Scene Understanding and Activity Recognition

Francois BREMOND

INRIA Sophia Antipolis – STARS team

Institut National Recherche Informatique et Automatisme

Francois.Bremond@inria.fr

http://www-sop.inria.fr/members/Francois.Bremond/

CoBTeK,

Nice University Hospital

Video Understanding

Objective: Designing systems for Real time recognition of human activities observed by various sensors (especially video cameras).

Challenge: Bridging the gap between numerical sensors and semantic events.

Approach: Spatio-temporal reasoning and knowledge management.

Examples of human activities:

for individuals *(graffiti, vandalism, bank attack, cooking)*

for small groups *(fighting)*

for crowd *(overcrowding)*

for interactions of people and vehicles *(aircraft refueling)*

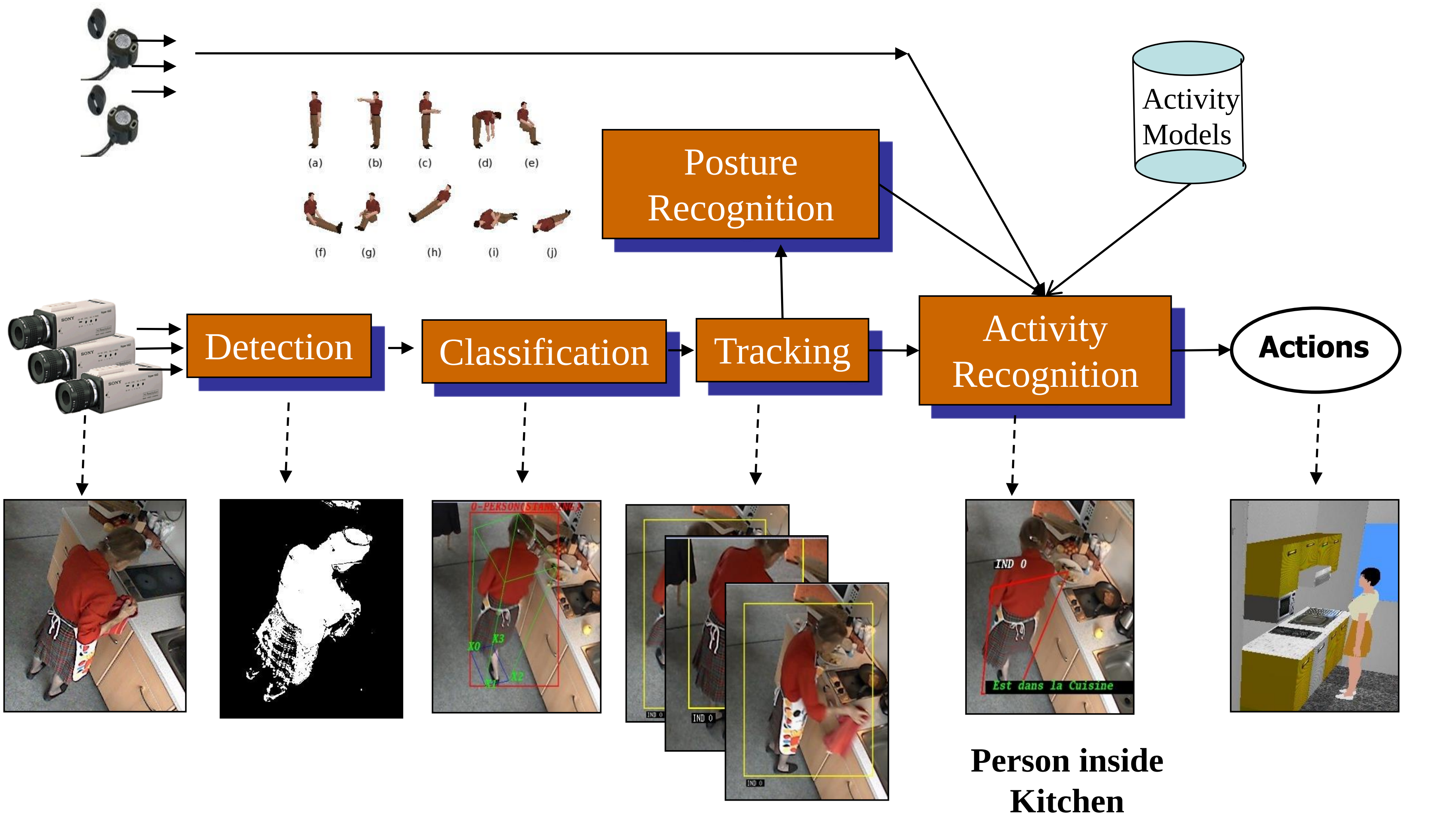
Video Understanding: Issues

Practical issues

- Video Understanding systems have poor performances over time, can be hardly modified and do not provide semantics

F. Porikli, et al., Video Surveillance: Past, Present, and Now the Future, IEEE-SPM forum, 30 (3) May 2013.

Generic Platform for activity understanding

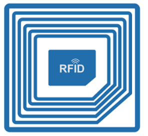


Background Subtraction & People Detection

Issues with People Detector:

- **Background subtraction:**
  - **Pros: Reducing processing time**
  - **Cons: Sensitive to illumination change, moving background, shadows, overlapping people…**
- **RGBD sensors**
  - **Pros:**
    - **Accurate human/head detector (occlusion)**
    - **Night and day (IR camera)**
    - **Privacy protection (Depth map)**
  - **Cons:**
    - **Sensitive to strong day light**
    - **Narrow field of view, accurate up to 4 meters**

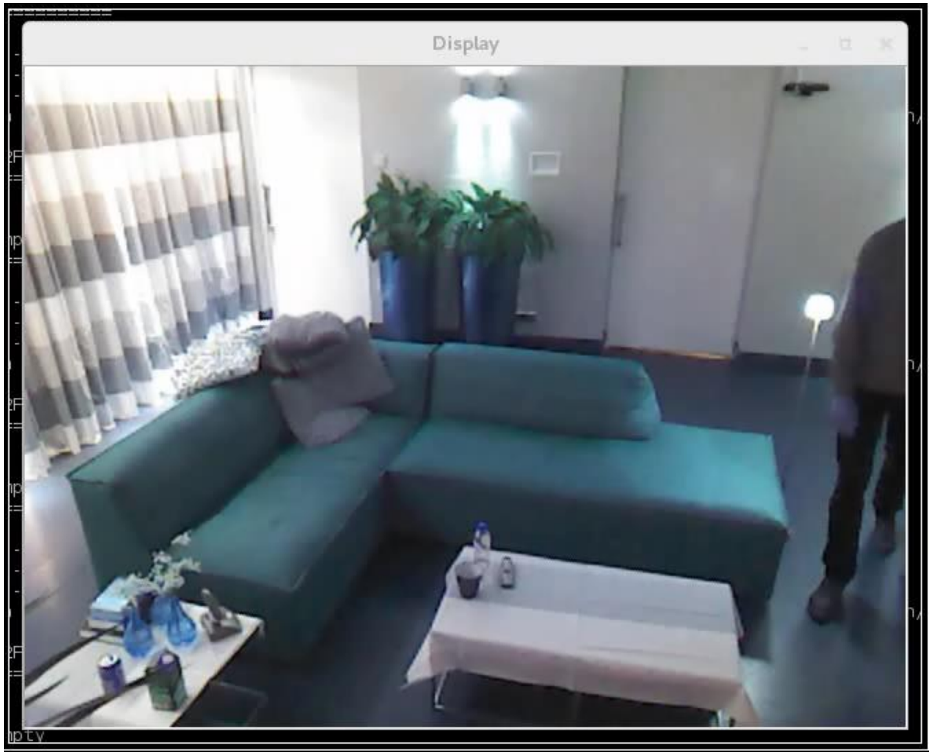
- **Wireless Sensors: beacon, smart-phone, RFID**
  - **Pros:**
    - **Human ID**
    - **Reliable (no lost ID track)**
  - **Cons:**
    - **Inaccurate (2 beacons define a zone of few meters), battery for 3 years**
    - **Cooperative (download an app on your cellular-phone, open your WiFi/Bluetooth)**
    - **Require WiFi hotspot - wireless LAN (WLAN) network, calibration step**

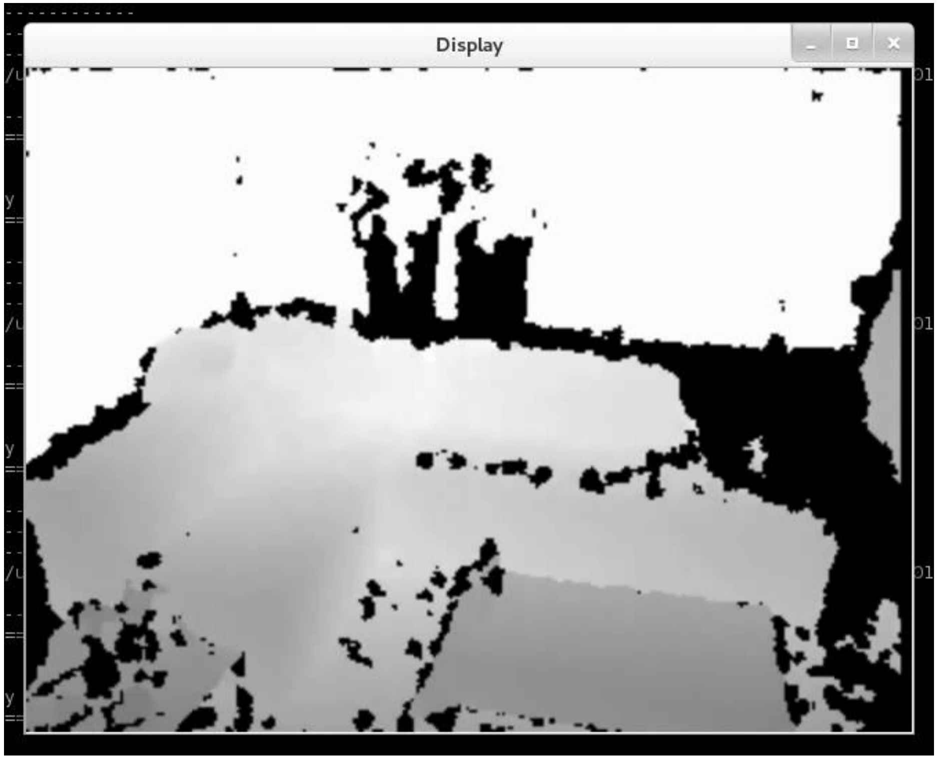


- **High Resolution, High Dynamic Range video cameras**
  - **Pros:**
    - **Accurate human/head detector (e.g. DPM, DCNN)**
    - **Inside/outside**
    - **GPU architecture**
  - **Cons:**
    - **Sensitive to training dataset**

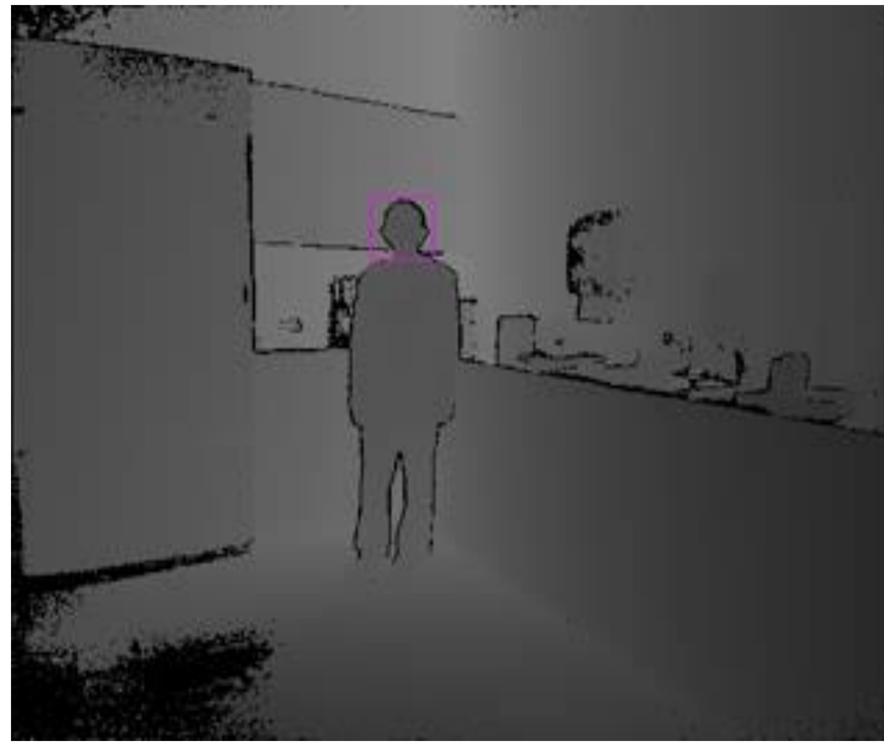
People/Head detection - Smart Room Dataset

Visualization of head detection.





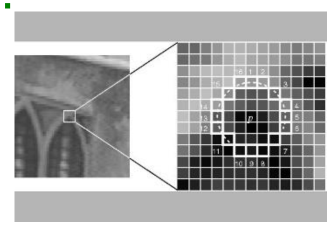
Head detection - Cornell University's kitchen dataset

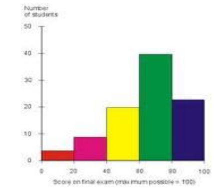


Pink : Skeleton
Red : Nghiem's result
Green : Our result

Background Subtraction & People Detection

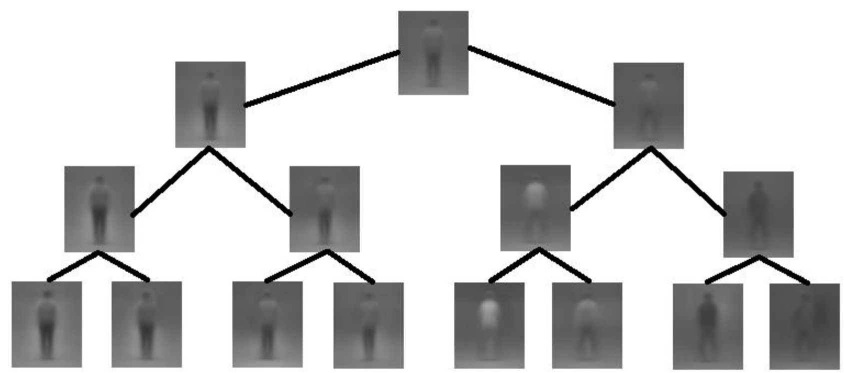
Issues with Local Descriptor for People Classifier:





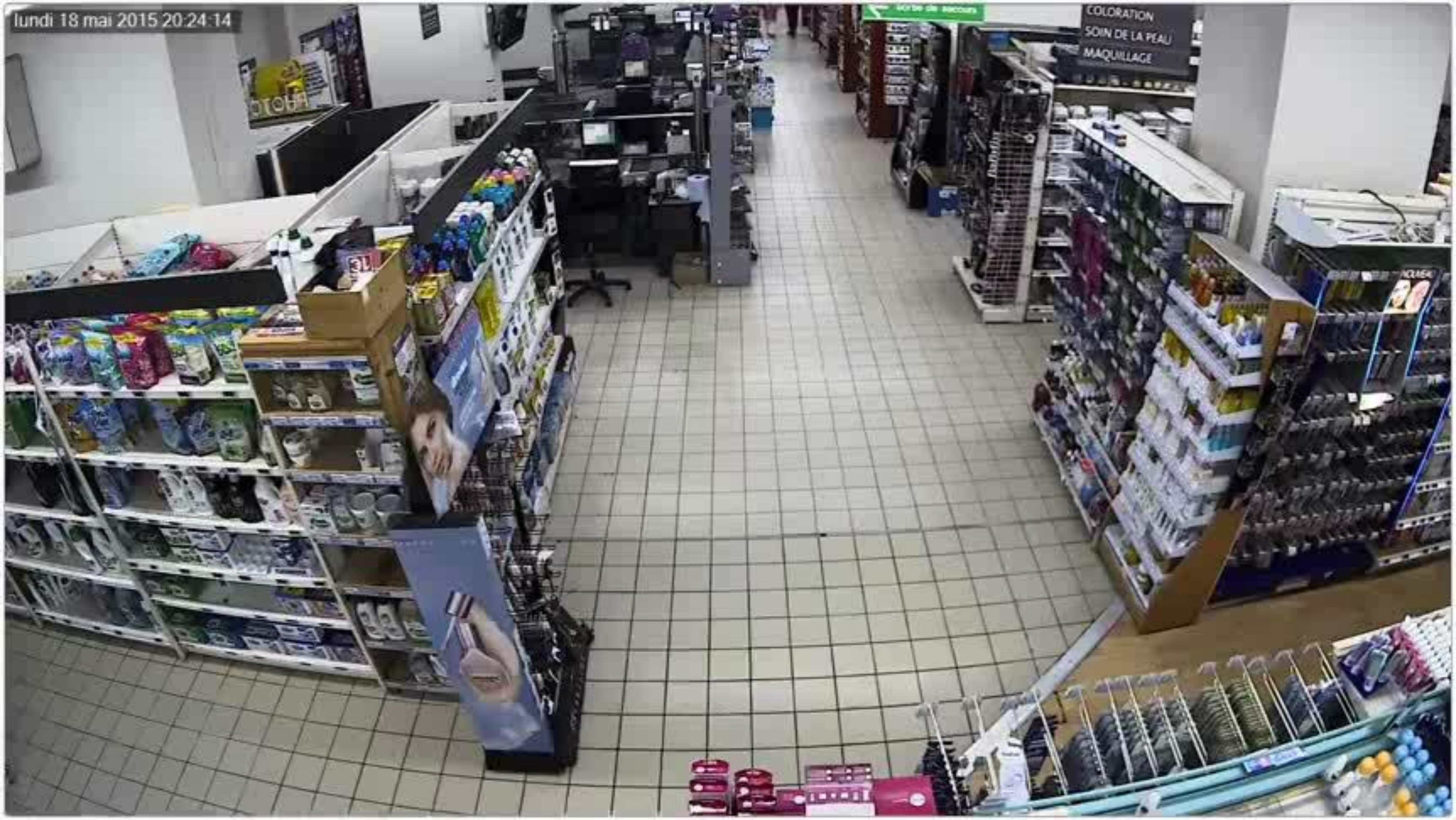
- **Features:**
  - HOG, LBP, Covariance Matrix, Haar, SIFT, Granules, deep features (DCNN)
- **Learning paradigm:**
  - Adaboost, Hierarchical trees, ensembles of SVM
- **<u>Training / testing databases:</u>**
  - Camera view point, distortion, resolution,
  - Occlusion, pose,
  - Background samples
  - Clustering the positive and negative samples

- **Processing time:**
  - Training (best feature selection)
  - Detection (scanning window sampling rate, multi-resolution)
- **Filtering:**
  - Overlapping scanning window, candidate selection
  - 3D constraint, motion segmentation (background subtraction),
- **Body parts:**
  - Global detection
  - Model based association, DPM
  - E.g. head, torso, legs ...



