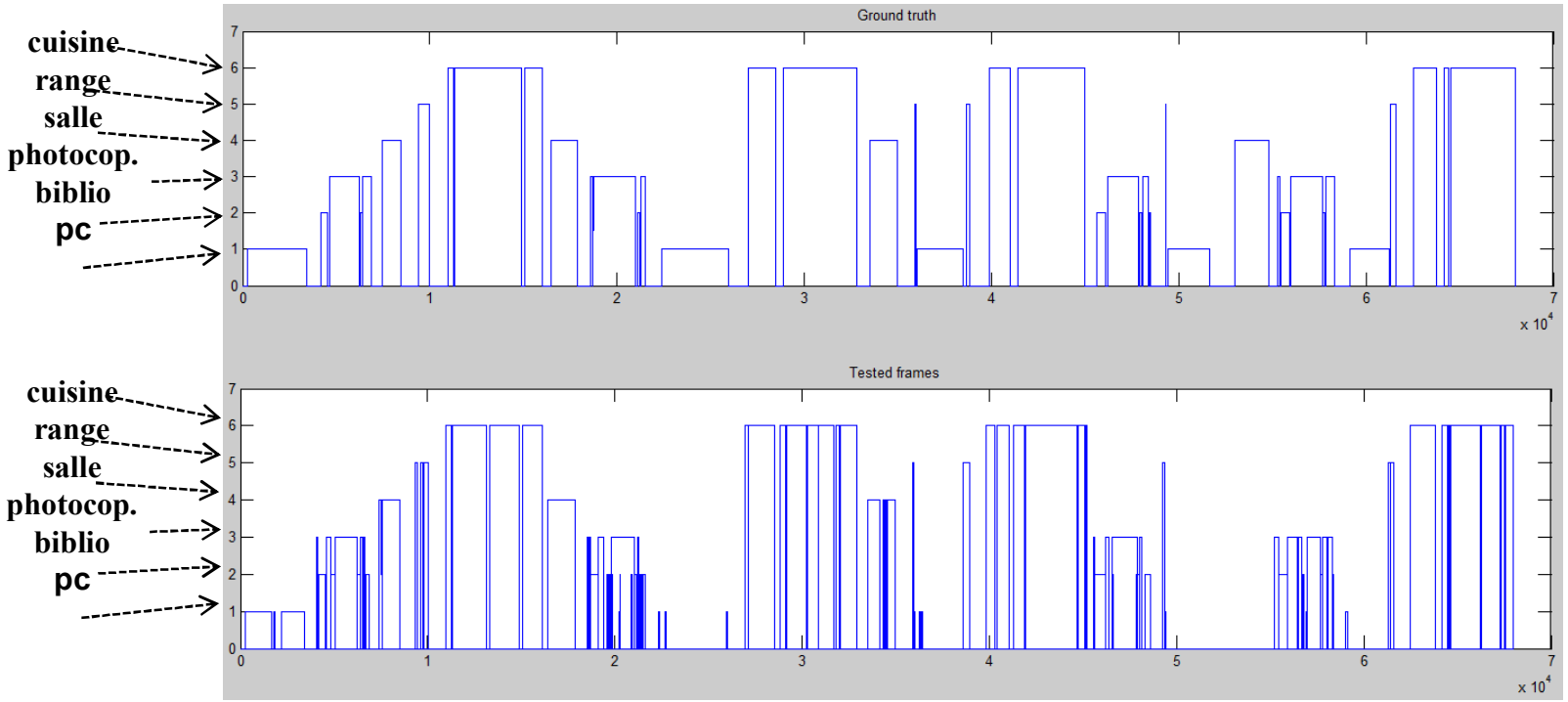



P = 0.9074

R = 0.5816

# Recognition performance

Majority filtering over 100 frames



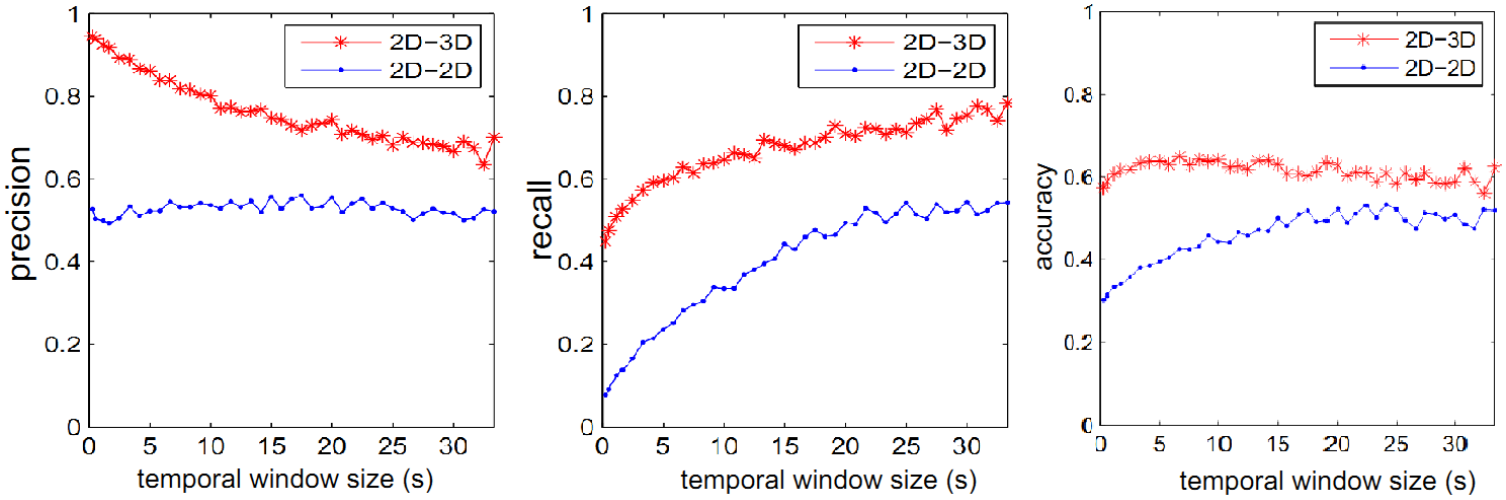

P = 0.8813

R = 0.6771

# Interest of 3D model for indexing

2D-3D matching: matching between test frames and 3D models

2D-2D matching: direct match between test frames and reference frames




(-) Overhead in building 3d model

(+) Much better precision/recall that purely 2D analysis

(+) Matching to consolidated model instead of large set of images

Visual positionning from wearable camera

# 5.2. OBJECT RECOGNITION WITH SALIENCY

- Many objects may be present in the camera field
- How to consider the object of interest?
- Our proposal: By using visual saliency

IMMED DB

# OUR APPROACH : MOLELING VISUAL ATTENTION

- Several approaches
  - Bottom-up or top-down
  - Overt or covert attention
  - Spatial or spatio-temporal
  - Scanpath or pixel-based saliency

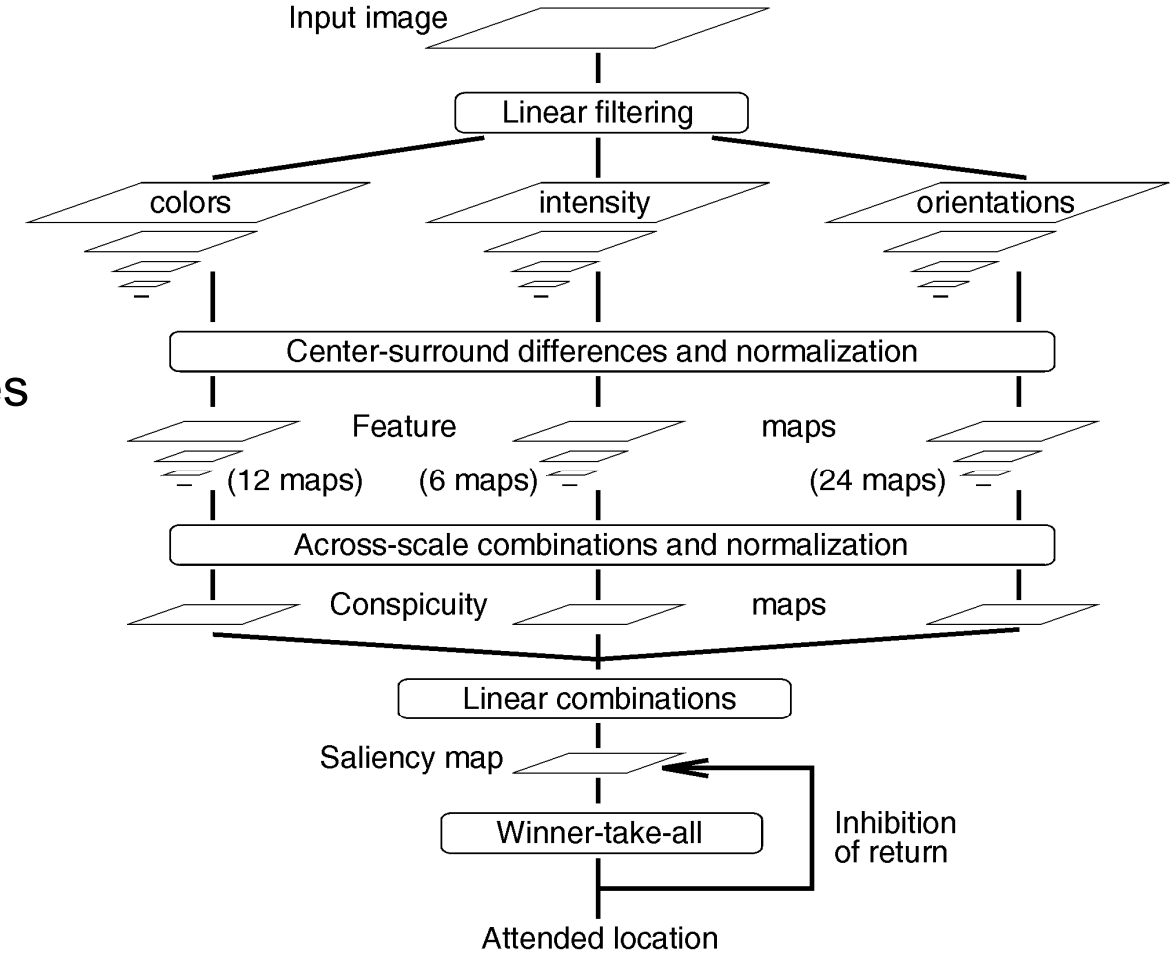
- Features
  - Intensity, color, and orientation (Feature Integration Theory [1]), HSI or L*a*b* color space
  - Relative motion [2]

- Plenty of models in the literature
  - In their 2012 survey, A. Borji and L. Itti [3] have taken the inventory of 48 significant visual attention methods

[1] Anne M. Treisman & Garry Gelade. A feature-integration theory of attention. Cognitive Psychology, vol. 12, no. 1, pages 97–136, January 1980.
[2] Scott J. Daly. Engineering Observations from Spatiovelocity and Spatiotemporal Visual Models. In IS&T/SPIE Conference on Human Vision and Electronic Imaging III, volume 3299, pages 180–191, 1 1998.
[3] Ali Borji & Laurent Itti. State-of-the-art in Visual Attention Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 99, no. PrePrints, 2012.

# ITTI'S MODEL

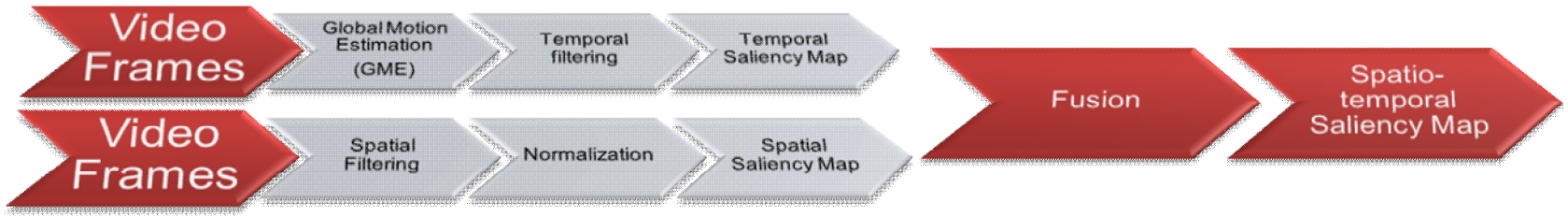
- The most widely used model
- Designed for still images
- Does not consider the temporal dimension of videos




[1] Itti, L.; Koch, C.; Niebur, E.; , "A model of saliency-based visual attention for rapid scene analysis , » *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol.20, no.11, pp.1254-1259, Nov 1998

# SPATIOTEMPORAL SALIENCY MODELING

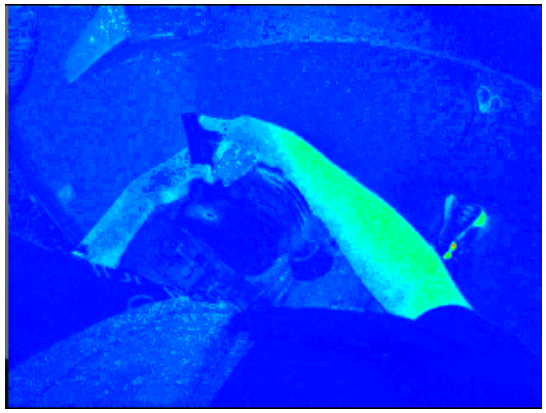
- Most of spatio-temporal bottom-up methods work in the same way[1], [2]
  - Extraction of the spatial saliency map (static pathway)
  - Extraction of the temporal saliency map (dynamic pathway)
  - Fusion of the spatial and the temporal saliency maps (fusion)




[1] Olivier Le Meur, Patrick Le Callet & Dominique Barba. Predicting visual fixations on video based on low-level visual features. Vision Researchearch, vol. 47, no. 19, pages 2483–2498, Sep 2007.
[2] Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin & Anne Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. International Journal of Computer Vision, vol. 82, no. 3, pages 231–243, 2009. Département Images et Signal.