Scenario recognition: Retails

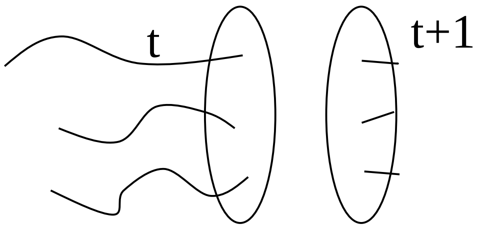
People detection and tracking using DPM on high resolution images



Tracking Parameter Control (Chau - Nguyen)

Multiple Objet Tracking in 2 steps:

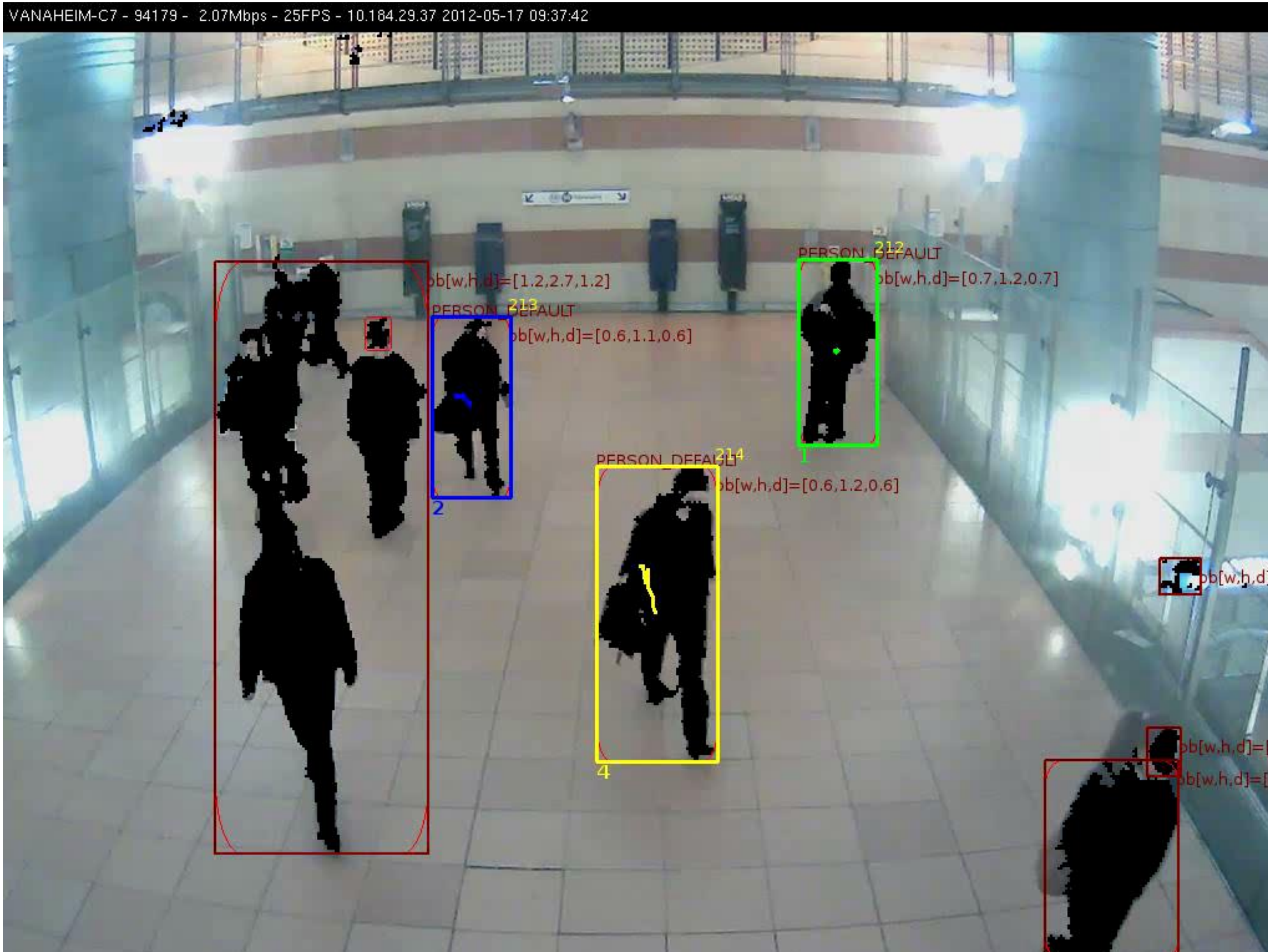
- Short term tracking: Object feature extraction and local data association between (t, t+1) to obtain short reliable tracklets
- Long term tracking: global association of tracklets through out the video



Two optimization techniques:

- Maximize the weights of the most **discriminant features** between a small set of object features
- Learn the optimal set of **tracking parameter** values :
  - Offline Learning of the best **parameters** for reference **videos** or **tracklets**
  - Online parameter tuning **retrieve** online the corresponding **parameters**

People detection and tracking



Video Understanding

- 3 types of **Human Activities** of interest and **Methods**:
  - Activities which can be well described or modeled by users (e.g. sitting)
  - ➢ Recognition engine using hand-crafted ontologies based on a priori knowledge (e.g. rules) predefined by users

  - Activities which can be collected by users through positive/negative samples representative of the targeted activities (e.g. falling)
  - ➢ Supervised learning methods based on positive/negative samples to build specific classifiers for the targeted activities

  - Rare activities which are unknown to the users and which can be observed only through large datasets (e.g. non motivated activities)
  - ➢ Unsupervised (fully automated) learning methods based on clustering of frequent activity patterns to discover new activity models

Action Recognition: supervised approaches

- Different Descriptors (STIP, HOG, HOF, MBHx,y…)
- Different Classifiers and Machine Learning Approaches (SVM, NN, BayesNet, statistical models …)
- Benefiting from well-clipped huge training sets, many approaches achieve reasonable performance and succeeded to improve SOTA

[1] Laptev and T. Lindeberg. Space-time interest points. In *ICCV* 2003

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld.Learning realistic human actions from movies *CVPR 2008*

[3] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories CVPR, June 2011

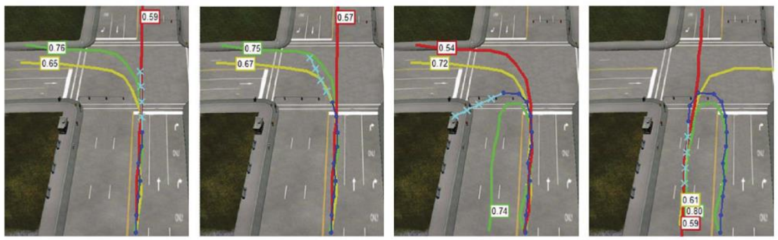
[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *2005. CVPR*

CONS

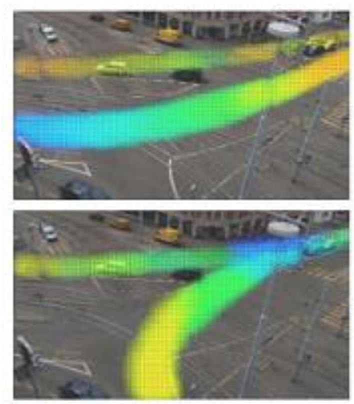
- Works good mostly on short term and well-clipped videos
- Localization problem in long videos (sliding window approaches)
- Doesn’t address complexity of composed motion like ADL, they are not really using the temporal relations of sub-events
- Needs annotation of large amount of data

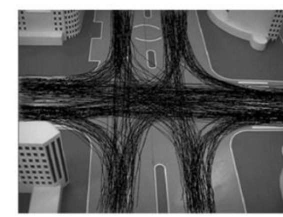
Action Recognition: unsupervised approaches

- Trajectory based
  - B. Morris and M. Trivedi. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach PAMI 2011
  - W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan
    A system for learning statistical motion patterns, PAMI2006

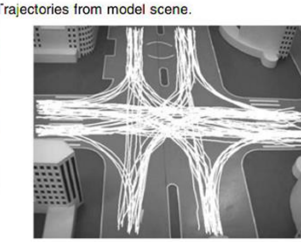


- Motion pattern
  - H. M. Dee, A. G. Cohn, and D. C. Hogg. Building semantic scene models from unconstrained video. *CVIU*, 2012
  - R. Emonet, J. Varadarajan, and J.-M. Odobez. Temporal Analysis of Motif Mixtures using Dirichlet Processes. *PAMI*, 2014





'rajectories from model scene.



CONS

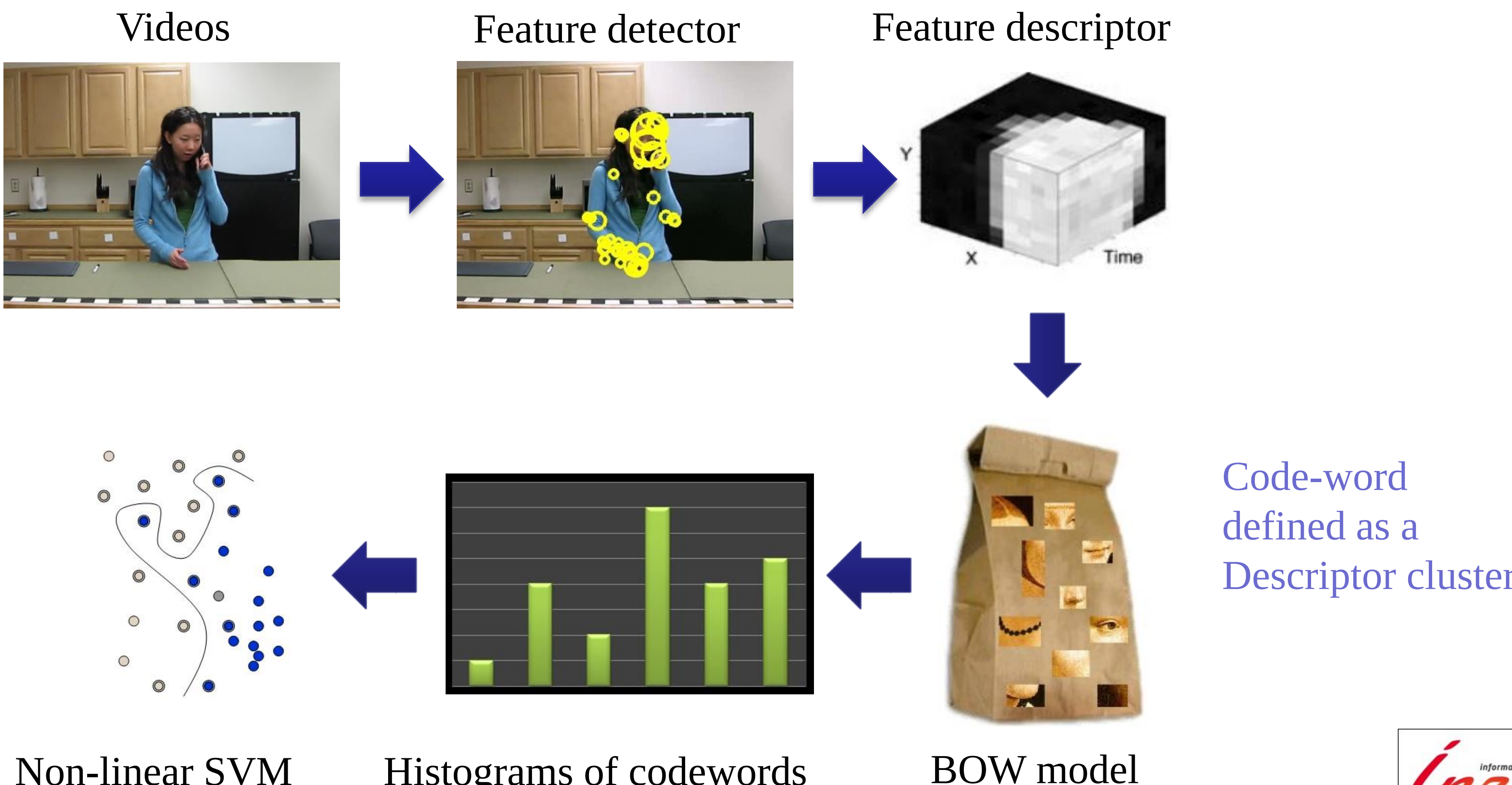
- Trajectories (global motion) cannot capture local motion patterns
- Since they use 2D motion pattern, there is no notion of persons and objects (semantics)
- Concurrency, works for traffic not for ADL
- Temporal and spatial structure required (repetitive events in traffic control)

Spatial and Temporal Localization

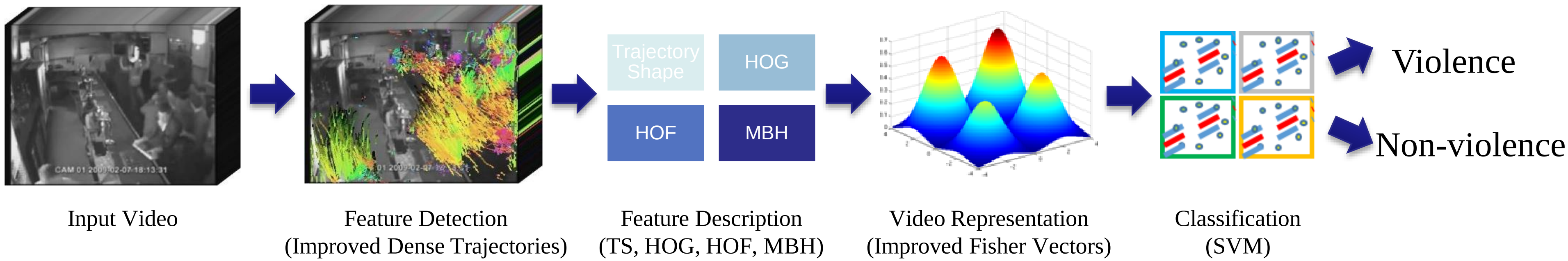
- Sliding window approaches, fixed-size clipping
  - Temporal segmentation [1]
  - Spatial segmentation [2,3]
  - Both [4,5]
- Problem: computationally expensive and therefore not appropriate for real-time activity recognition scenarios in real-world settings like long-term ADL

[1] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. *CVPR 2009*
[2] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff. Action recognition and localization by hierarchical space-time segments. *ICCV 2013*
[3] M. Jain, J. Van Gemert, H. Jegou, P. Bouthemy, and C. G. Snoek. Action localization with tubelets from motion. In *CVPR 2014*
[4] G. Willems, J. H. Becker, T. Tuytelaars, and L. J. Van Gool. Exemplar-based action recognition in video. In *BMVC* 2009
[5] K. Avgerinakis, A. Briassouli, and I. Kompatsiaris. Activity detection using sequential statistical boundary detection (ssbd). CVIU, 2015

Action Recognition using Bag of Words



Violence Recognition Framework, P. Bilinski



- We represent positions of local features in a video normalized manner, so that the video size does not significantly change the magnitude of the feature position vector.

$$p_i = \left[\frac{1}{v_w n_i} \sum_{j=1}^{n_i} x_{ij}, \frac{1}{v_h n_i} \sum_{j=1}^{n_i} y_{ij}, \frac{1}{v_l n_i} \sum_{j=1}^{n_i} t_{ij} \right]$$

- We also consider using the unity based normalization to reduce the influence of motionless regions at the boundaries of a video, so that the large motionless regions do not significantly change the magnitude of the feature position vector.

$$\bigvee_j max(p_{:j}) \neq \min(p_{:j}): \; p'_{ij} = \frac{p_{ij} - \min(p_{:j})}{max(p_{:j}) - \min(p_{:j})}$$

Local descriptor: | d_1 | d_2 | ... | $d_{k/2}$ | | | | d_k |

+ PCA: | d_1 | d_2 | ... | $d_{k/2}$ |

+ normalized spatial position: | x | y | z | d_1 | d_2 | ... | | $d_{k/2}$ |

p

Dataset: Violent-Flows (Crowd Violence \ Non-violence)

Violence

Pub

Steet

Steet

Football Stadium

Steet

Volleyball Arena

School

Movies Analysis

Non-violence

Football Stadium

Football Stadium

246 videos with real-world footage of crowd violence, collected from YouTube.
Variety of scenes, *e.g.* streets, football stadiums, volleyball and ice hockey arenas, and schools. 5-folds CV.

Violent-Flows: Results & Comparisons (MCA)

Results

Approach	Size	Accuracy (%)
Baseline	1	93.5
Ours: STIFV	~1	**96.4**
IFV 1×1×2	2	94.0
IFV 1×2×1	2	94.3
IFV 2×1×1	2	94.3
IFV 1×1×3	3	93.5
IFV 1×3×1	3	94.3
IFV 3×1×1	3	93.5
IFV 2×2×2	8	93.5
IFV 2×2×3	12	93.1
IFV 2×2×1	4	93.9
IFV 2×1×2	4	93.5
IFV 1×2×2	4	93.9

Comparison with the state-of-the-art

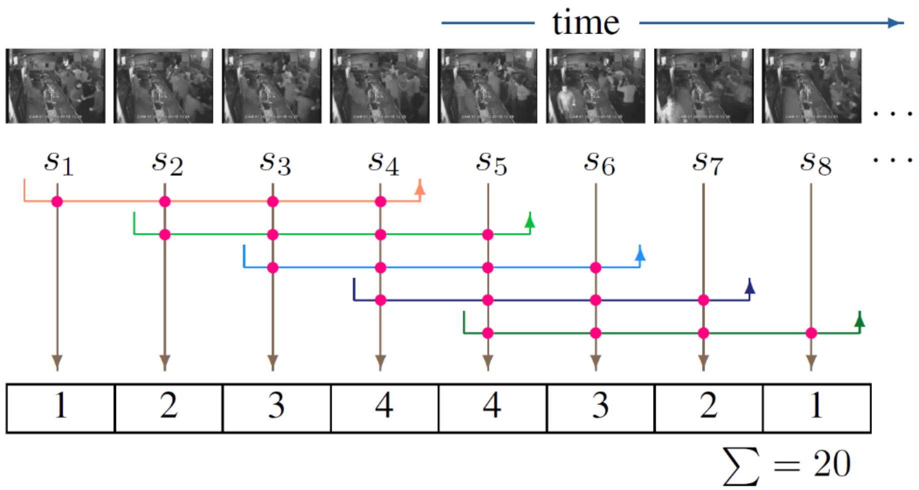
Approach	Accuracy (%)
HNF [Laptev *et al.*, CVPR'08]	56.5
HOG [Laptev *et al.*, CVPR'08]	57.4
HOF [Laptev *et al.*, CVPR'08]	58.3
LTP [Yeffet and Wolf, ICCV'09]	71.5
Jerk [Datta *et al.*, ICPR'02]	74.2
Interaction Force [Mehran *et al.*, CVPR'09]	74.5
ViF [Hassner *et al.*, CVPRW'12]	81.3
HOT [Mousavi *et al.*, WACV'15]	82.3
F^{L} | F^{Cv} [Mohammadi *et al.*, AVSS'15]	85.4
Our Approach	**96.4**

11↑

STIFV outperforms existing techniques on 3 violence recognition datasets: Violent-Flows, Movies, Hockey Fight