# SPATIAL SALIENCY MODEL

- Based on the sum of 7 color contrast descriptors in HSI domain [1][2]
  - Saturation contrast
  - Intensity contrast
  - Hue contrast
  - Opposite color contrast
  - Warm and cold color contrast
  - Dominance of warm colors
  - Dominance of brightness and hue



The 7 descriptors $V_\delta$ are computed for each pixels $s_i$ of a frame I using the 8 connected neighborhood.

The spatial saliency map $S^{SP}$ is computed by: $S^{SP}(s_i) = \frac{1}{7}\sum_{\delta=1}^{7} V_\delta(s_i)$

Finally, $S^{SP}$ is normalized between 0 and 1 according to its maximum value $S_{max}$

$$S^{SP'}(s_i) - S^{SP}(s_i)/S_{max}$$

[1] M.Z. Aziz & B. Mertsching. Fast and Robust Generation of Feature Maps for Region-Based Visual Attention. Image Processing, IEEE Transactions on, vol. 17, no. 5, pages 633 –644, may 2008.
[2] Olivier Brouard, Vincent Ricordel & Dominique Barba. Cartes de Saillance Spatio-Temporelle basées Contrastes de Couleur et Mouvement Relatif. In Compression et representation des signaux audiovisuels, CORESA 2009, page 6 pages, Toulouse, France, March 2009.

# TEMPORAL SALIENCY MODEL

The temporal saliency map is extracted in 4 steps [Daly 98][Brouard et al. 09][Marat et al. 09]

The optical flow is computed for each pixel $s_i$ of frame i.

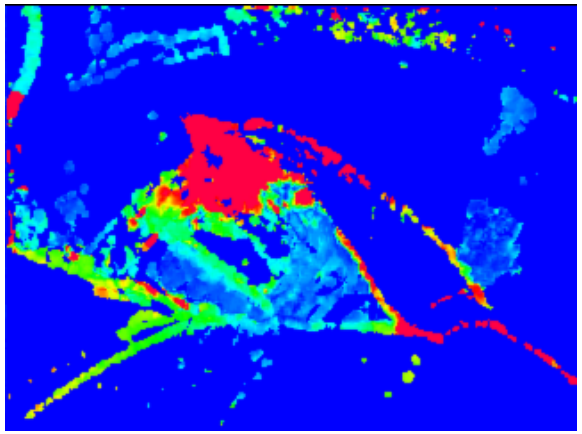
The motion is accumulated in $\overrightarrow{V_\theta}(s_i)$ and the global motion $\overrightarrow{V_G}(s_i)$ is estimated.

The residual motion is computed:

$$\overrightarrow{V_R}(s_i) = \overrightarrow{V_\theta}(s_i) - \overrightarrow{V_G}(s_i)$$

Finally, the temporal saliency map $S^T(s_i)$ is computed by filtering the amount of residual motion in the frame.

$$S^T(s_i) = \begin{cases} \frac{1}{7}\overrightarrow{V_R}(s_i) & if \quad 0 \le \overrightarrow{V_R}(s_i) < \vec{v}_1 \\ 1 & if \quad \vec{v}_1 \le \overrightarrow{V_R}(s_i) < \vec{v}_2 \\ \frac{1}{60}\overrightarrow{V_R}(s_i) + \frac{8}{5} & if \quad \vec{v}_2 \le \overrightarrow{V_R}(s_i) < \vec{v}_{max} \\ 0 & if \quad \overrightarrow{V_R}(s_i) \ge \vec{v}_{max} \end{cases}$$

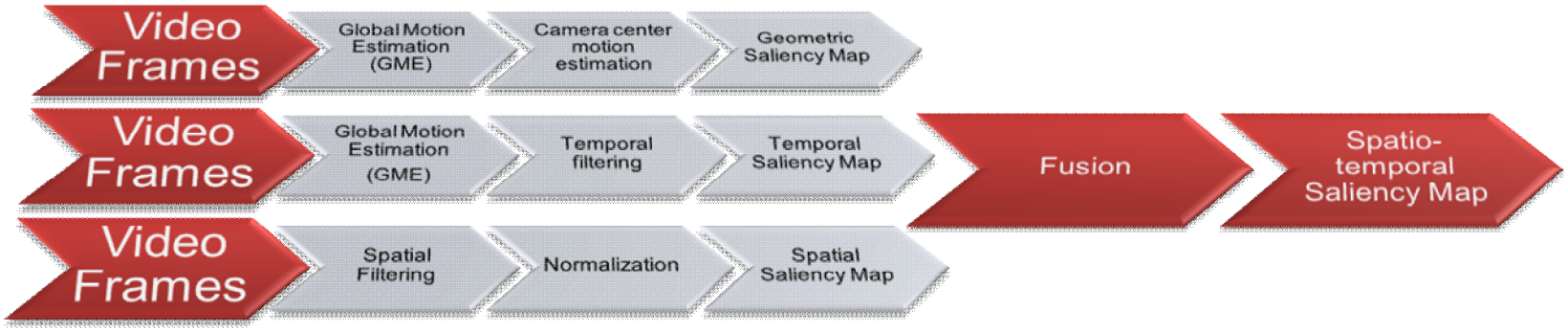


with $\vec{v}_1 = 6\ deg./s$ , $\vec{v}_2 = 30\ deg./s$ and $\vec{v}_{max} = 80\ deg./s$

# SALIENCY MODEL IMPROVEMENT

- Spatio-temporal saliency models were designed for edited videos

- Not well suited for unedited egocentric video streams
- Our proposal:
  - Add a geometric saliency cue that considers the camera motion anticipation




1. H. Boujut, J. Benois-Pineau, and R. Megret. Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion. In A. Fusiello, V. Murino, and R. Cucchiara, editors, Computer Vision ECCV 2012, IFCV WS

# GEOMETRIC SALIENCY MODEL

- 2D Gaussian was already applied in the literature [1]
  - “Center bias”, Busswel, 1935 [2]
  - Suitable for edited videos
- Our proposal:
  - **Train the center position as a function of camera position**
  - **Move the 2D Gaussian center according to camera center motion.**
  - Computed from the global motion es $dx_i = a_1 + a_2 x + a_3 y$ $dy_i = a_4 + a_5 x + a_6 y$
  - Considers the anticipation phenomenon [Land et al.].