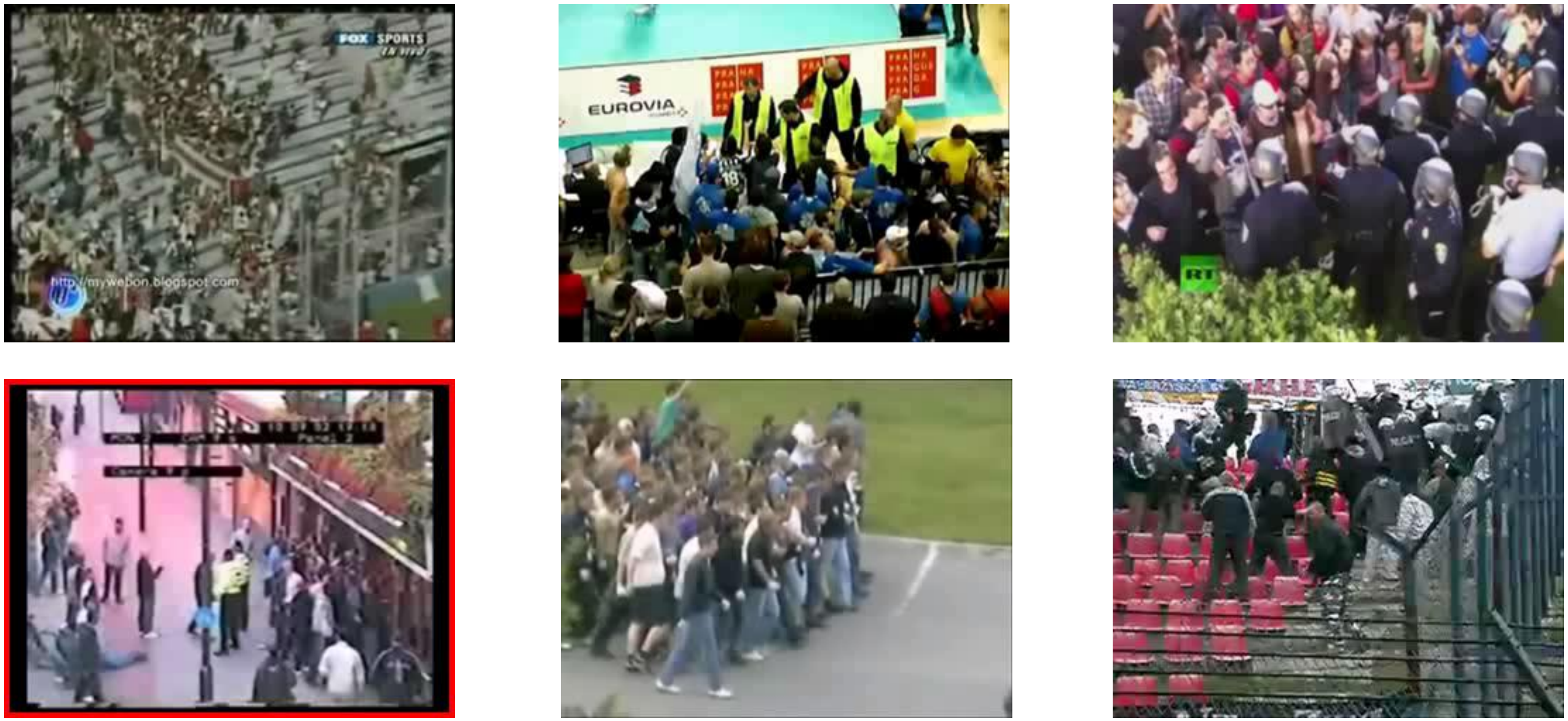
Sliding Window

- We search for a range of frames which contains a violence.
- We base our approach on the temporal sliding window which evaluates video sub-sequences at varying locations and scales.
- 1 scale only:



- Improved Fisher Vectors with summed area table / KDD-trees

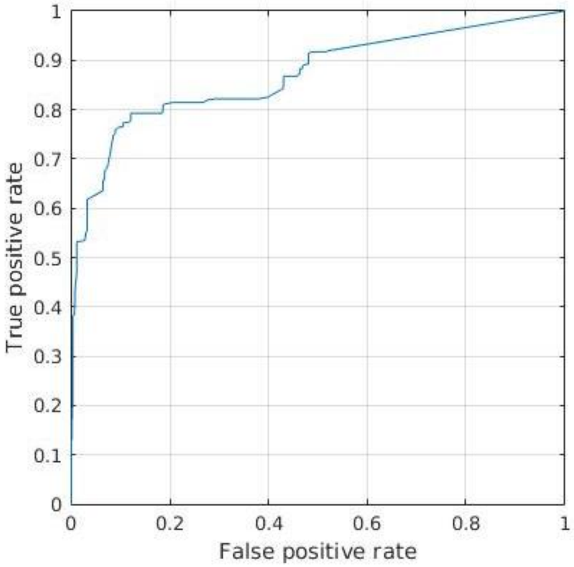
Dataset: Violent-Flows 21 (Crowd Violence \ Non-violence 21)

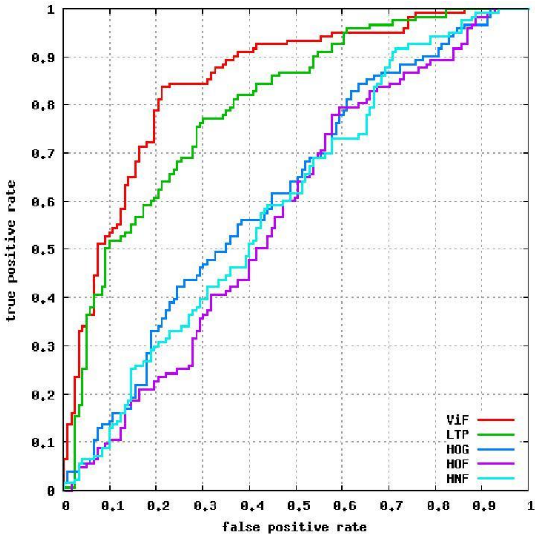


21 videos with real-world video footage of crowd violence, collected from YouTube. They begin with non-violent behavior, which turns to violent mid-way through the video.

The training is performed using 227 out of 246 videos from the Violent-Flows dataset; 19 videos are removed as they are included in the detection set.

Violent-Flows 21: Results & Comparison (ROC, AUC, fps)





Approach	AUC
LTP	79.9
HOG	61.8
HOF	57.6
HNF	59.9
ViF	85.0
Ours	**87.0**

Process	Processing Time (fps)
Feature Extraction (Improved Dense Trajectories)	5.7
Sliding Window	9.28
Ours: Fast Sliding Window	**99.21**

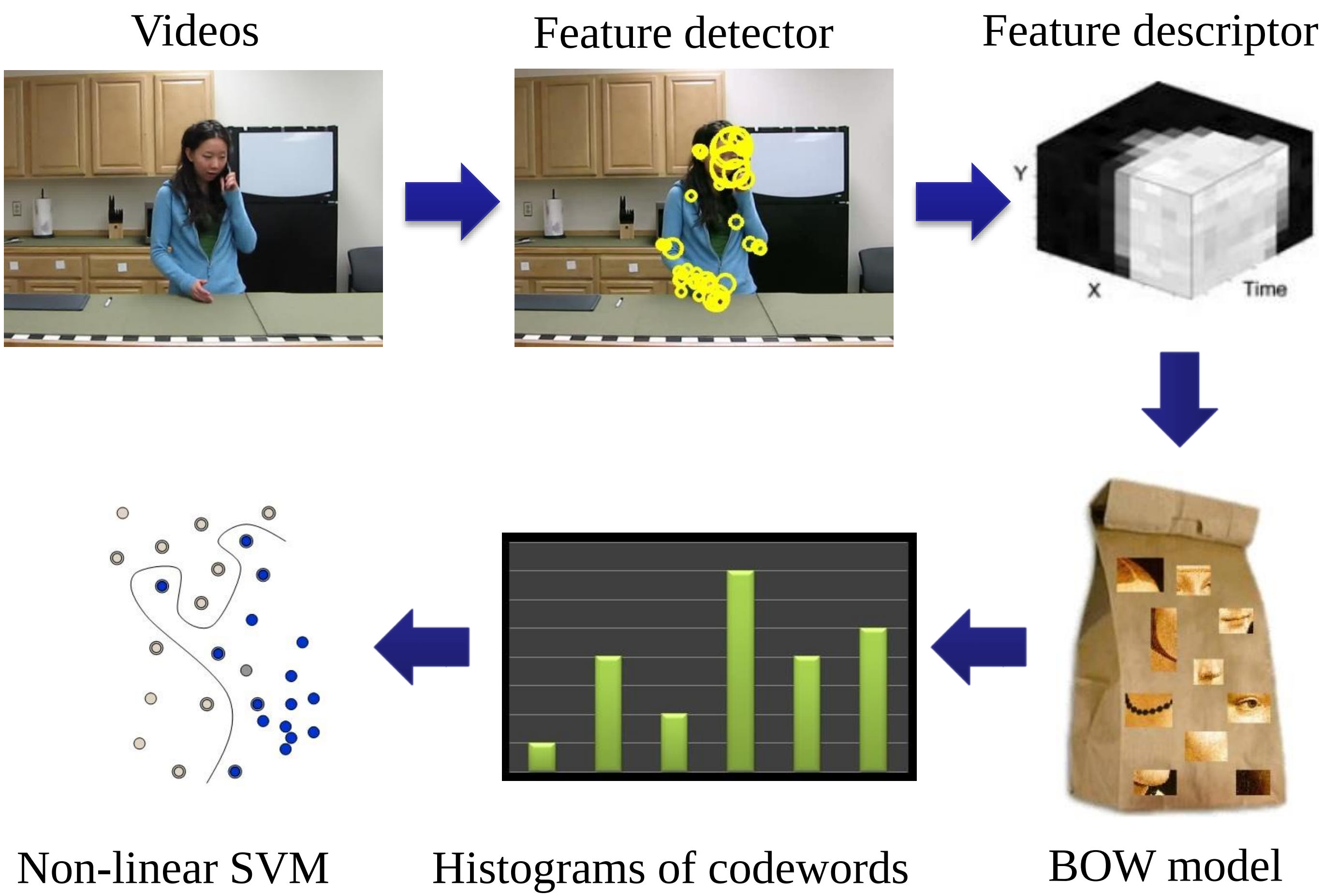
Reduce the memory usage (a lot of motion, dense features):
e.g. 130k features in a 35sec. video = 1.6M floats to store per second = 29x IFV with 128 Gaussians.

SofA: limitations of BoW

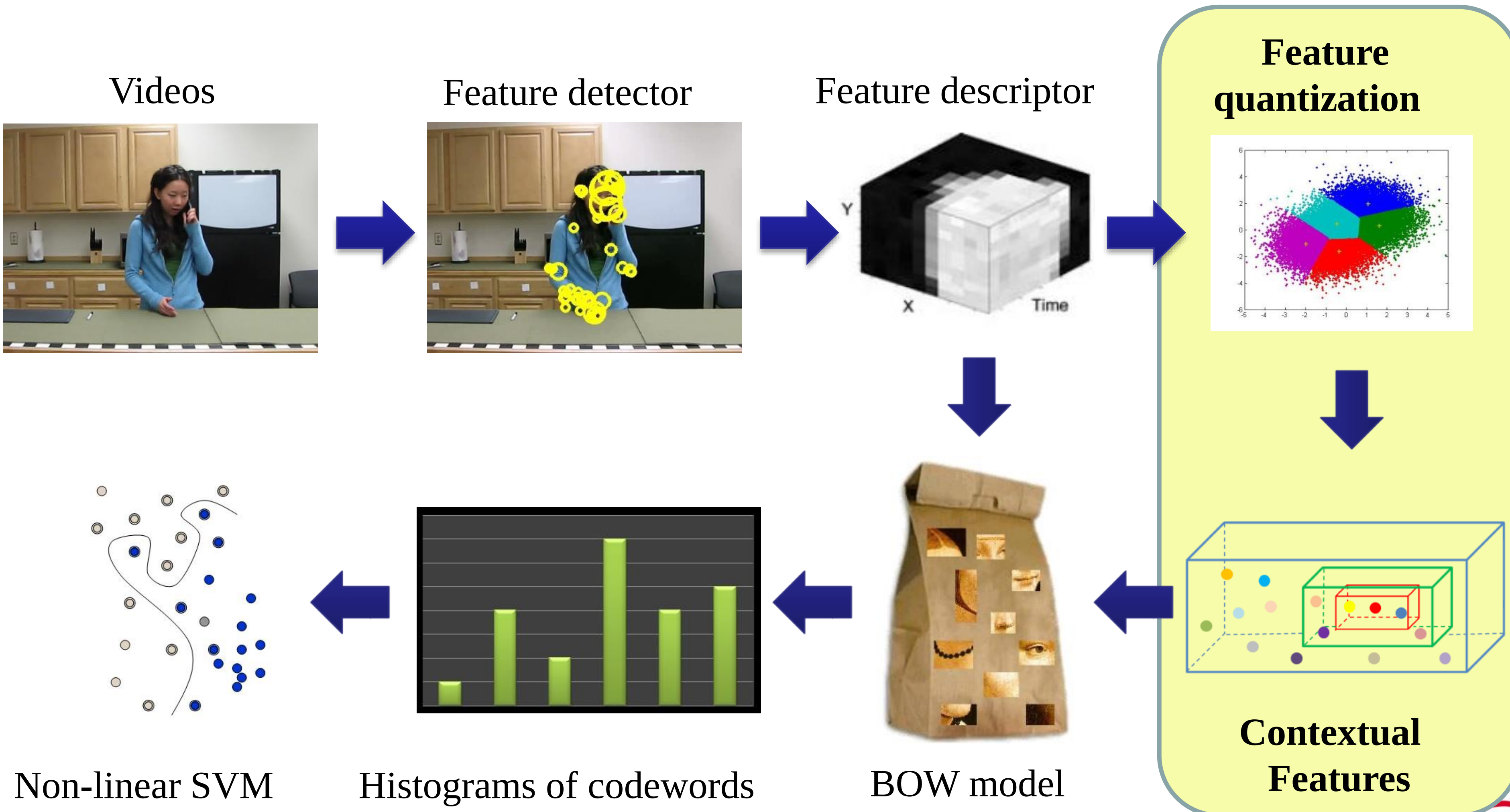
- Recent methods:
  - have focused on capturing global and local statistics of features
  - mostly ignore relations between the features
    - especially, spatio-temporal order of features

- Our goal is to propose a novel representation of CF:
  - overcoming limitations of BOW, i.e. capturing:
    - Global statistics of features
    - Local statistics of features
    - Pairwise relationship between features
    - Order of local features
  - to enhance the discriminative power of features and improve action recognition performance

Contextual Statistics
of Space-Time Ordered Features
for Human Action Recognition (Piotr BILINSKI)



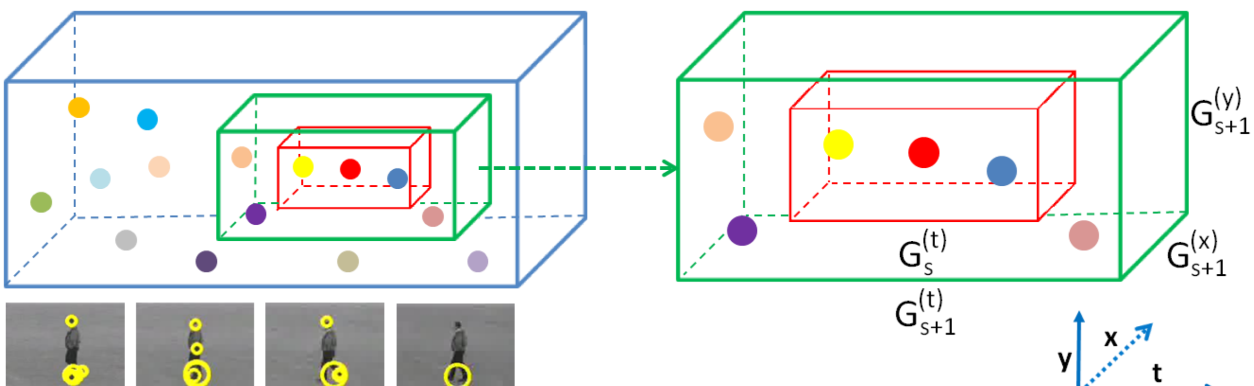
Overview of our approach



Contextual Features

Video

Quantized local features
(features assigned to visual words)



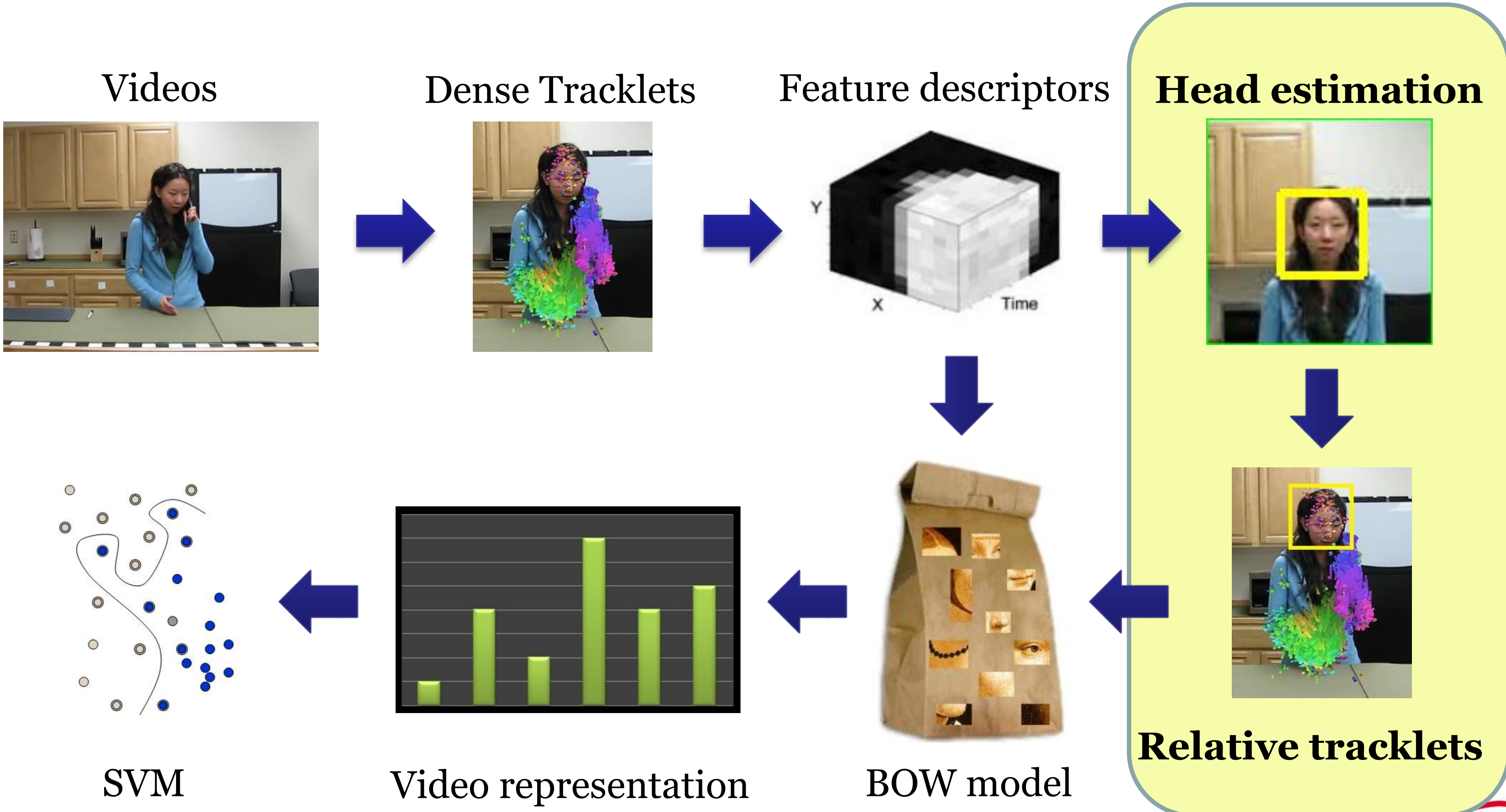
Multi-scale figure-centric neighbourhoods