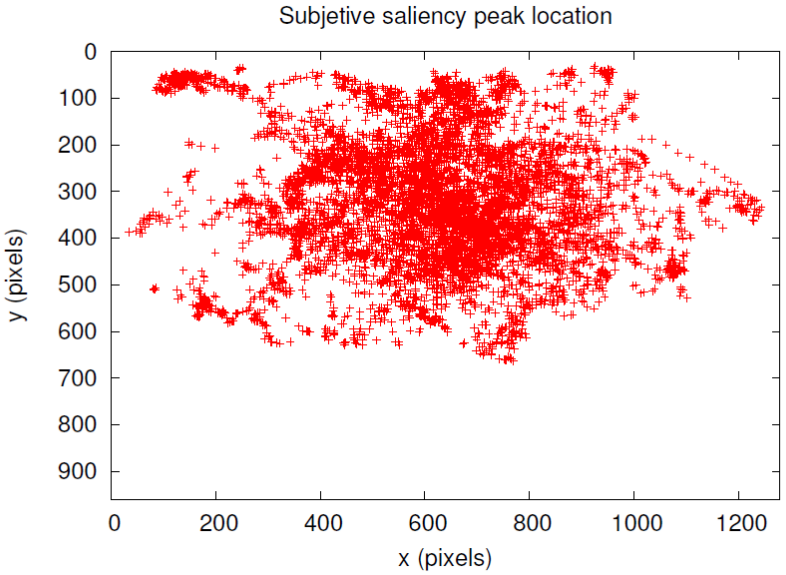
Geometric saliency map

[1] Tilke Judd, Krista A. Ehinger, Frédo Durand & Antonio Torralba. Learning to predict where humans look. In ICCV, pages 2106–2113. IEEE, 2009.
[2] Michael Dorr, et al. Variability of eye movements when viewing dynamic natural scenes. Journal of Vision (2010), 10(10):28, 1-17

# GEOMETRIC SALIENCY MODEL

- The saliency peak is never located on the visible part of the shoulder
- Most of the saliency peaks are located on the 2/3 at the top of the frame
- So the 2D Gaussian center is set at:

$$x_0 = \frac{width}{2} \qquad y_0 = \frac{height}{3}$$



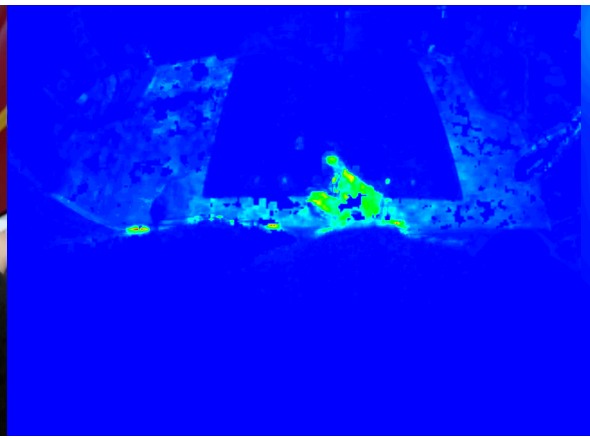

Saliency peak on frames from all videos of the eye-tracker experiment

$$S_g(I) = e^{-\left(\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2}\right)}$$

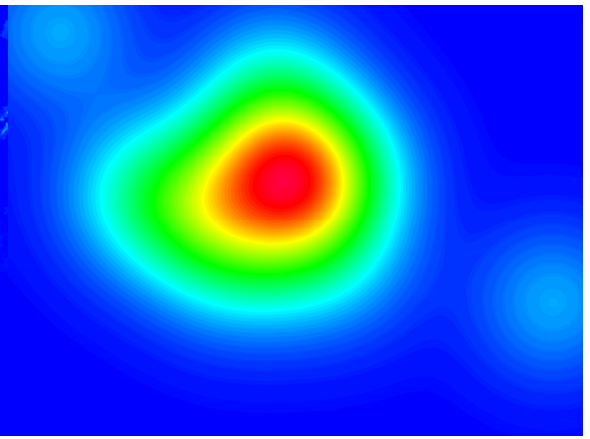
$$x = \frac{width}{2} \qquad y = \frac{height}{2}$$

# SALIENCY FUSION

Frame



Spatio-temporal-geometric saliency map



Subjective saliency map

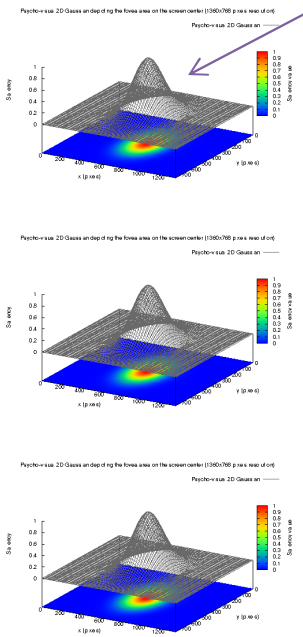
# SUBJECTIVE SALIENCY

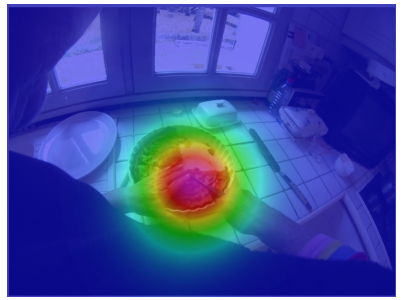
D. S. Wooding method, 2002

(was tested over 5000 participants)

Eye fixations
from the eye-tracker

2D Gaussians
(Fovea area = 2° spread)





Subjective
saliency map

# HOW PEOPLE WATCH VIDEOS FROM WEARABLE CAMERA?

- Psycho-visual experiment

- Gaze measure with an Eye-Tracker (Cambridge Research Systems Ltd. HS VET 250Hz)

- 31 HD video sequences from IMMED database.
- Duration 13’30’’
- 25 subjects (5 discarded)
- 6 562 500 gaze positions recorded
- We noticed that subject anticipate camera motion

# EVALUATION ON IMMED DB

Normalized Saliency Scanpath (NSS) correlation method Metric

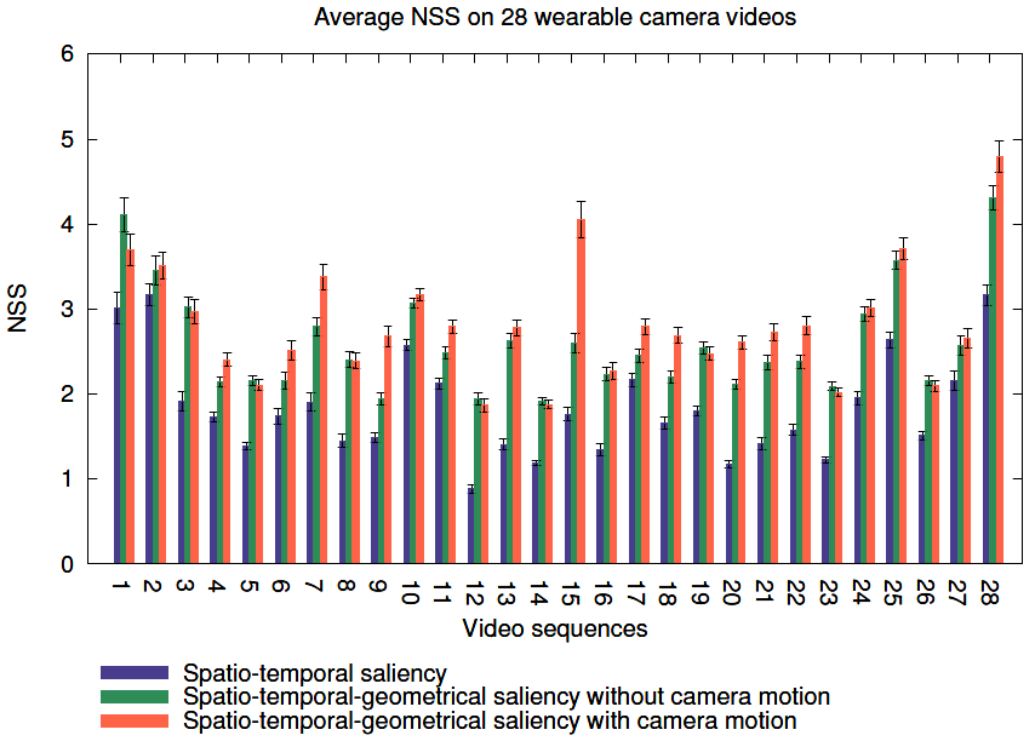
$$NSS_l = \frac{\overline{S_{subj_l} \times S^N_{obj_l}} - \overline{S_{obj_l}}}{\sigma\left(S_{obj_l}\right)}$$

Comparison of:

- Baseline spatio-temporal saliency
- Spatio-temporal-geometric saliency **without** camera motion
- Spatio-temporal-geometric saliency **with** camera motion

Results:

- **Up to 50% better than spatio-temporal saliency**
- **Up to 40% better than spatio-temporal-geometric saliency without camera motion**




H. Boujut, J. Benois-Pineau, R. Megret: « Fusion of Multiple Visual Cues for Visual Saliency Extraction from Wearable Camera Settings with Strong Motion ». ECCV Workshops (3) 2012: 436-445

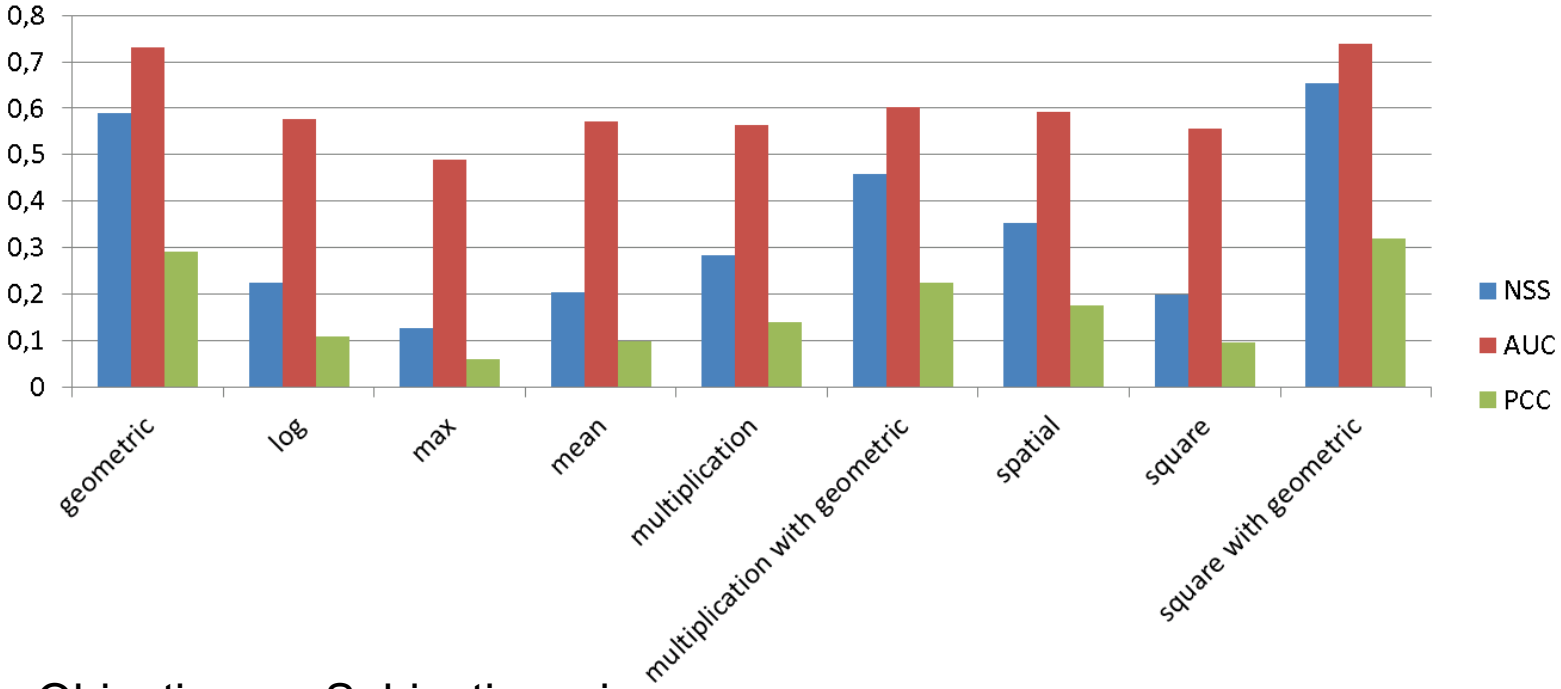
# EVALUATION ON GTEA DB

- IADL dataset
- 8 videos, duration 24’43’’
- Eye-tracking mesures for actors and observers
  - Actors (8 subjects)
  - Observers (31 subjects) 15 subjects have seen each video
- SD resolution video at 15 fps recorded with eyetracker glasses

A. Fathi, Y. Li, and J. Rehg. Learning to recognize daily actions using gaze. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, Computer Vision ECCV 2012