ADL - Results

Method	Year	Recognition Rate (%)
Matikainen *et al.* [24]	2010	70%
Satkin *et al.* [29]	2010	80%
Banabbas *et al.* [4]	2010	81%
Raptis *et al.* [28]	2010	82.67%
Messing *et al.* [25]	2009	89%
Wang *et al.* [34]	2011	96% (93.8% for KTH)
Our method		**93.33%**

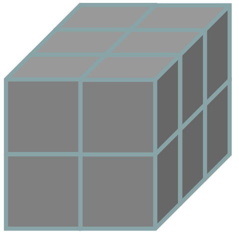
STIPx2 and 1 Person out validation <> test

Relative Dense Tracklets for Human Action Recognition (Piotr BILINSKI)



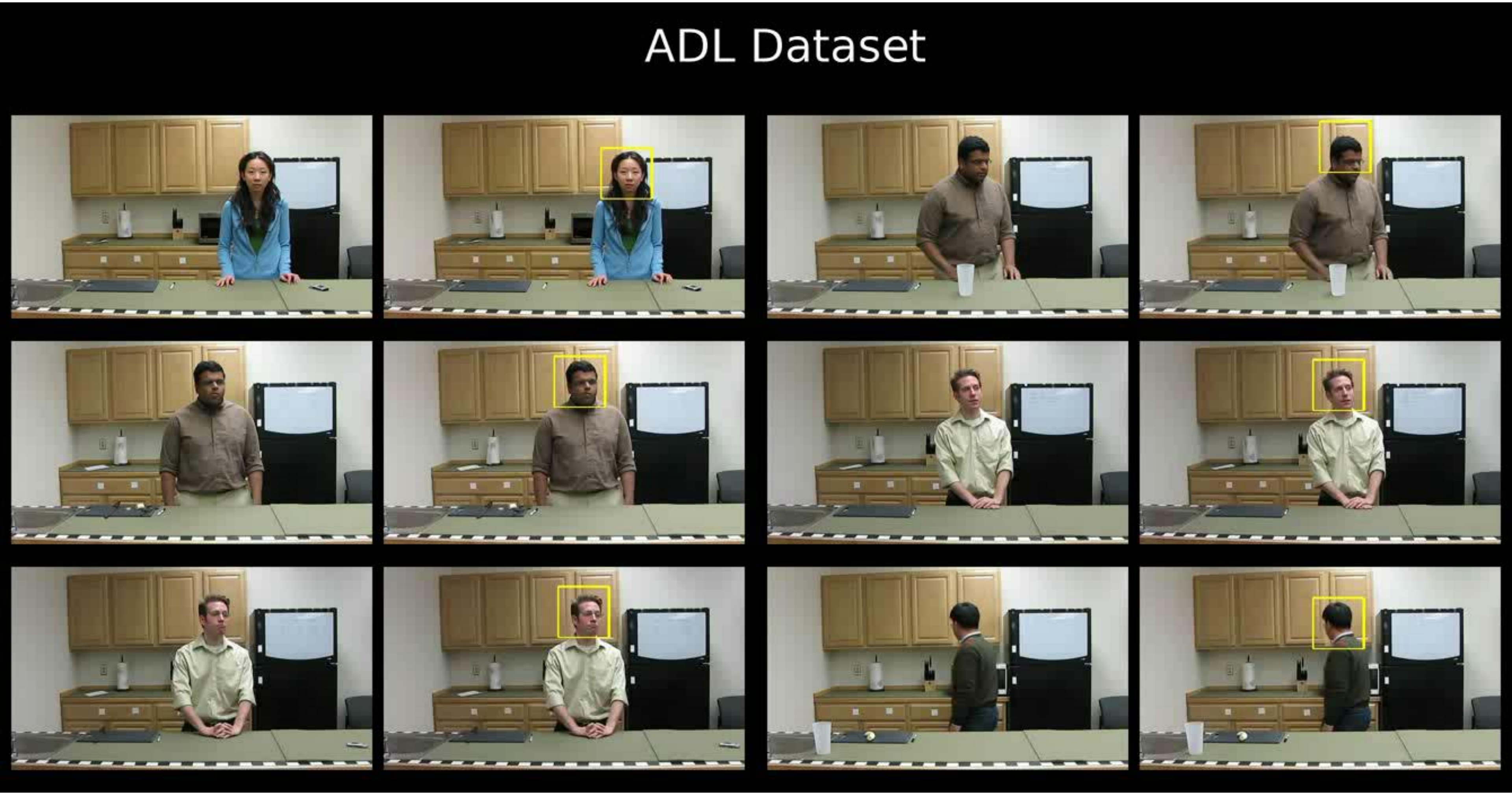
Relative Dense Tracklet Descriptors

- Shape Multi-Scale Tracklet (SMST) Descriptor
  - encodes a local motion pattern of a tracklet as its displacement vectors normalized by the sum of the magnitudes of these displacement vectors.
- HOG and HOF descriptors:
  - encode appearance around tracklets.
  - For each tracklet we define a grid (2×2×3).
  - For each cell of a grid we compute a histogram.
  - HOG – capture local visual appearance.
  - HOF – capture local motion appearance.



- Relative Multi-Scale Tracklet (RMST) Descriptor
  - encodes shape of a tracklet with respect to the estimated head trajectory.
- Combined Multi-Scale Tracklet (CMST) Desc.
  - Combination of SMST and RMST.

Action Recognition using ADL: Benchmarking video dataset



ADL Dataset – Results

Method	Accuracy
SMST	76.67%
RMST	78.67%
CMST	**88.00%**
CMST + HOG-HOF	**92%**

Method	Accuracy
Matikainen *et al.*	70%
Satkin *et al.*	80%
Banabbas *et al.*	81%
Raptis *et al.*	82.67%
Messing *et al.*	89%
Wang *et al.*	96% (93.8% for KTH)
Our method	**92%**

Head + Tracklet

Hospital Action Dataset

8 actions (semi-guided): playing cards, matching ABCD sheets of paper, reading, sitting down and standing up, turning back, standing up and moving ahead, walking back and forth.

55 older people : NC/ MCI/ AD patients.

Spatial resolution: 640×480.

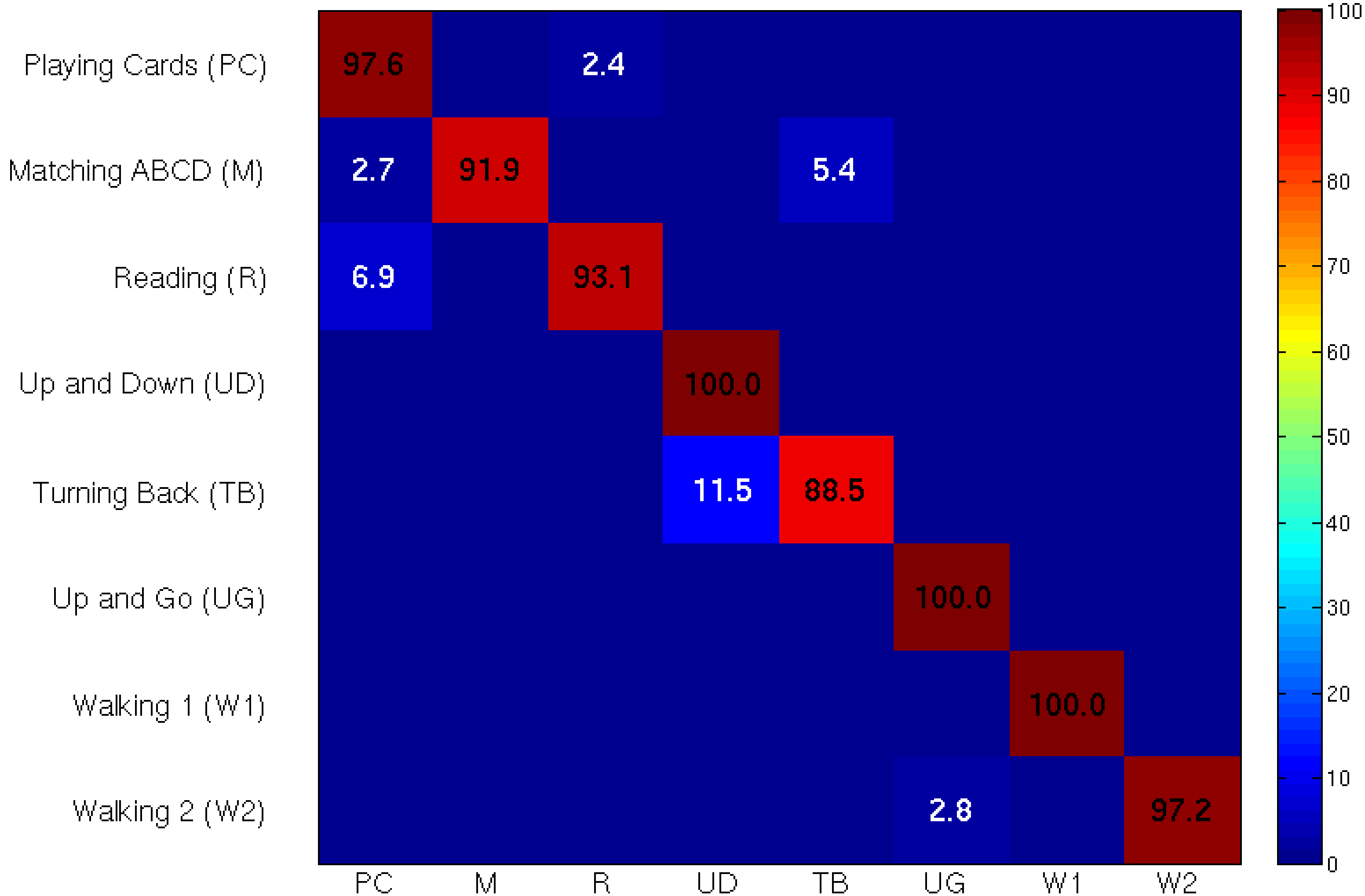
Frame rate: 20 fps.

Challenges: different shapes, sizes, genders and ethinicities of people, occlusions, and multiple people (sometimes both patient and doctor are visible).

Evaluation Scheme: 5-people-fold cross-validation.

Action Recognition using Nice hospital video dataset

Hospital Action Dataset – Results 1



Issues in Action Recognition

Many parameters to tune

- Different detectors (Hessian, Dense sampling, STIP, IDT, context...)
- Different parameters of descriptors (grid size, ...)
- Different clustering algorithms (kmeans++,…)
- Different classifiers (k-NN, linear-SVM, ...)
- Different pooling algorithms (Soft assignment, sparse coding, Fisher Kernels, Naïve Bayes Nearest Neighbour,…)

Performance depends on training sets

- Different resolutions of videos
- Generic to other datasets (IXMAS, UCF Sports , Hollywood, Hollywood2, YouTube, ...)

Still open challenges

- Finer actions, more discriminative, without context...

Issues in Action Recognition

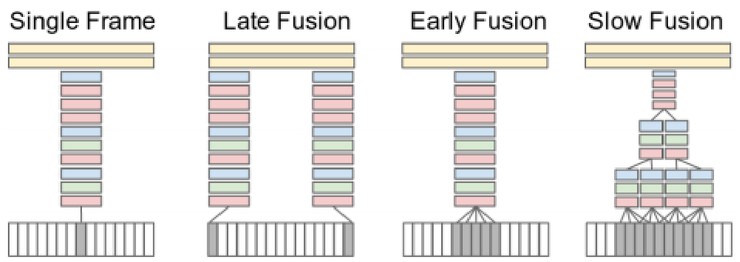
Deep Convolutional Neural Networks

Images

- Large Annotated data (Imagenet)
- Architecture Suitable for Images with good resolution

Videos: How to capture motion information in CNN ?

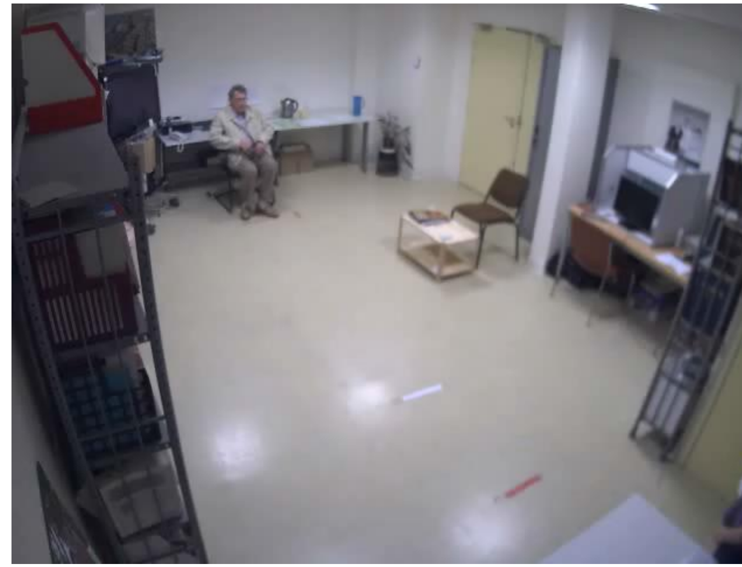
- Stacking of frames

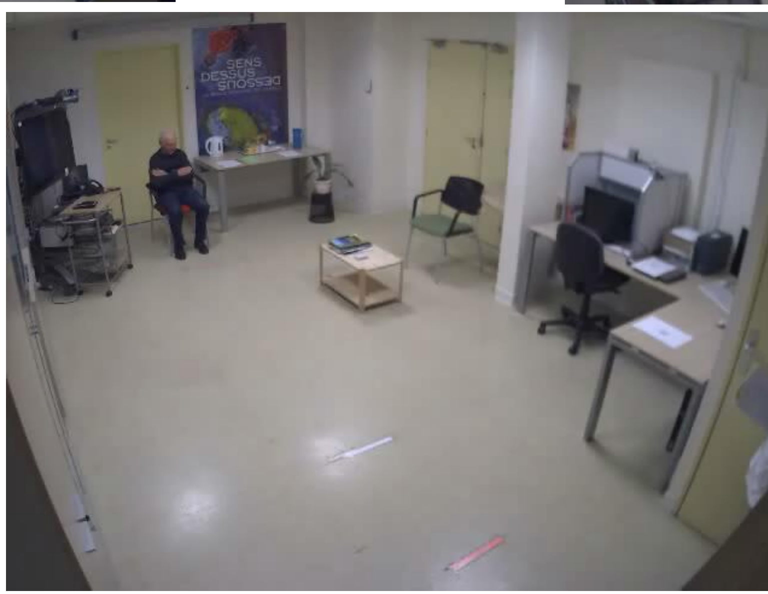


- Capture motion independently: several stream CNNs
  - One ConvNet to capture static visual information.
  - Another ConvNet to capture motion information (like Optical Flow, but expensive)
  - Other Nets to capture motion on longer scales
- Trajectory-Pooled Deep-Convolutional Descriptors using Improved Dense Trajectories

Issues in Action Recognition

- Finer actions, more discriminative (NC, MCI, AD)

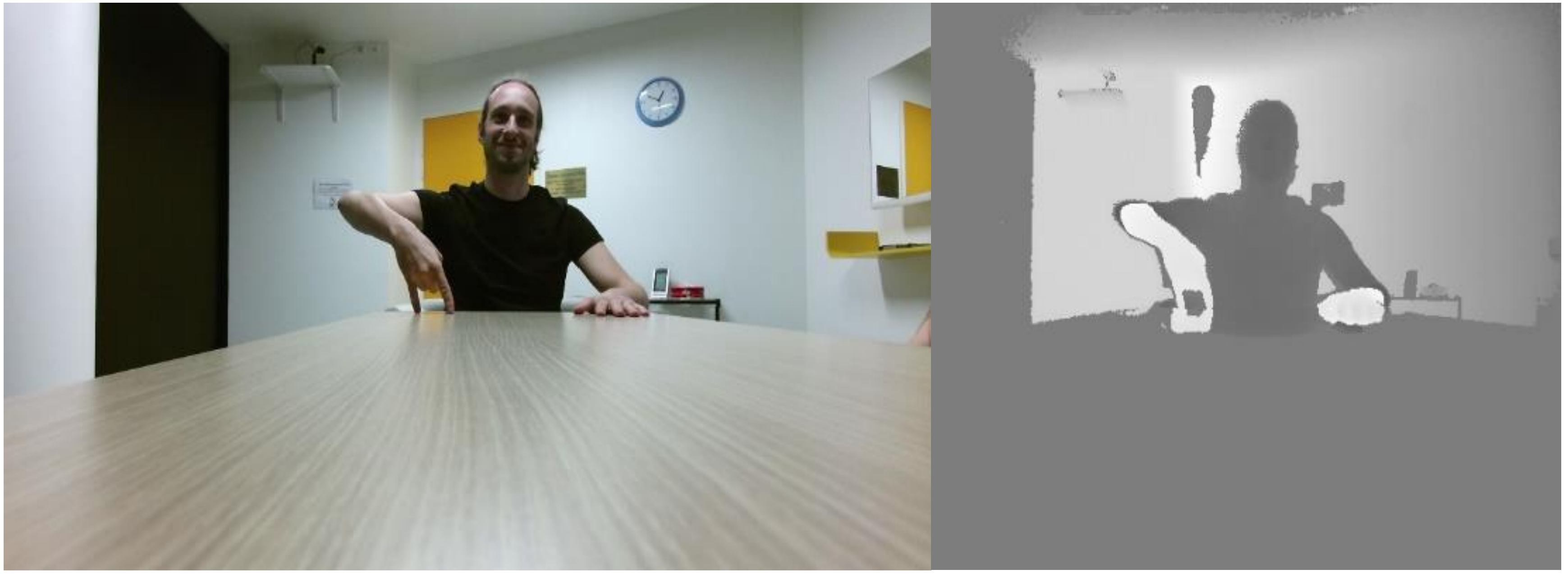




AD versus NC	
Playing Cards	69%
Up and Go	66%
Reading	44%

Issues in Action Recognition

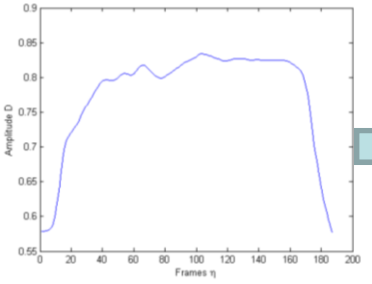
- Finer actions, smiling, talking, grim, gender, age, praxis



Gender recognition using smile: Dynamics based on Facial Landmarks

Can a smile reveal your gender?
P. Bilinski, A. Dantcheva, F. Bremond

Facial landmark detection



Signal displacement of facial landmarks

Feature	Definition			
	General	Onset	Apex	Offset
Duration		$\frac{\eta(D^+)}{\omega}$	$\frac{\eta(D^a)}{\omega}$	$\frac{\eta(D^-)}{\omega}$
Duration Ratio		$\frac{\eta(D^+)}{\eta(D)}$		$\frac{\eta(D^-)}{\eta(D)}$
Maximal Amplitude	$max(D)$			
STD of Amplitude	$std(D)$			
Mean Amplitude		$mean(D^+)$	$mean(D^a)$	$mean(D^-)$
Total Amplitude		$\sum(D^+)$		$\sum(\vert D^-\vert)$
Net Amplitude	$\sum(D^+) - \sum(\vert D^-\vert)$			
Amplitude Ratio		$\frac{\sum(D^+)}{\sum(D^+)+\sum(D^-)}$		$\frac{\sum(D^-)}{\sum(D^+)+\sum(D^-)}$
Maximal Speed		$max(V^+)$		$max(V^-)$
Mean Speed		$mean(V^+)$		$mean(V^-)$
Maximum Acceleration		$max(A^+)$		$max(A^-)$
Mean Acceleration		$mean(A^+)$		$mean(A^-)$
Net Ampl., Duration Ratio	$\frac{(\sum(D^+)-\sum(\vert D^-\vert))\omega}{\eta(D)}$			

Statistics of signal displacement

Feature Selection
Min-Redundancy
Max-Relevance
(mRMR)

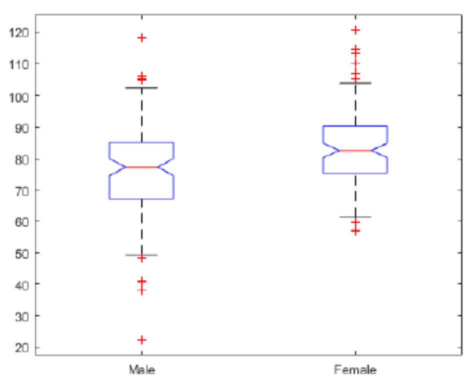
Classification

Male

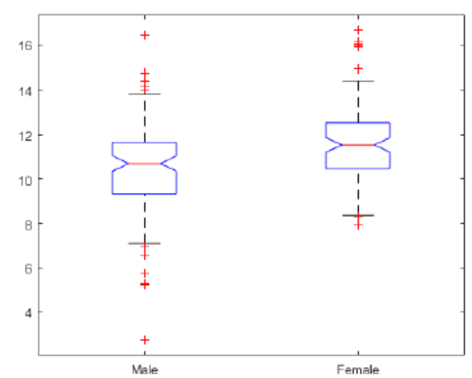
Female

Gender recognition using smile: Pertinent features (dynamics based on facial landmarks)

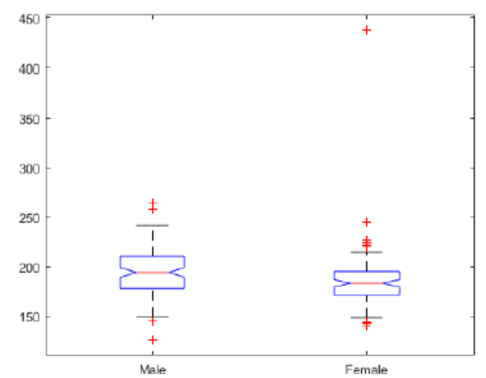
- Adolescents: females show longer Duration Ratio (Offset) and Duration (Onset) on the right side of the mouth and a larger Amplitude Ratio (Onset) on the left side of the mouth, than males.
- In adults, females show: a larger Mean Amplitude (Apex) of mouth opening, a higher Maximum Amplitude on the right side of the mouth, as well as a shorter Mean Speed Offset on the left side of the mouth, than males.



(a) D_{11} Mean Amplitude Apex



(b) D_8 Maximum Amplitude

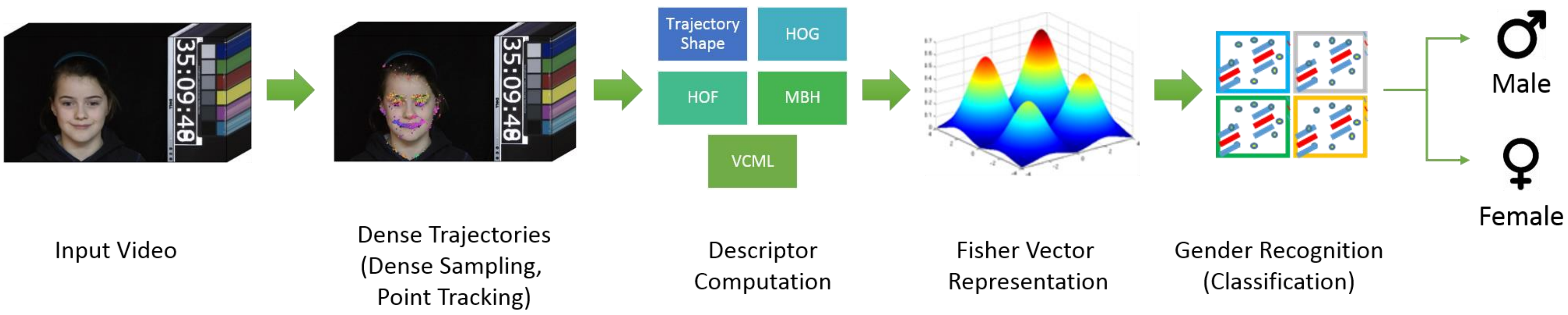


(c) D_9 Mean Speed Offset

[13] Dantcheva, A.; Bremond, F.: Gender estimation based on smile-dynamics, in IEEE TIFS, 2016.

Gender recognition using smile: Proposed method based on IDT and FV

Spatio-temporal features based on dense trajectories [8] represented by a set of descriptors encoded by Fisher Vectors [9].



[8] Wang, J.; Li, J.; Yau, W.; Sung, E.: Boosting dense SIFT descriptors and shape contexts of face images for gender recognition. CVPRW, 2010.
[9] Perronnin, F.; Sanchez, J.; Mensink, T.: Improving the Fisher Kernel for large-scale image classification. ECCV, 2010.

Gender recognition using smile: Dense Trajectories: visualization

Gender recognition using smile:
Results : true gender classification rates

Age (Subject amount)	**≤20 (148)**	**>20 (209)**
OpenBR	52.3%	75.6%
how-old.net	55.5%	92%
COTS (appearance based)	76.9%	92.5%
Dynamics based on facial landmarks	59.4%	67.8%
COTS + Dynamics based on facial landmarks	76.9%	**93%**
Motion-based descriptors	77.7%	80.1%
Proposed Method (IDT+FV)	**86.3%**	91%

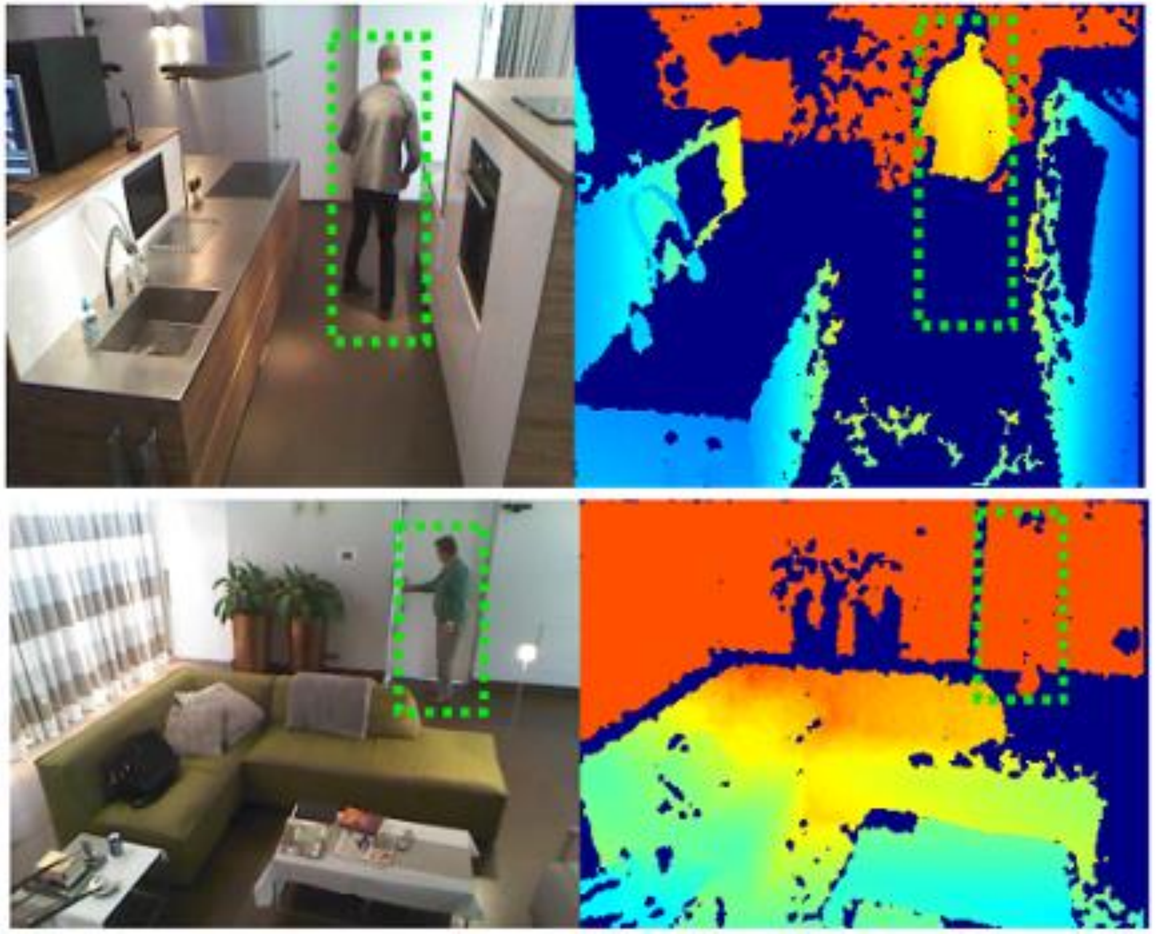
Activity recognition using RGBD sensor Motivation – skeleton based methods & dense trajectories - M. Koperski

State-of-the Art:

Data-set	Dense Trajectories* [%]	Skeleton based method [%]
MSRDailyActivity3D	78.44	85.80 *[Wu et al., CVPR'12]*
CAD-60	66.31	74.10 *[Wu et al., CVPR'12]*
CAD-120	80.19	84.70 *[Koppula et al., CVPR'13]*

* Based on Wang et al., CVPR'11

Activity recognition using RGBD sensor
Motivation – when skeleton detection fails

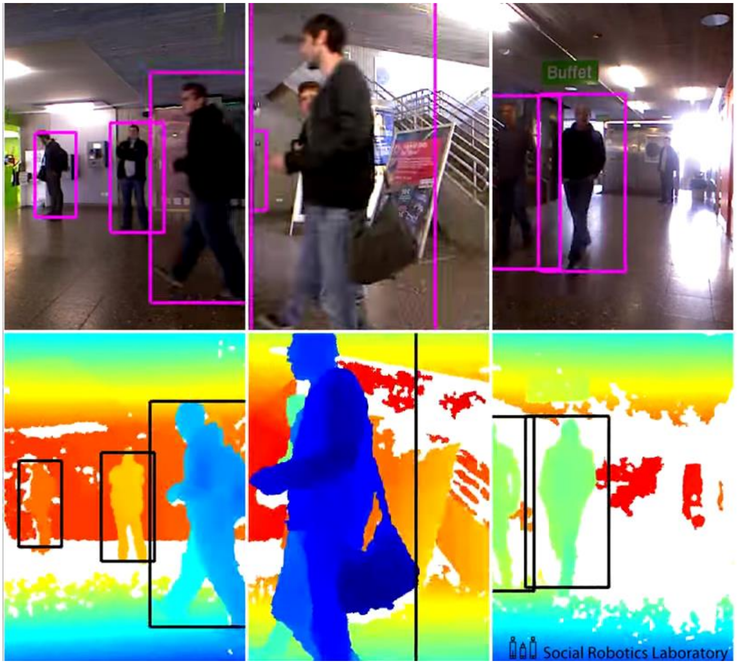


Skeleton Detection
Fails

Activity recognition using RGBD sensor
Proposed solution

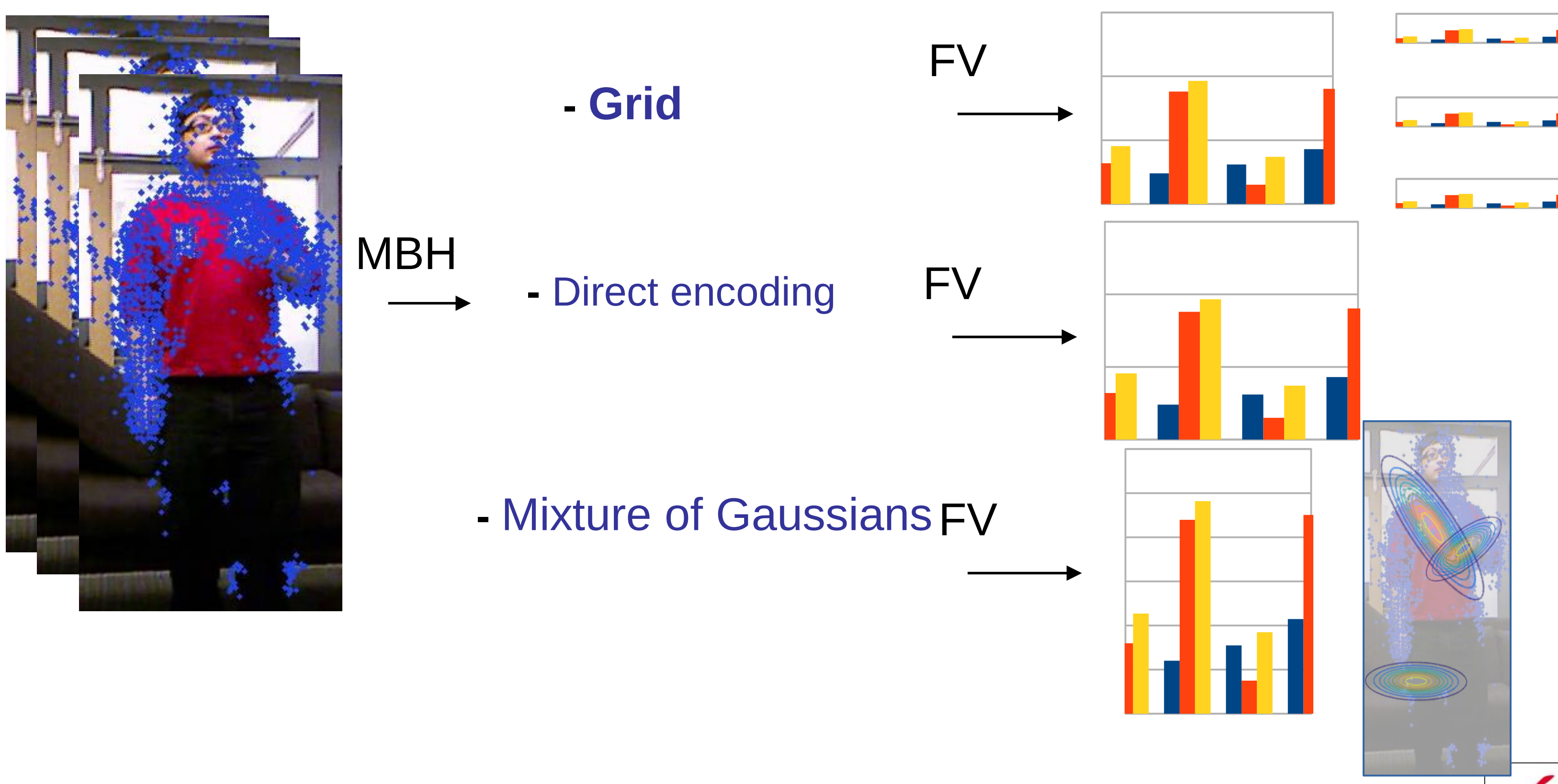
Does not require skeleton detection

- People detection in place of skeleton detection
- Detection based on RGB and depth data
- Dataset: L.Spinello, K. Arras "People detection in RGB-D Data"



Activity recognition using RGBD sensor
Proposed solution – motion features spatial-layout

Motion features spatial-layout : 3 approaches



Activity recognition using RGBD sensor Results

1. We validate our approach on 3 public data-sets :
    a) CAD-60 – 60 videos, 12 actions, 4 subjects,
    b) CAD-120 – 120 videos, 10 actions, 4 subjects,
    c) MSRDailyActivity3D – 360 videos, 16 actions, 10 subjects
2. We use 3x1 grid for GridHOG (cross-validated)
3. We use 3x1 grid for motion features spatial-layout modeling (cross-validated)

Results – MSRDailyActivity3D

Method	Accuracy [%]
NBNN* *[Sedinari et al. CVPRW'14]*	70.00
HON4D* *[Oreifej et al. CVPR'13]*	80.00
STIP+skeleton* *[Zhu et al. I&VC'14]*	80.00
SSFF* *[Shahroudy et al. ISCCSP'14]*	81.90
DSCF* *[Xia et al. CVPR'13]*	83.60
Actionlet Ensemble* *[Wu et al. CVPR'12]*	85.60
RGGP + fushion* *[Liu et al. IJCAI'13]*	85.80
Super Normal* *[Yang et al. CVPR'14]*	86.26
DCSF + joint* *[Xia et al. CVPR'13]*	88.20
BHIM *[Kong et al. CVPR'15]*	**86.88**
Our Approach	85.95

* method which requires skeleton detection

Results – CAD-60

Method	Accuracy [%]
Order Sparse Coding* *[Ni et al. ECCV'12]*	65.30
Object Affordance* *[Koppula et al. ICML'13]*	71.40
HON4D* *[Oreifej et al. CVPR'13]*	72.70
Actionlet Ensemble* *[Wu et al. CVPR'12]*	74.70
Joule SVM* *[Hu et al. CVPR'15]*	84.10
STIP *[Zhu et al. IV&C'14]*	62.50
Our Approach	**80.36**

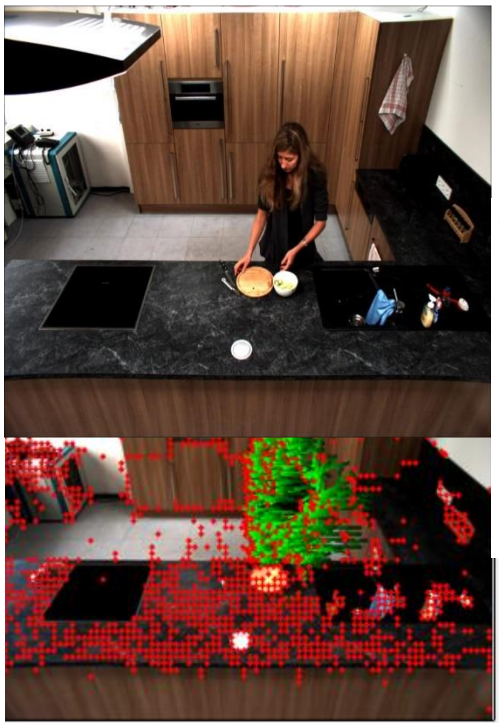
Results – CAD-120

Method	Accuracy [%]
Object Affordance* *[Koppula et al. ICML'13]*	84.70
STS* *[Koppula et al. ICML'13]*	93.50
Salient Proto-Objects *[Rybok et al. WACV'13]*	78.20
Our Approach	**85.48**