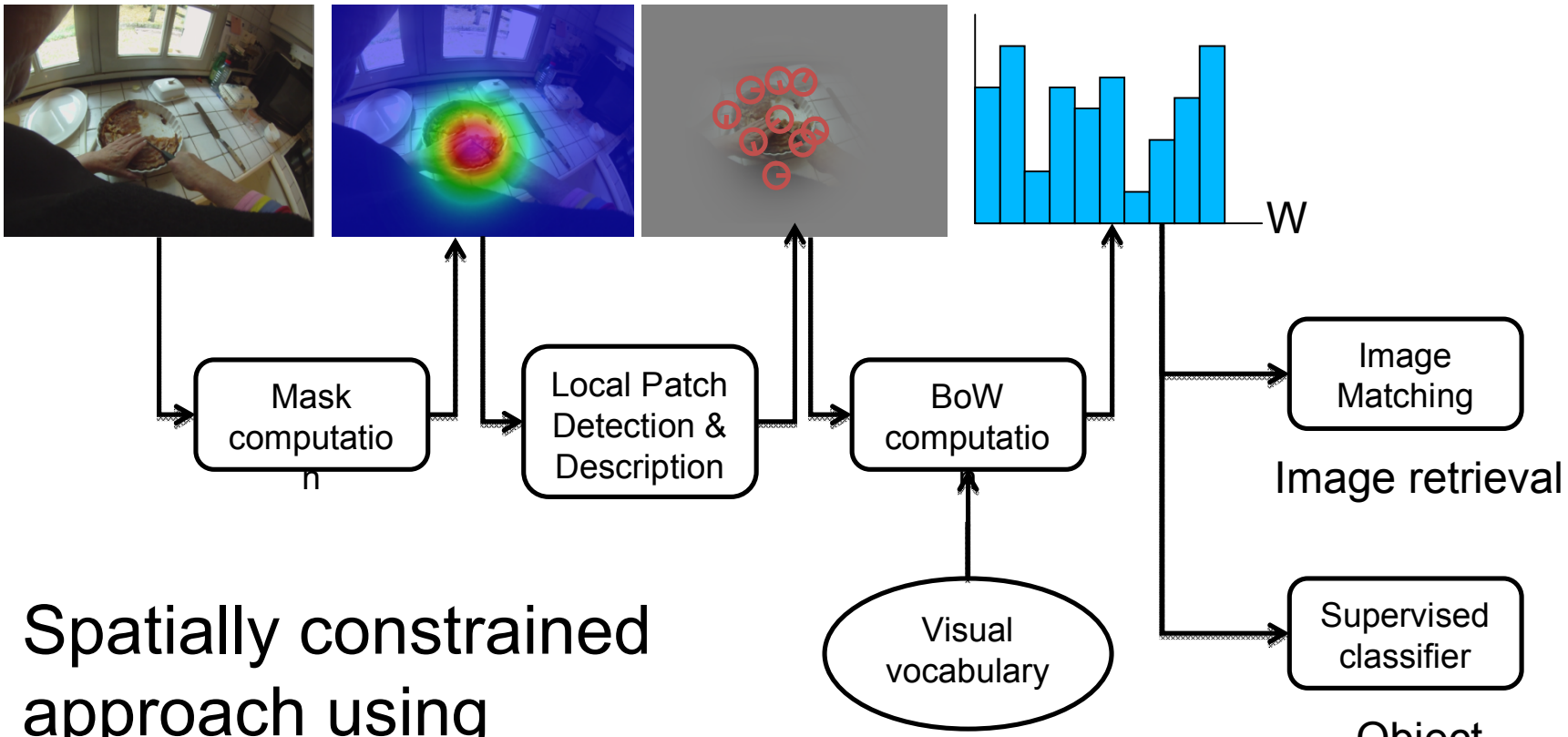
# EVALUATION ON GTEA DB

- “Center bias” : camera on looking glaces, head mouvement compensates
- Correlation is database dependent




Objective vs. Subjective -viewer

# PROPOSED PROCESSING PIPELINE FOR OBJECT RECOGNITION




Spatially constrained approach using saliency methods

# GTEA DATASET (CONSTRAINED SCENARIO)

- GTEA [1] is an ego-centric video dataset, containing 7 types of daily activities performed by 14 different subjects.
- The camera is mounted on a cap worn by the subject.
- Scenes show 15 object categories of interest.
- Data split into a training set (294 frames) and a test set (300 frames)

| appetizers | bowl | brioche | brioches_bag | cutlery |
|---|---|---|---|---|
| 9/9 | 12/12 | 90/89 | 14/15 | 15/16 |
| glass | ham_closed | jam | milk | peanut_butter |
| 9/9 | 12/12 | 20/21 | 14/15 | 30/30 |
| plate | slice_of_cheese | stack_of_cheese | stack_of_plates | vegetables |
| 13/13 | 17/18 | 9/10 | 10/11 | 20/20 |

Categories in GTEA dataset with the number of positives in train/test sets

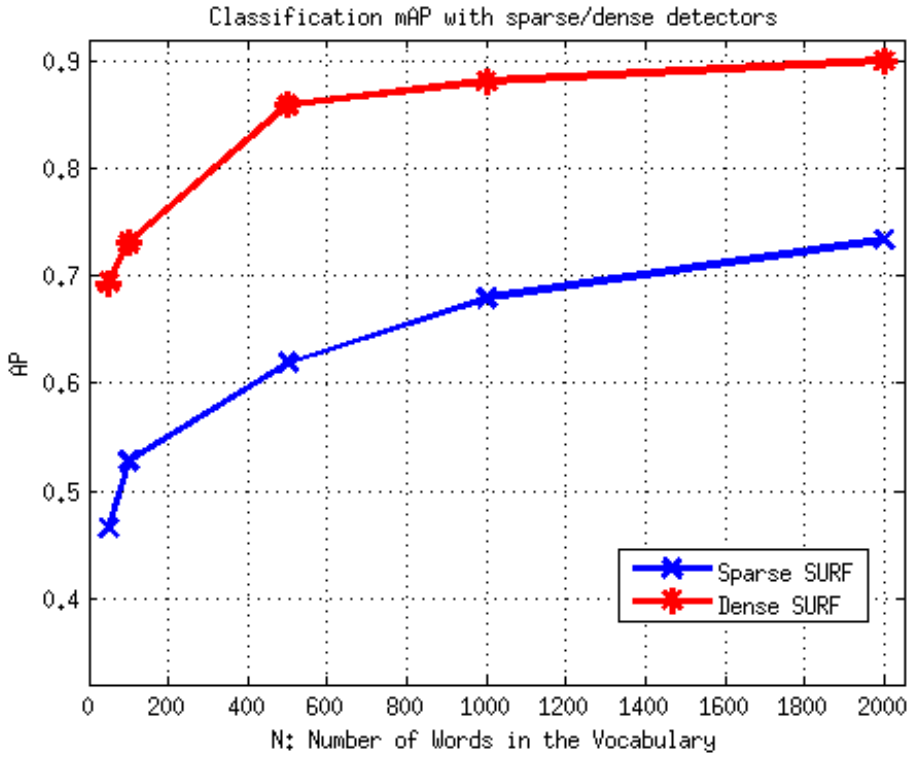
[1] Alireza Fathi, Yin Li, James M. Rehg, Learning to recognize daily actions using gaze, *ECCV 2012*.