* method which requires skeleton detection

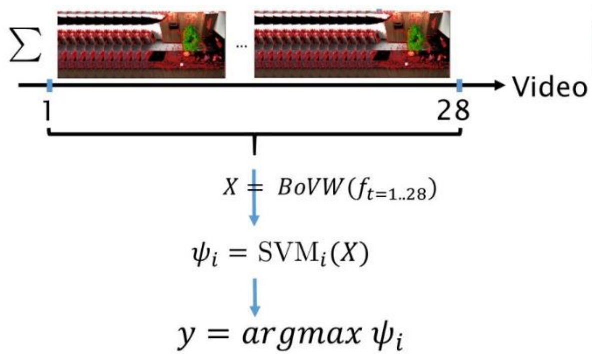
Semi-supervised understanding of complex activities from temporal concepts, C. Crispim

Image -> features

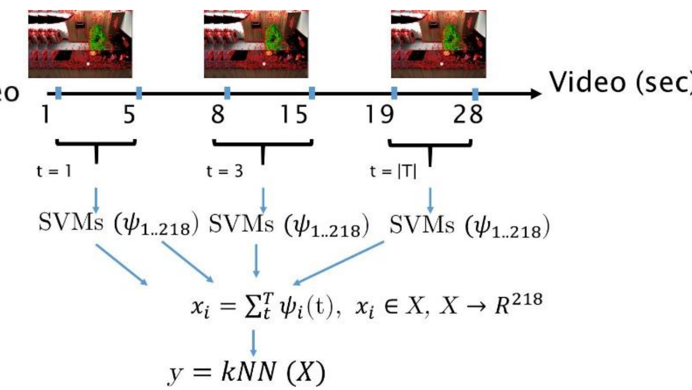


Lack of attention to temporal and composite relations

Flat, feature-based



Flat, concept-based



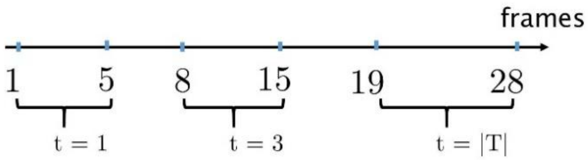
$$\psi_i \in \{tomato, knife, take, stir, bread\ loaf, \dots,\}_{218}$$

Find probabilistic representation
of an activity video
given temporal composite concepts

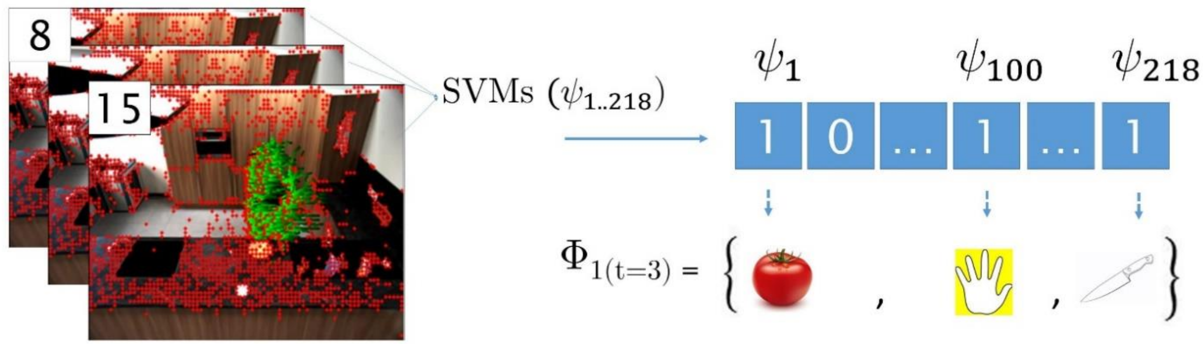
?=

Semi-supervised understanding of complex activities from temporal concepts

1) Video temporal segmentation



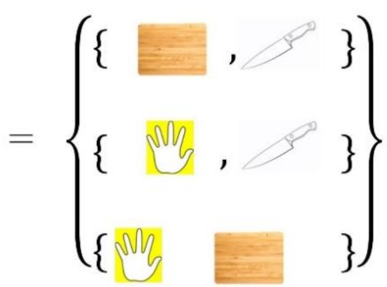
2) Concept recognition at time segment t:



3) Composite concept generation at t

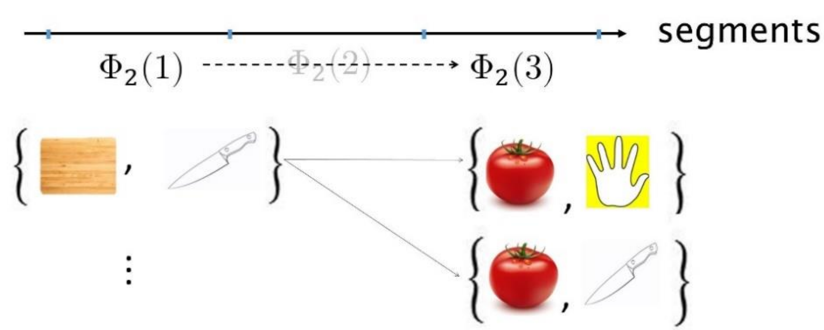
C_k^N: k-combinations of elements in N

$$\Phi_2(t=1) = \quad C_k^{\Phi_1(t=1)}$$



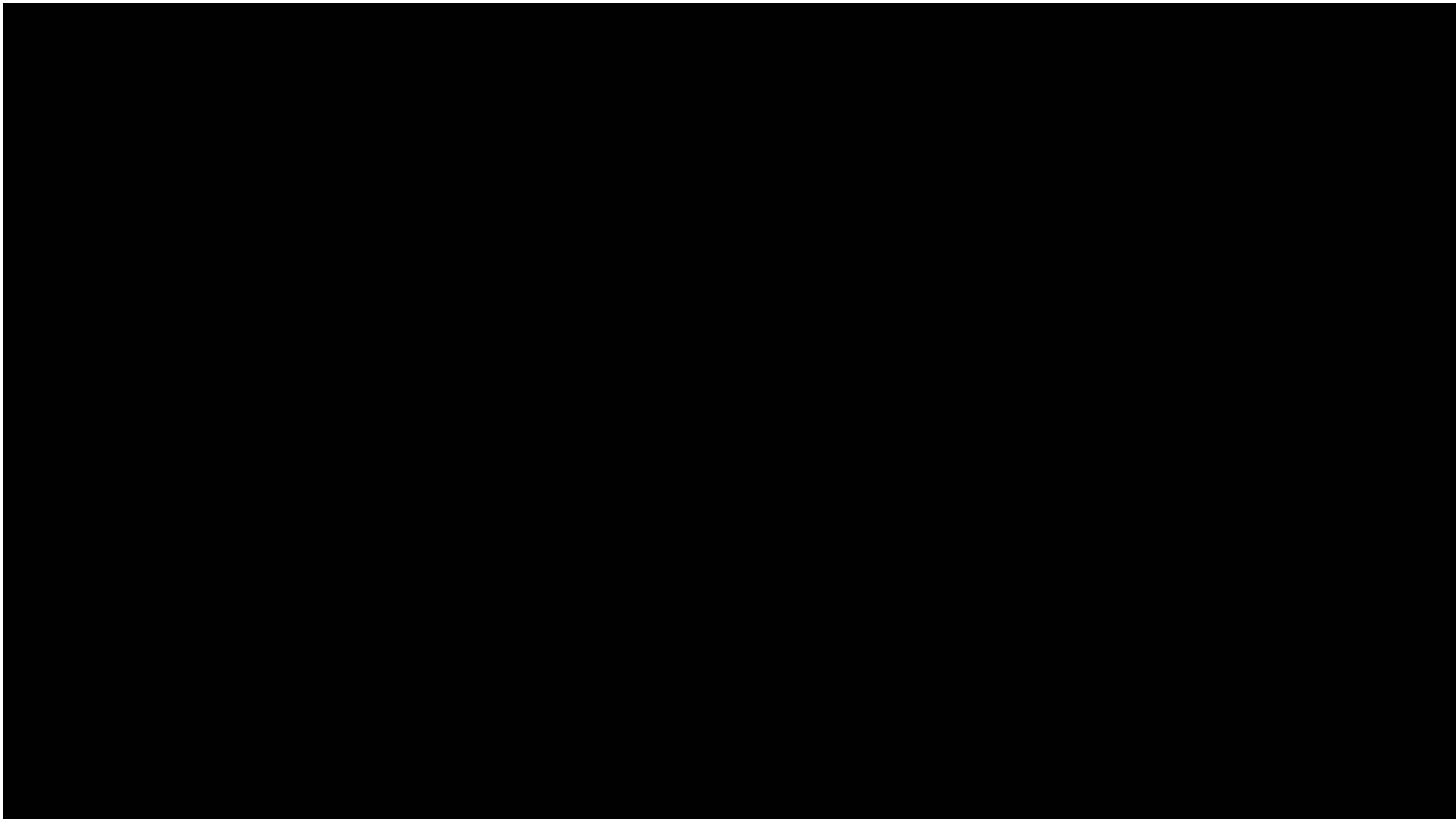
4) Temporal composites between segments

$$\Gamma_2\,(1,3) = \Phi_2(1) \times \Phi_2(3)$$



Semi-supervised understanding of complex activities from temporal concepts

Cooking Composite data set [Rohrbach, et al., ECCV 2012].



Monitoring of Activities of Daily Living for Older People

Motivation : Increase autonomy and quality of life

- Enable older adults to live longer, autonomously in their preferred environment.
- Reduce costs for public health systems.
- Relieve family members and caregivers.

Objectives :

- Detecting alarming situations *(eg. Falls)*
- Assess the degree of frailty of older people (impact of therapies).
- Detecting changes in behavior
  (*missing activities, disorder, interruptions, repetitions, inactivity*).
- Building a video library of reference behaviors characterizing people frailty.

Approach : designing activity recognition systems

Event Recognition based on Knowledge

Design a language for event recognition:

An **event** is mainly constituted of five parts:

- Physical objects: all real world objects present in the scene observed by the cameras

  Mobile objects, contextual objects, zones of interest

- Components: list of states and sub-events involved in the event
- Forbidden Components: list of states and sub-events that must not be detected in the event
- Constraints: symbolic, logical, spatio-temporal relations between components or physical objects
- Action: a set of tasks to be performed when the event is recognized

A language to model complex events

- **Language combining multi-sensor information**

EVENT (Use Fridge,

Physical Objects ((p: Person), (Fridge: Equipment), (Kitchen: Zone))

Components ((c1: Inside zone (p, Kitchen))

Detected by video camera

(c2: Close_to (p, Fridge))

(c3: Bending (p))

(c4: Opening (Fridge))

Detected by contact sensor

(c5: Closing (Fridge)))

Constraints ((c1 before c2)

(c3 during c2)

(c4:time + 10s < c5:time))**)**

Event recognition results

- Recognition of the "Having meal" event for a 84 old woman

Discussion about the obtained results

+ Results of recognition of 6 daily activities for 5*4=20 hours

Activity	GT	TP	FN	FP	Precision	Sensitivity
Use fridge	65	54	11	9	86%	83%
Use stove	177	165	11	15	**92%**	**94%**
Sitting on chair	66	54	12	15	78%	82%
Sitting on armchair	56	49	8	12	80%	86%
Prepare lunch	5	4	1	3	**57%**	**80%**
Wash dishes	16	13	3	7	65%	81%

- Errors occur at the border between living-room and kitchen
- Mixed postures such as bending and sitting due to segmentation errors

Discussion about the obtained results

+ Good recognition of a set of activities and human postures (video cameras)

Activity	GT	TP	FN	FP	Precision	Sensitivity
Use fridge	65	54	11	9	86%	83%
Use stove	177	165	11		**92%**	**94%**
Sitting on chair	66	54	12	15	78%	82%
Sitting on armchair	56	49	8	12	80%	86%
Prepare lunch	5	4	1	3	**57%**	**80%**
Wash dishes	16	13	3	7	65%	81%

Bag on chair

2 instances of the event

Cold meal

- Errors occur at the border between living-room and kitchen
- Mixed postures such as bending and sitting due to segmentation errors

Monitoring Activities at Nice Hospital

- **Medical staff & healthy younger**
  - 22 people (more female than male)
  - Age: ~ 25-35 years
  - Medical staff
- **Older persons (normal control)**
  - 20 (woman & man)
  - Age: ~ 60-85 years
- **Alzheimer patients**:
  - 21 AD people (woman & man)
  - 19 MCI (mild cognitive impairment) and mixed
  - Age: ~ 60-85 years

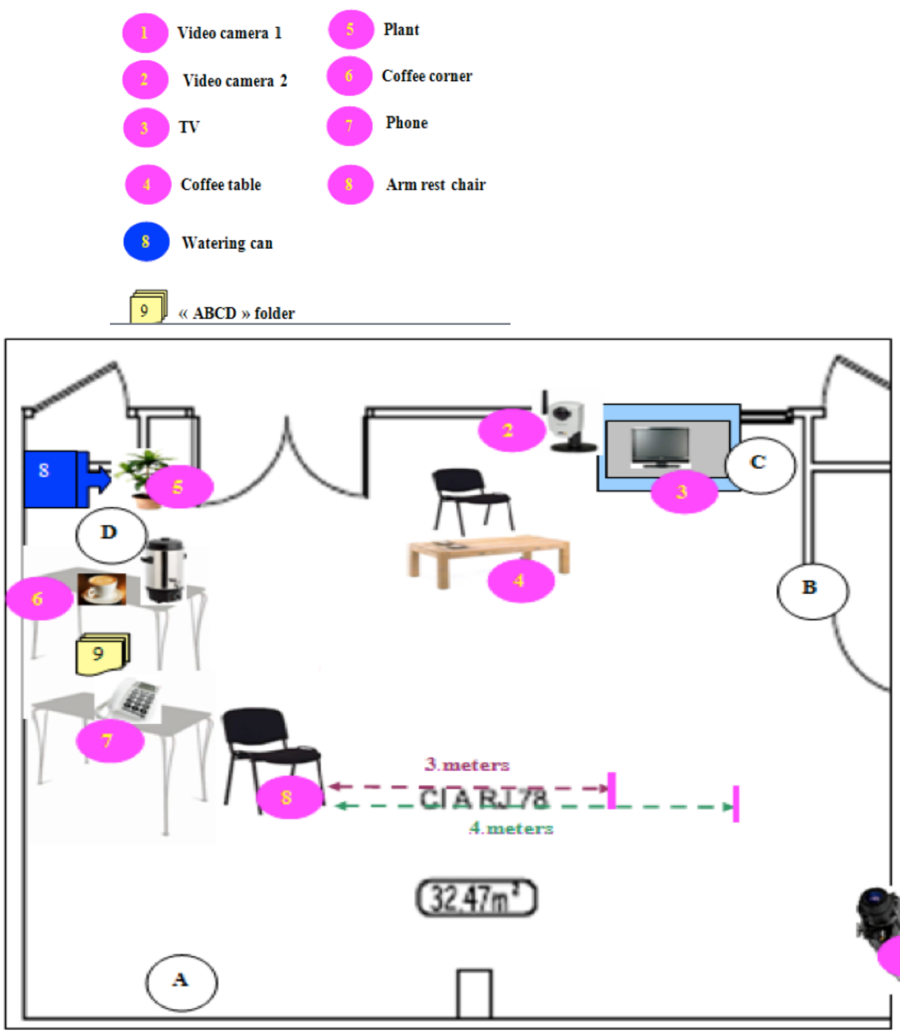


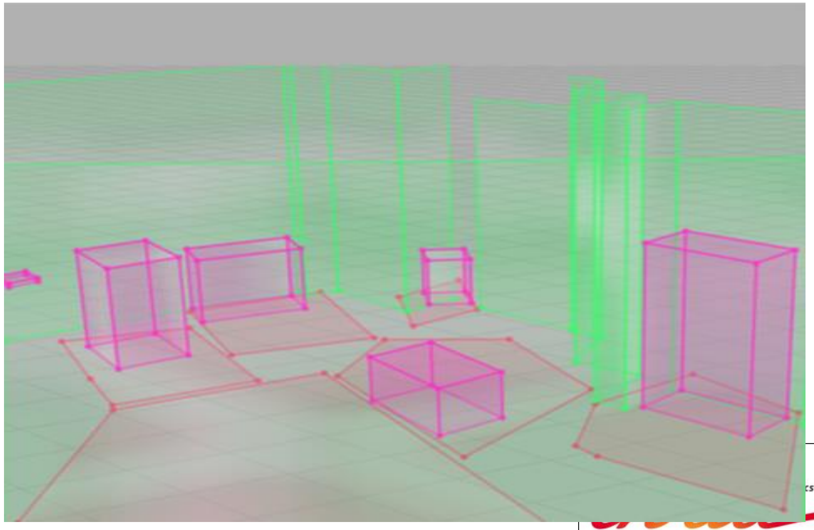



- **Activities monitored by various sensors**:
  - 2D RGB video cameras,
  - 3D RGBD video cameras,
  - inertial sensors : Actiwach/ motionPod
  - Stress sensors (impedance)
  - Microphones

- **3 Medical Protocols**
  - Protocol1: ~1year (2010-2011) **36** (18NC/6MCI/12AD) persons recruited
  - Protocol2: ~1year(2011-2012) **79** (29NC/36MCI/14AD) persons recruited
  - Protocol3: start on 06/2012 - **150** (50NC/50AD/50MCI) persons expected

CMRR in Nice Hospital Screening of AD patients



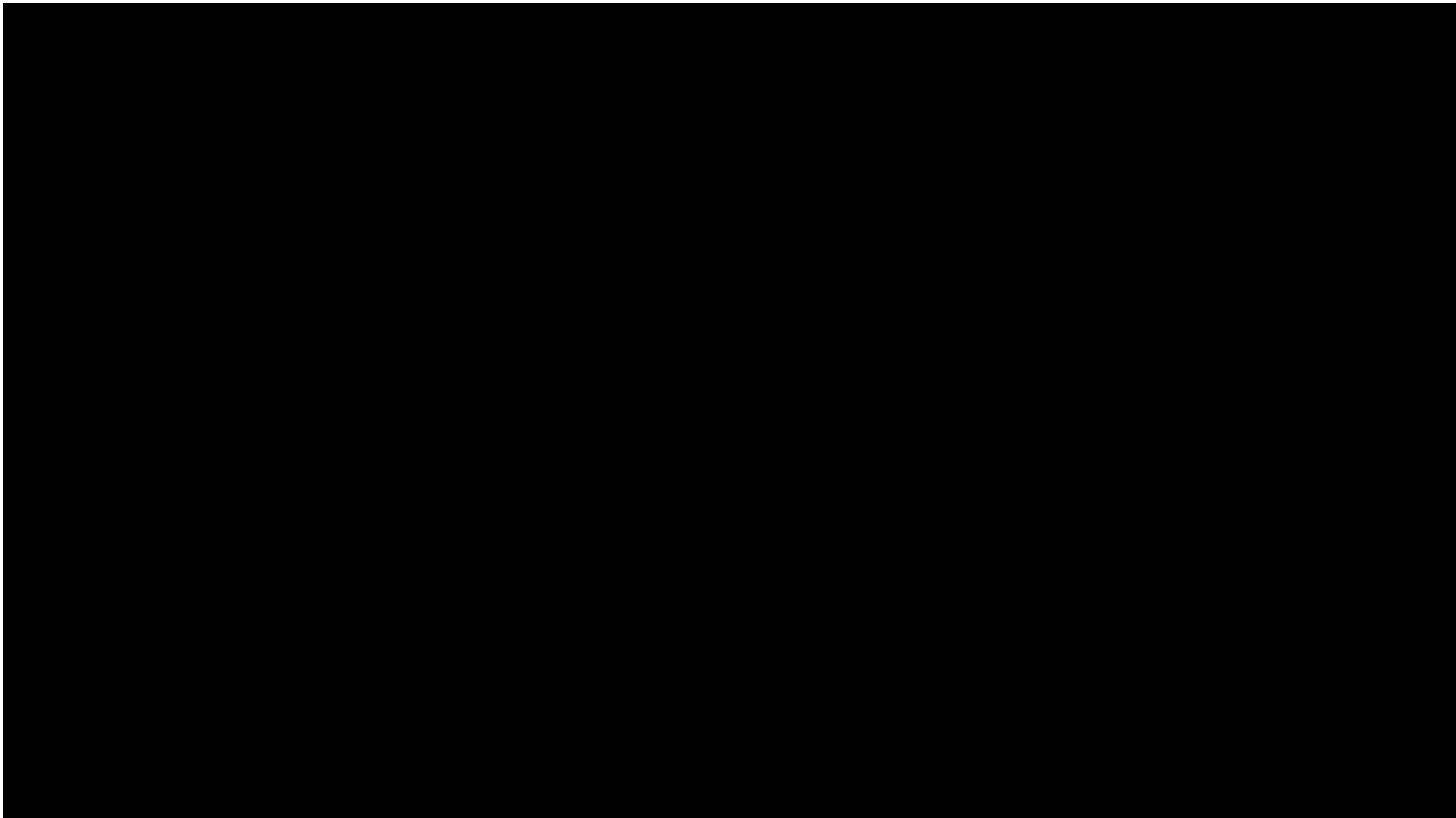


Activity monitoring in Nice Hospital with AD patients

Recognition of the "stand-up & walking" activity.