# ASSESSMENT OF VISUAL SALIENCY IN OBJECT RECOGNITION

We tested various parameters of the model:

- Various approaches for local region sampling.
    1. Sparse detectors (SIFT, SURF)
    2. Dense Detectors (grids at different granularities).
- Different options for the spatial constraints:
    1. Without masks: global BoW
    2. Ideal manually annotated masks.
    3. Saliency masks: geometric, spatial, fusion schemes…
- In two tasks:
    1. Object retrieval (image matching): mAP ~ 0.45
    2. Object recognition (learning): mAP ~ 0.91

# SPARSE VS DENSE DETECTION



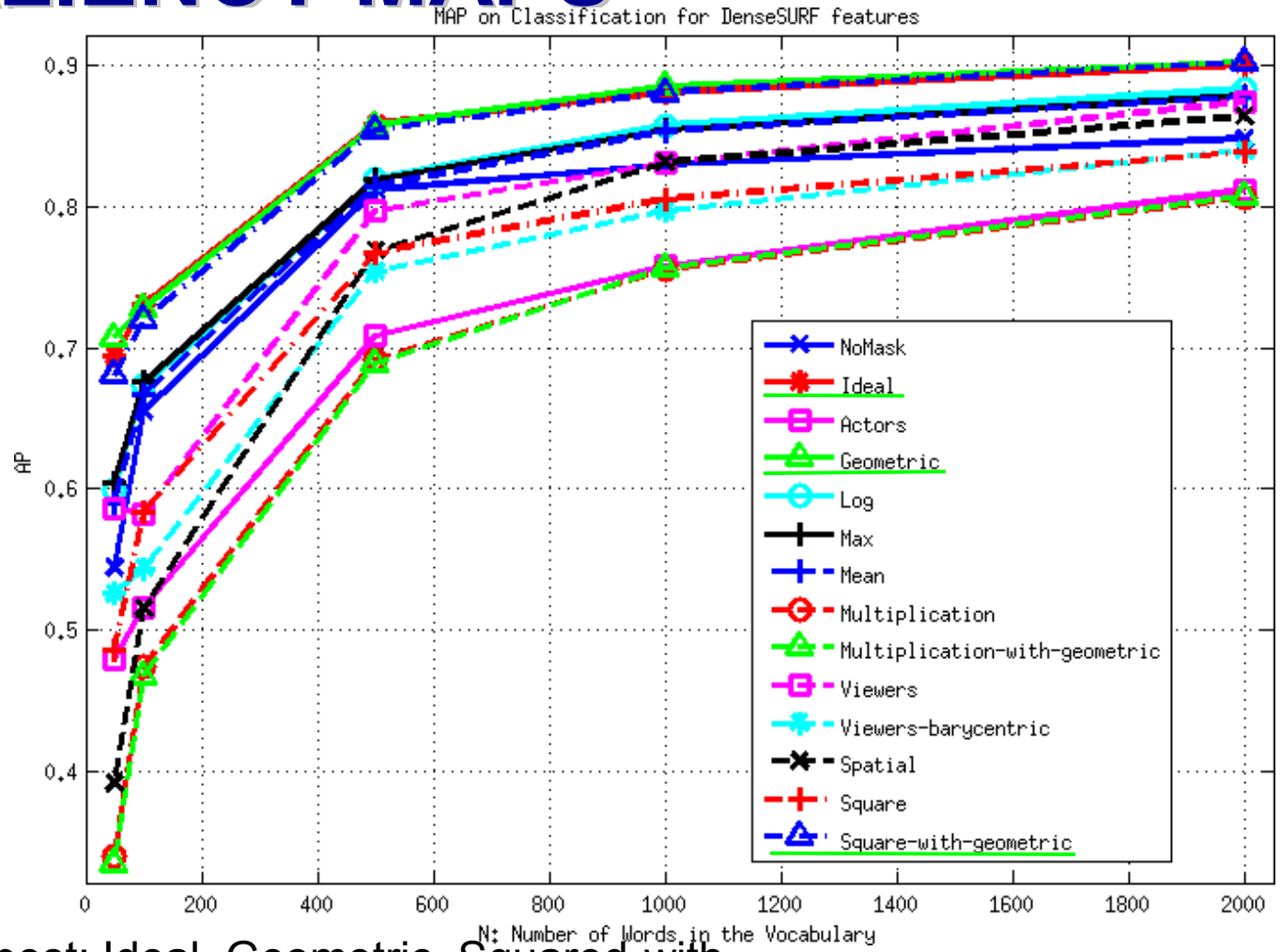

| Detector + Descriptor | Avg Pt Nbr |
|---|---|
| Sparse SURF + SURF | 137.9 |
| Dense + SURF | 1210 |

Experimental setup:

- GTEA database.
- 15 categories.
- SURF descriptor.
- Object recognition task.
- SVM with $\chi 2$ kernel.
- Using ideal masks.

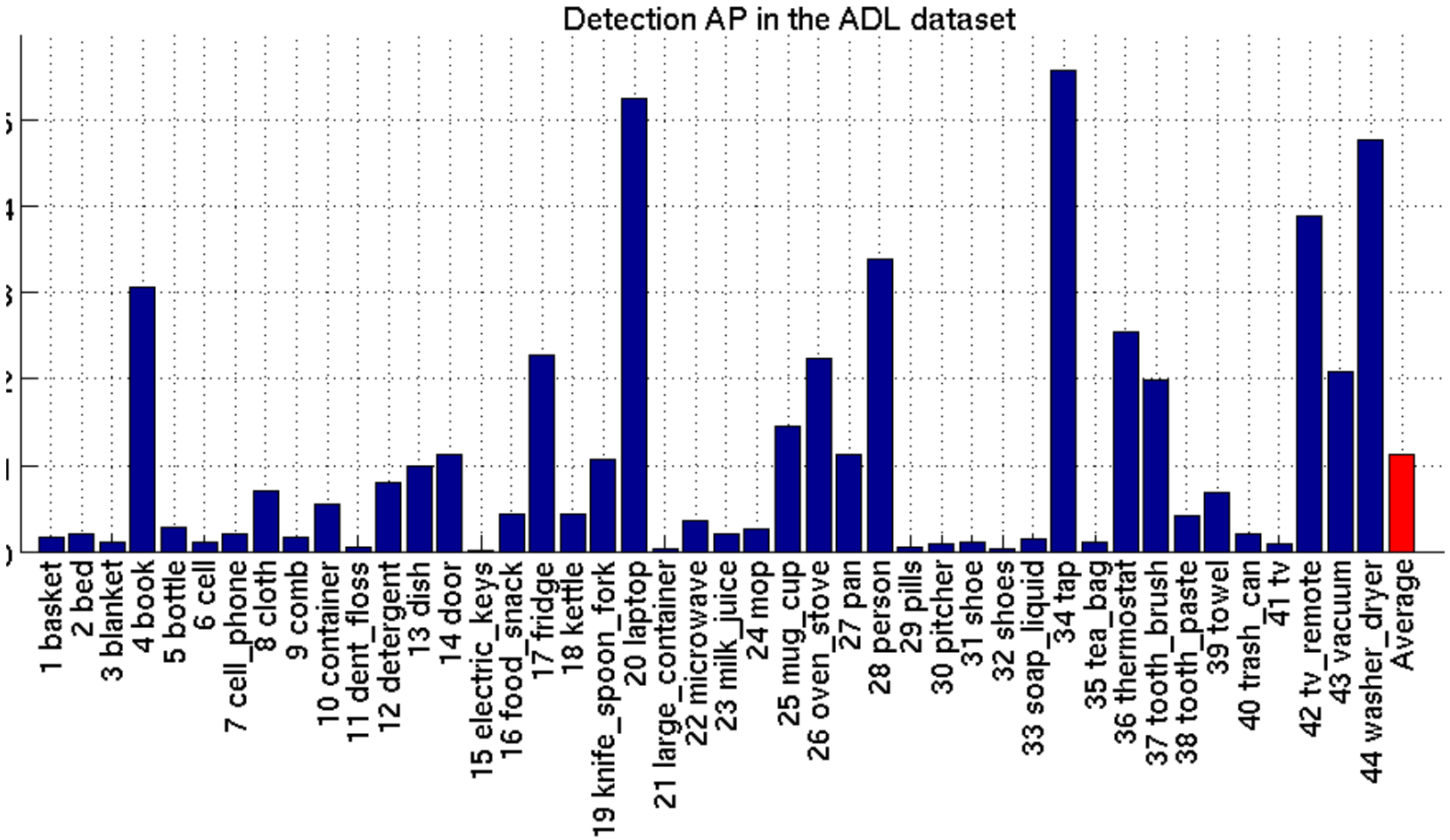
Dense detection greatly improves the performance at the expense of an increase on the computation time (more points to be processed)