Activity monitoring in Nice Hospital with AD patients

Visualization of older adult performance while accomplishing the semi-guided tasks.



Activity monitoring in Greece Hospital with AD patients

Visualization of older adult performance while accomplishing the semi-guided tasks.

Activity monitoring in Greece Hospital with AD patients

Visualization of older adult performance while accomplishing the semi-guided tasks.

Activity monitoring at ICP with AD patients

Visualization of older adult performance while accomplishing the semi-guided tasks.

Experimental Results: Summary of Patient Activities Physician Interface

Scenario: Semi-guided

Current patient / Reference older people

Current patient		Healthy	MCI	Alzheimer
PREPARE_DRINK		HEALTHY	MCI	ALZHEIMER
- Frequency (times):	4	2±1.08	1.08±0.76	1.25±0.45
- Duration (s):	46.2	42.94±22.50	51.94±36.49	33.61±30.39
TALK_ON_PHONE		HEALTHY	MCI	ALZHEIMER
- Frequency (times):	1	2.11±0.83	2.04±0.79	2±1.03
- Duration (s):	4.6	37.54±12.31	42.84±16.57	43.48±15.08
READ		HEALTHY	MCI	ALZHEIMER
- Frequency (times):	1	0.94±0.23	0.96±0.79	0.55±0.61
- Duration (s):	7.9	57.19±15.33	73.9	184
PREPARE_DRUG_BOX		HEALTHY	MCI	ALZHEIMER
- Frequency (times):	2	1±0	1.08±0.57	0.94±0.80
- Duration (s):	17.2	82.68±24.55	113.40±48.20	82.93±50.29
WATER_PLANT		HEALTHY	MCI	ALZHEIMER
- Frequency (times):	5	1±0	0.6±0.64	1.14±0.38
- Duration (s):	41.5	7.03	6.61±2.27	5.66±1.87

European FP7 Project Dem@Care (end Dec 2015)

- **Experiments: Pilot1 @Lab (France, Thessaloniki) & Pilot2 @Nursing-Home (France, Ireland) & Pilot3 @Home (Ireland, Sweden, Thessaloniki, France):**
  - Objectives:
    - Monitoring of the **5 functional areas**: **Sleep** (diurnal/nocturnal), **ADL/IADLs**, **Physical Exercise**, **Social Interaction**, **Mood**
  - Clinical Motivations : autonomy
    - Clinician benefits: Maintain comprehensive views of the status and progress of PwD's health in order to increase the **early detection** rate of **functional decline and other disorders** in older adults
    - PwD/Caregiver benefits:
      - Real-time **alerts**, Receive **adaptive feedback** and **personalized support**
  - Tested sensors (to be updated):
    - Video camera: RGB ambient (Axis®)/embedded, GoPro®) video camera, SenseCam®(Image, ambient light, T°C), RGBD video camera (Kinect®)
    - Audio: Ambient and embedded microphone
    - Accelerometers/Physiological sensors: BodyMedia SenseWear Pro3® (Skin conductance, 2D accelerometer, T°C), Philips DTI-2®(Skin conductance, 3D accelerometer, T°C, ambient light), Wireless Inertial Measurement Unit devices (accelerometry, gyroscope data)
    - Environmental sensors: Power consumption, Presence sensor, Sleep sensor

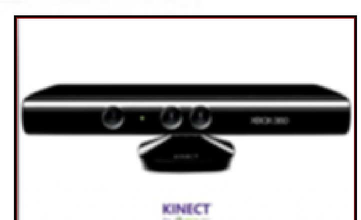
Dem@Care Sensors



- **Wearable sensors:**
  - Physiological: (WIMU), DTI – 2
  - Life- logging sensors: (SenseCam)
  - Audiovisual: wearable microphone, GoPro camera


- **Ambient sensors:**
  - Gear 4 Sleep Clock, Aural, Bedit…
  - Static camera: Sony Kinect, ASUS RGB-D

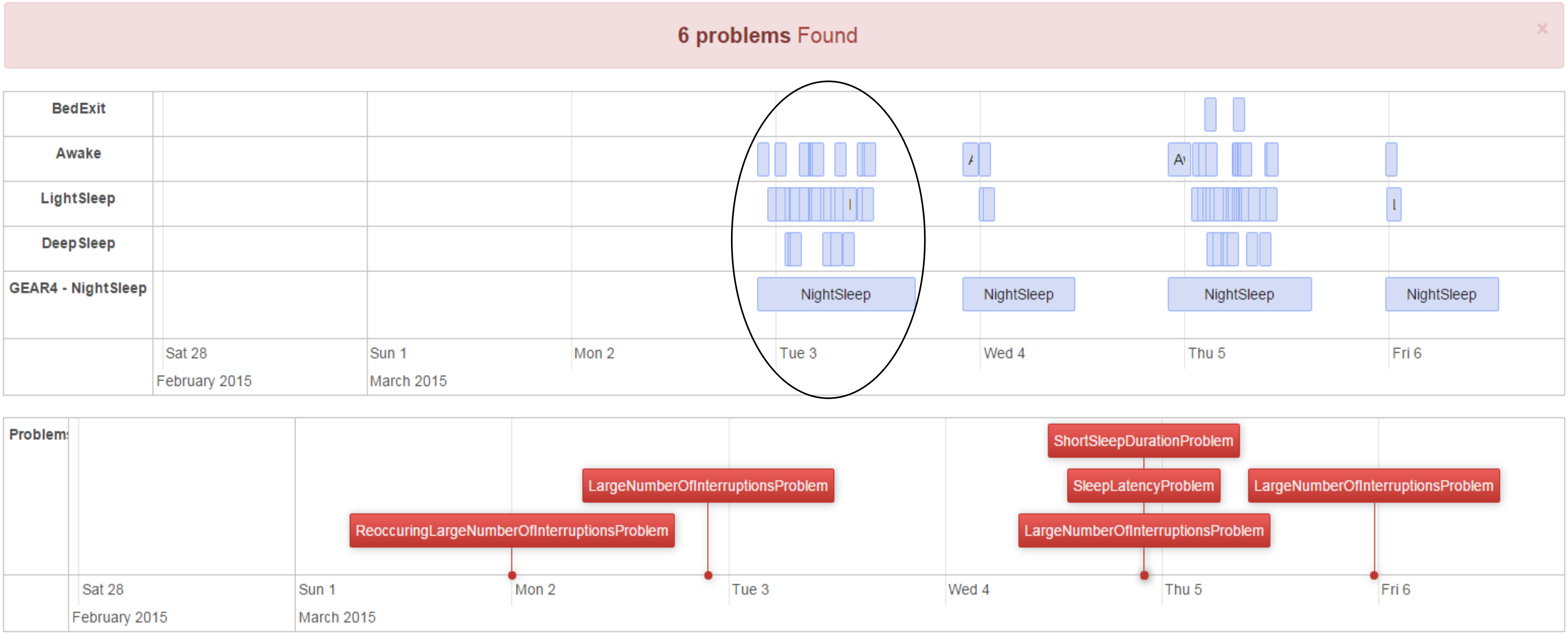
- **off-the-self sensors**
  - Accelerometer
  - Power, water monitoring
  - Motion, pressure sensor
  - RFID tags attached to objects



Activity monitoring in Nursing Home with AD patients

Visualization of bed exit at night.

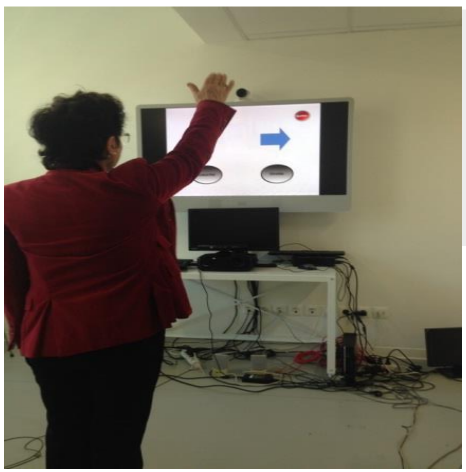
Dem@Care Clinician Interface : sleep window

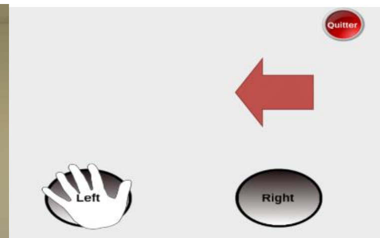


General Problem detection :

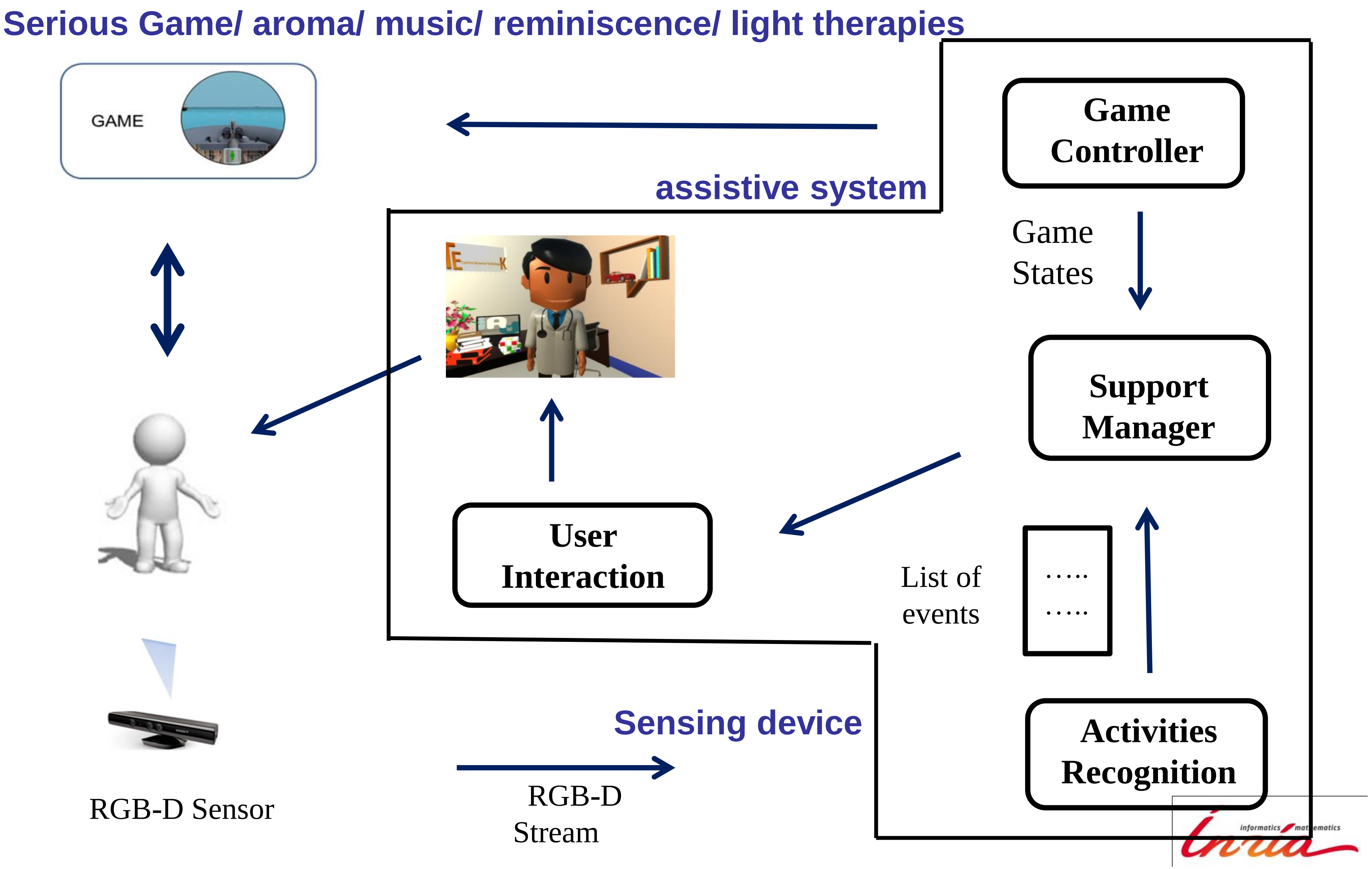
- LargeNumberOfSleepInterruptions: > 2 night sleep interruptions
- ShortSleepDuration: night sleep duration < 7 h
- SleepLatency > 30 minutes
- NapProblem: nap duration > 30 minutes
- ReoccuringLargeNumberOfSleepInterruptions:
- more than three LargeNumberOfInterruptions problems in a week.
- ReoccuringShortSleepDuration: more than three ShortSleepDuration problems in a week.
- Nocturia: > 3 night bathroom visits - Gear4 + CAR fusion

Stimulation using Serious Games and other interventions

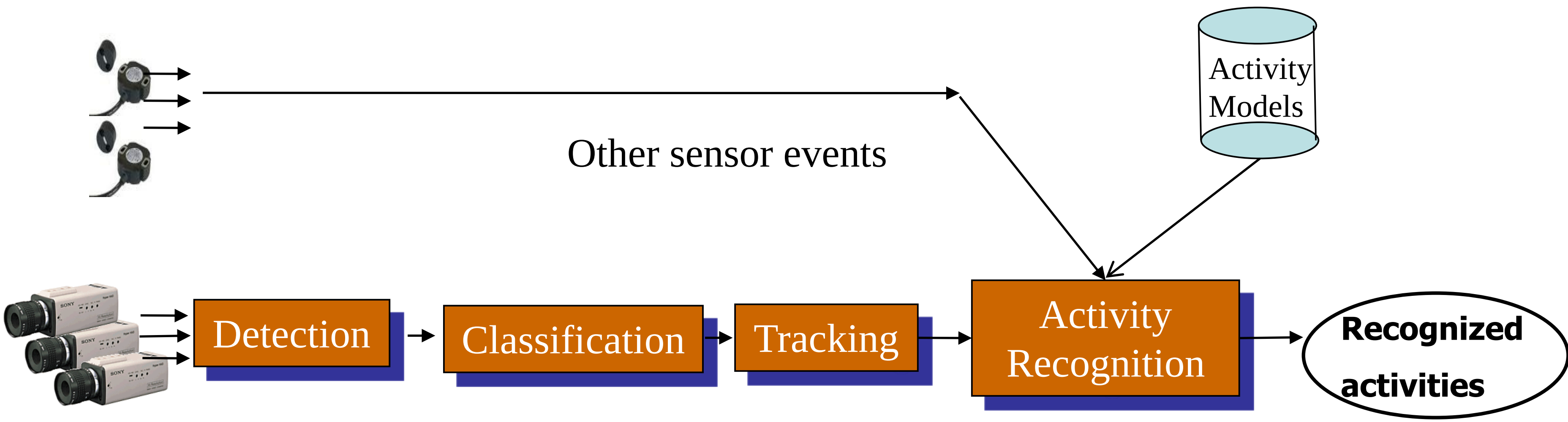




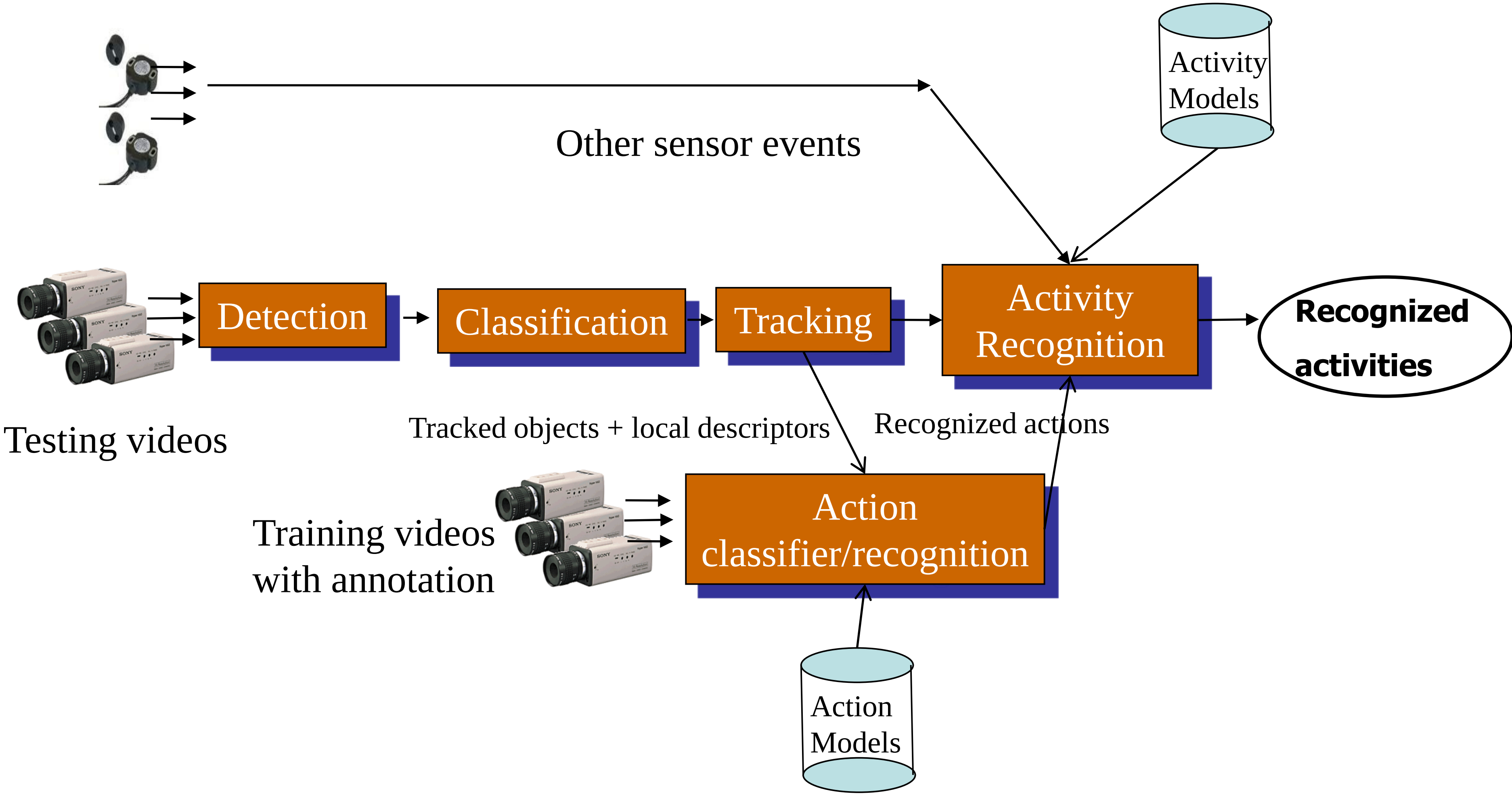
An assistive system to improve game usability for patients with cognitive disorders