# OBJECT RECOGNITION WITH SALIENCY MAPS




The best: Ideal, Geometric, Squared-with-geometric,

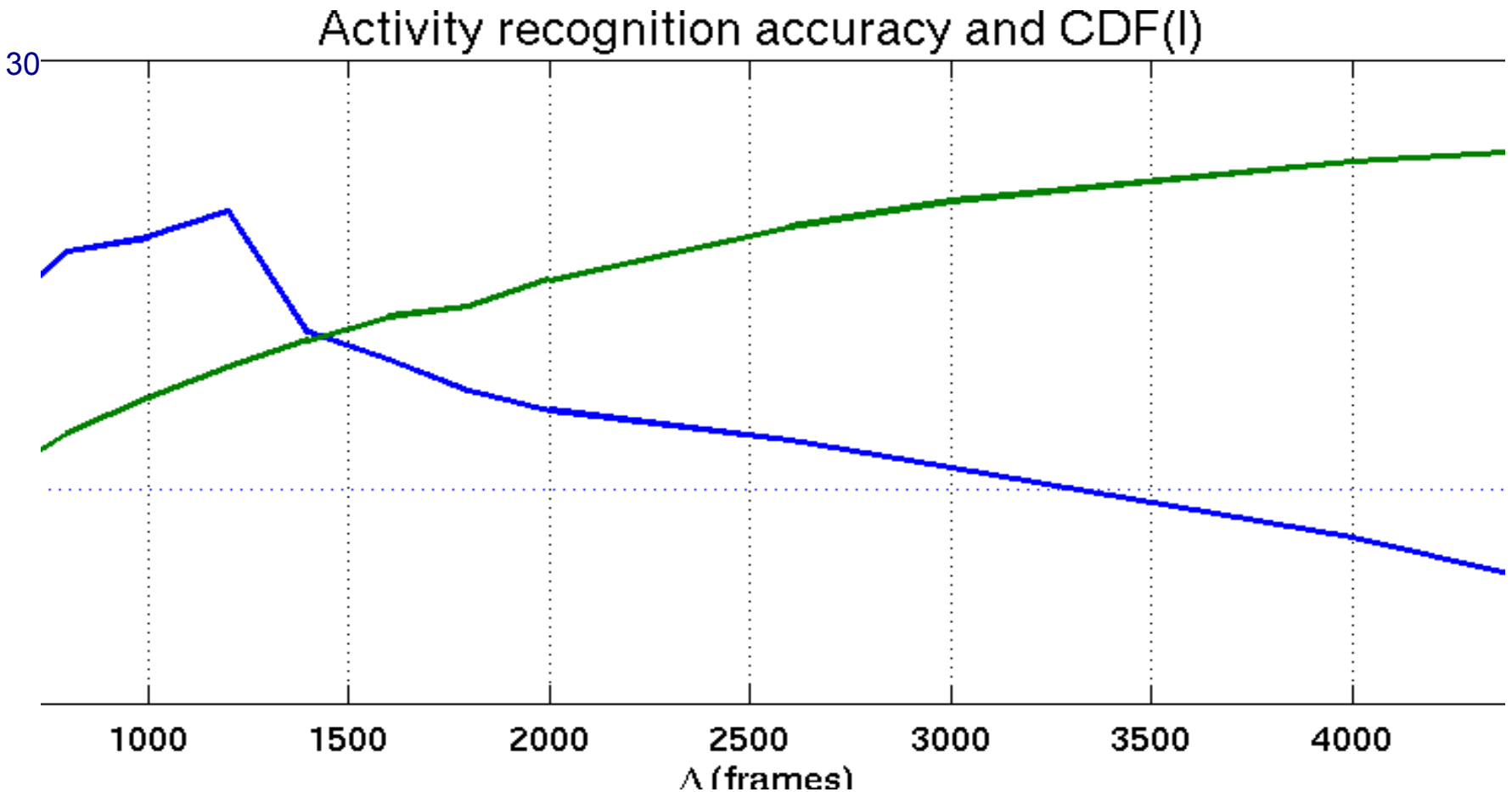
# Object recognition on ADL dataset (unconstrained scenario)



# Approach

Temporal pyramid matching

- Features :
  - probability vector of objects *O*
  - probability vector of places (2D approach) *P*
  - Early fusion by concatenation *F*=*O*+*P*


Activity recognition accuracy and CDF(l)
30
1000
1500
2000
2500
3000
3500
4000
Λ (frames)

# Resultas of activity recognition (ADL)

Activity recognition accuracy for our approach computed at Frame and Segment level, respectively

| Approach | Avg Frame Accuracy | Avg Segment Accuracy |
|---|---|---|
| Our Objects | 24.2 | 40.5 |
| Our Places | 19.7 | 11.1 |
| Our Objects + Places | **26.9%** | **41.3%** |
| Pirsiavash et al. | 23 | 36.9 |

# 6. CONCLUSION AND PERSPECTIVES(1)

- Spatiotemporal saliency correlation is database dependent for egocentric videos
  - Need A VERY GOOD motion estimation ( now Ransac)

- Saliency maps are a good ROIs for object recognition in the task of interpretation of wearable video content.
- The model of activities as a combination of objects and location is promising

# Conclusion and Perspectives (2)

Integration of 3D localization

Tests on Dem@care constrained and unconstrained datasets (Lab and Home)

Learning of « manupulated objects » vs « Active Objects ».