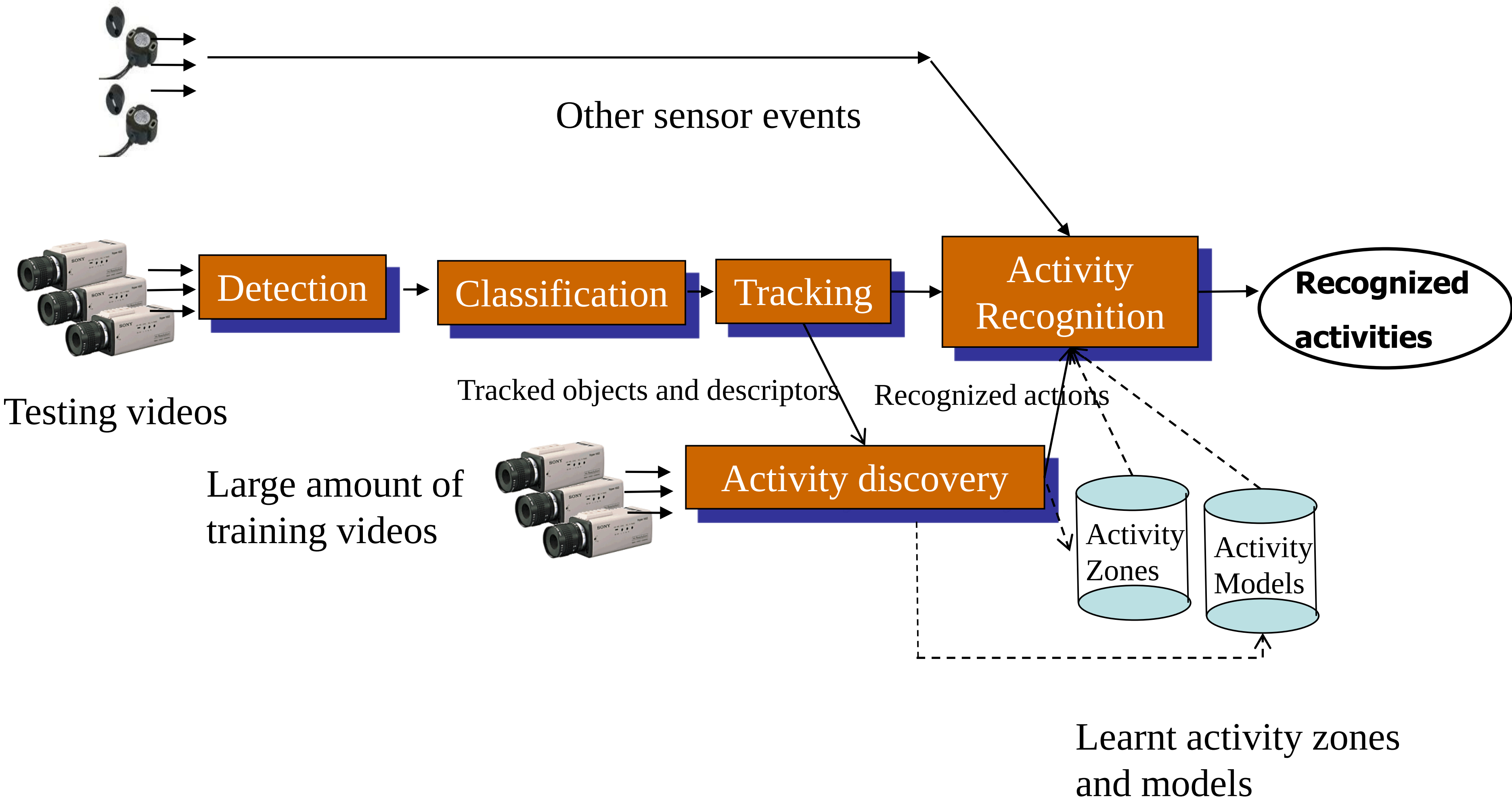
Generic Platform for activity understanding



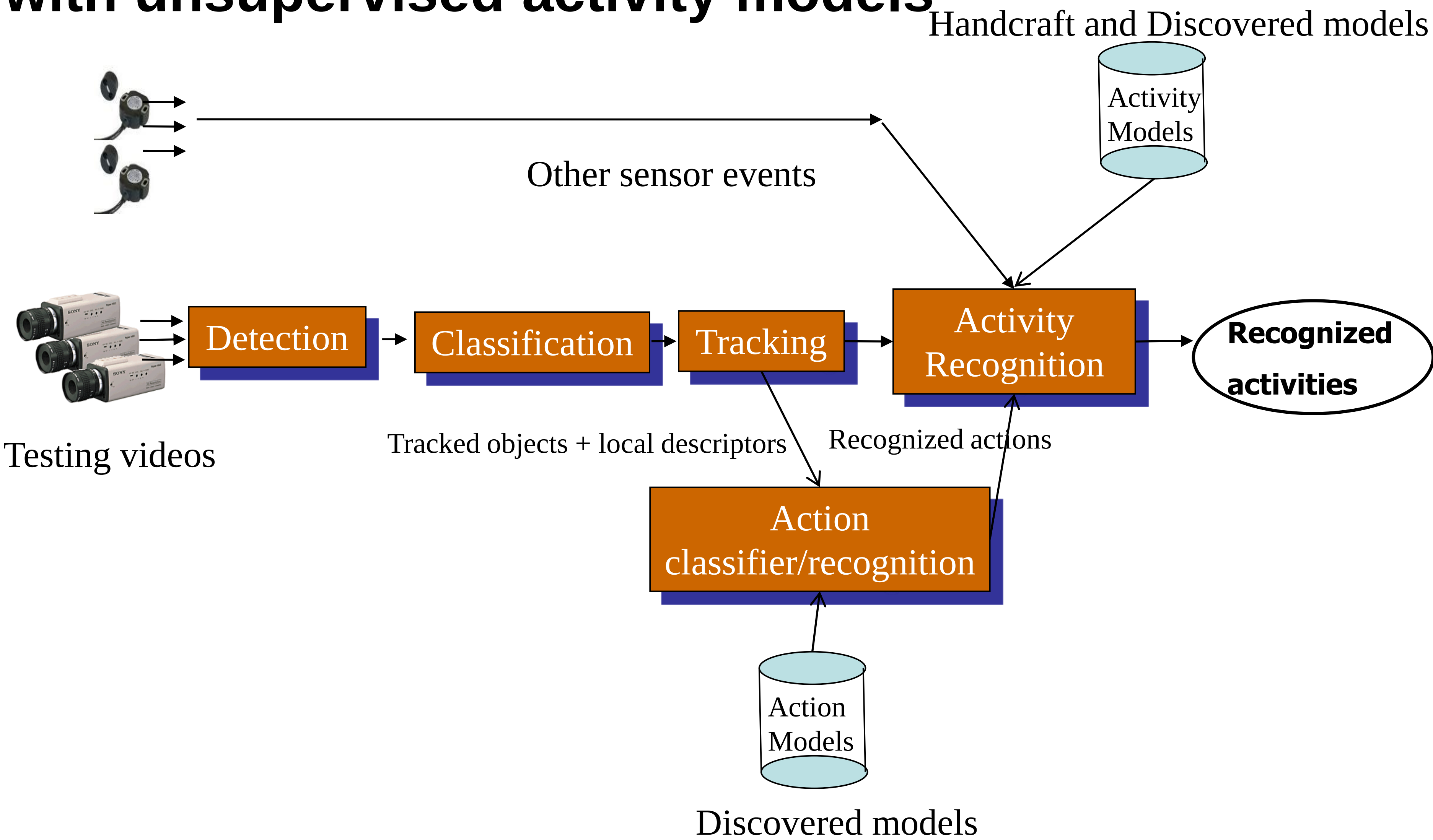
Generic Platform for activity understanding with supervised learnt actions



Generic Platform for activity understanding with unsupervised activity models



Generic Platform for activity understanding with unsupervised activity models

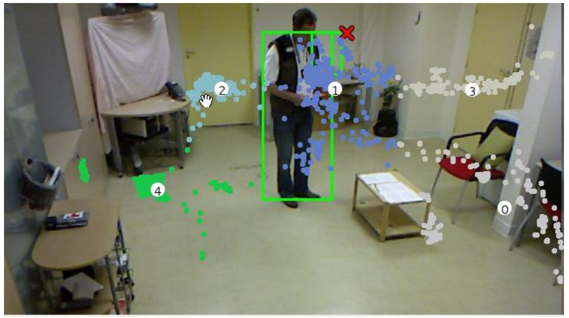


Discovering Activities

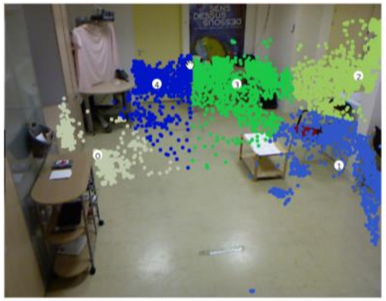
Zone Learning (Important Scene Regions) – F. Negin

Person Tracking

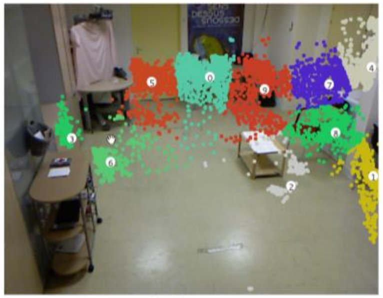
- Detect person using depth images
- Global Trajectory: track center of mass of detected person



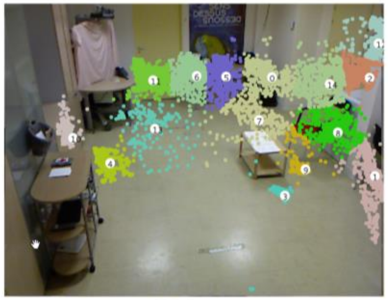
- Collect trajectories of all subjects in training set
- Cluster all trajectory points in different resolutions using k-means algorithm to find scene regions



5 clusters



10 clusters

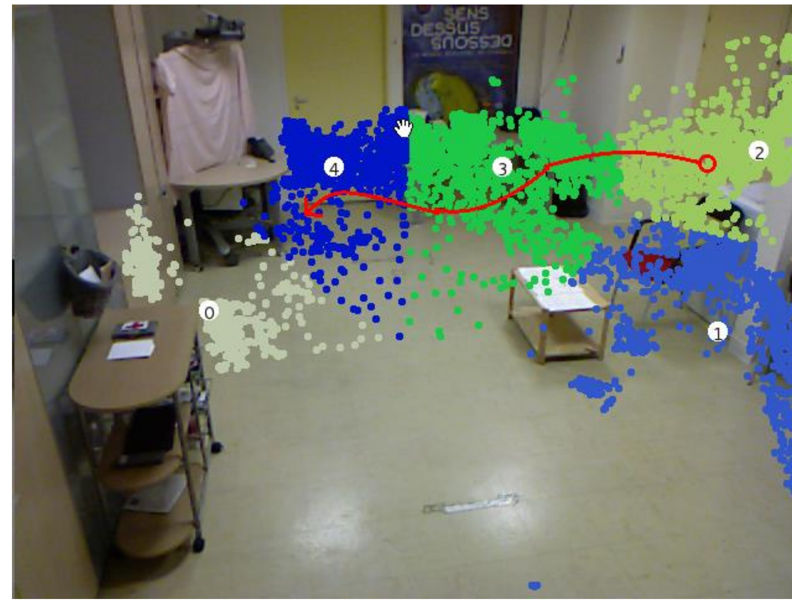


15 clusters

Discovering Activities - Activity Detection

$$Primitive\ Event = \ Change_{P-Q}$$

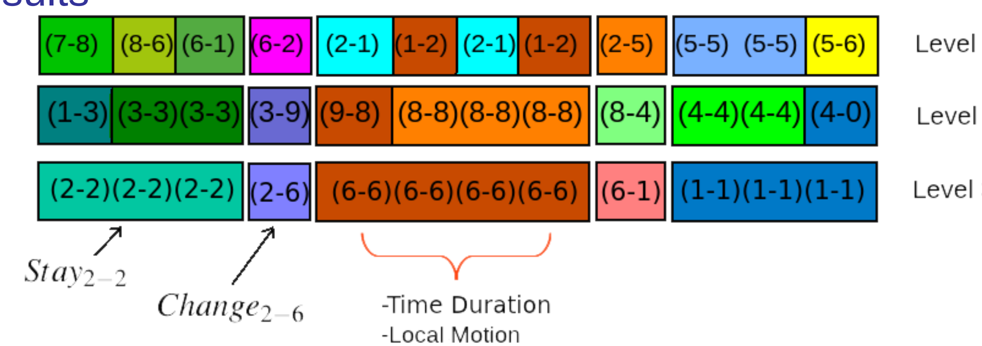
$$Primitive\ State = \ Stay_{P-P}$$



Align Tracking Information With Scene Regions

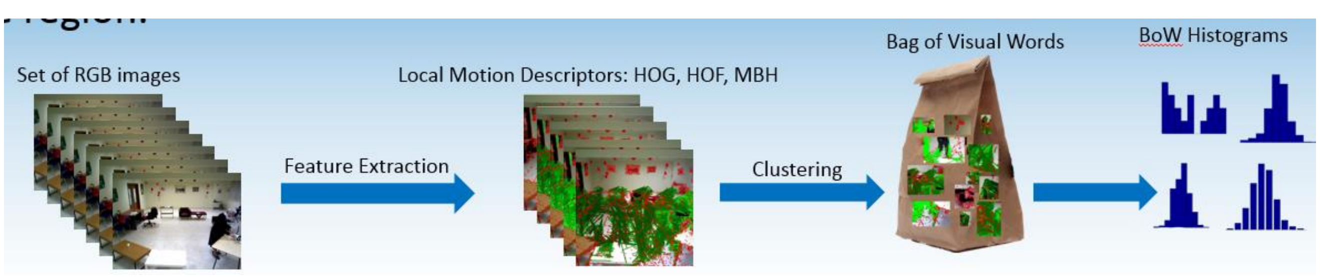
(2-2) … (2-2)(2-3)(3-3) … (3-3)(3-4)(4-4) …

Combining primitives in higher granularity results a composite event sequence called: **Discovered Activities**



Discovering Activities
Local Motion Descriptors

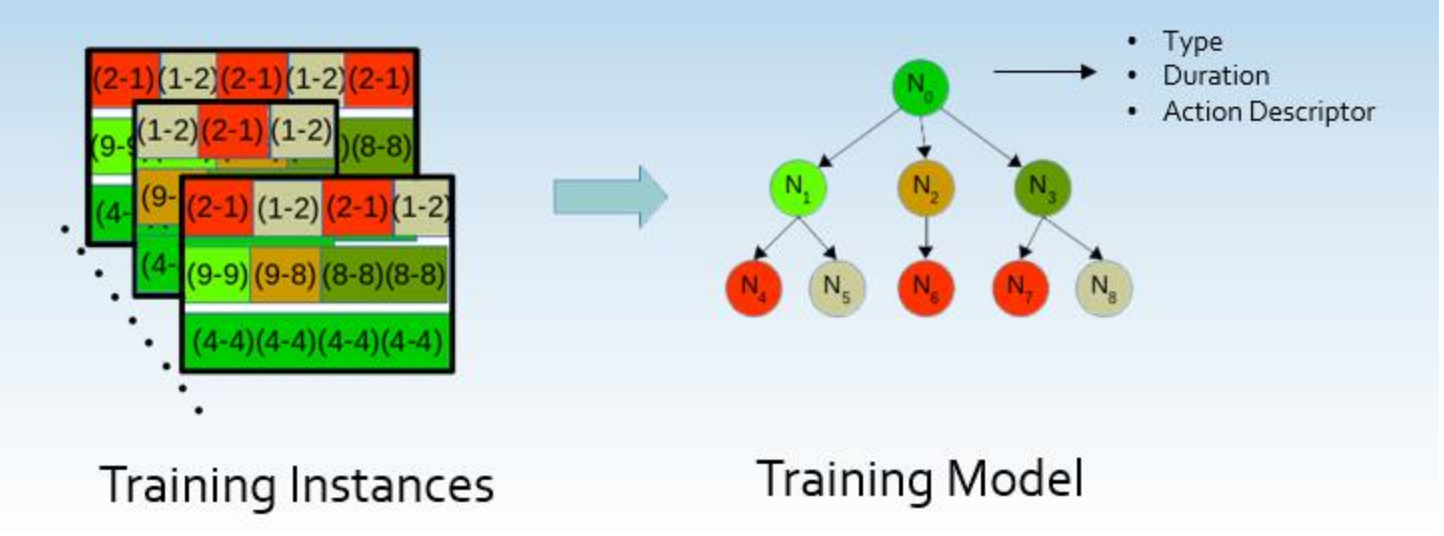
- Extract descriptors (Improved Dense Trajectories) for every discovered activity
- Calculate histograms using BoVW
- Labeling by the user (accelerated by a clustering step)
- Train a supervised classifier SVM per action class



Supervised Classifier

Discovering Activities
Training of the ACTIVITY MODELS

- Combination of structural information (global) of discovered activities and BoW histograms labels (local)



Model Training

Model Training

Subjects' trajectories

Model Training

Subjects' trajectories

Model Training

Trajectory clustering
To define scene regions

Extract information

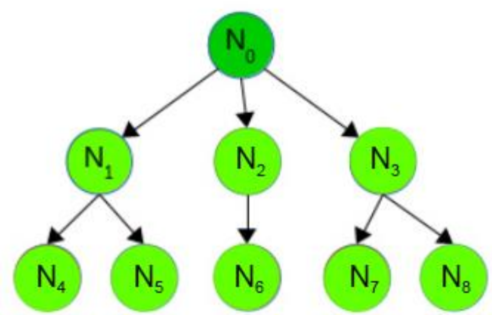
Time distribution μ, δ

Type: stay, change

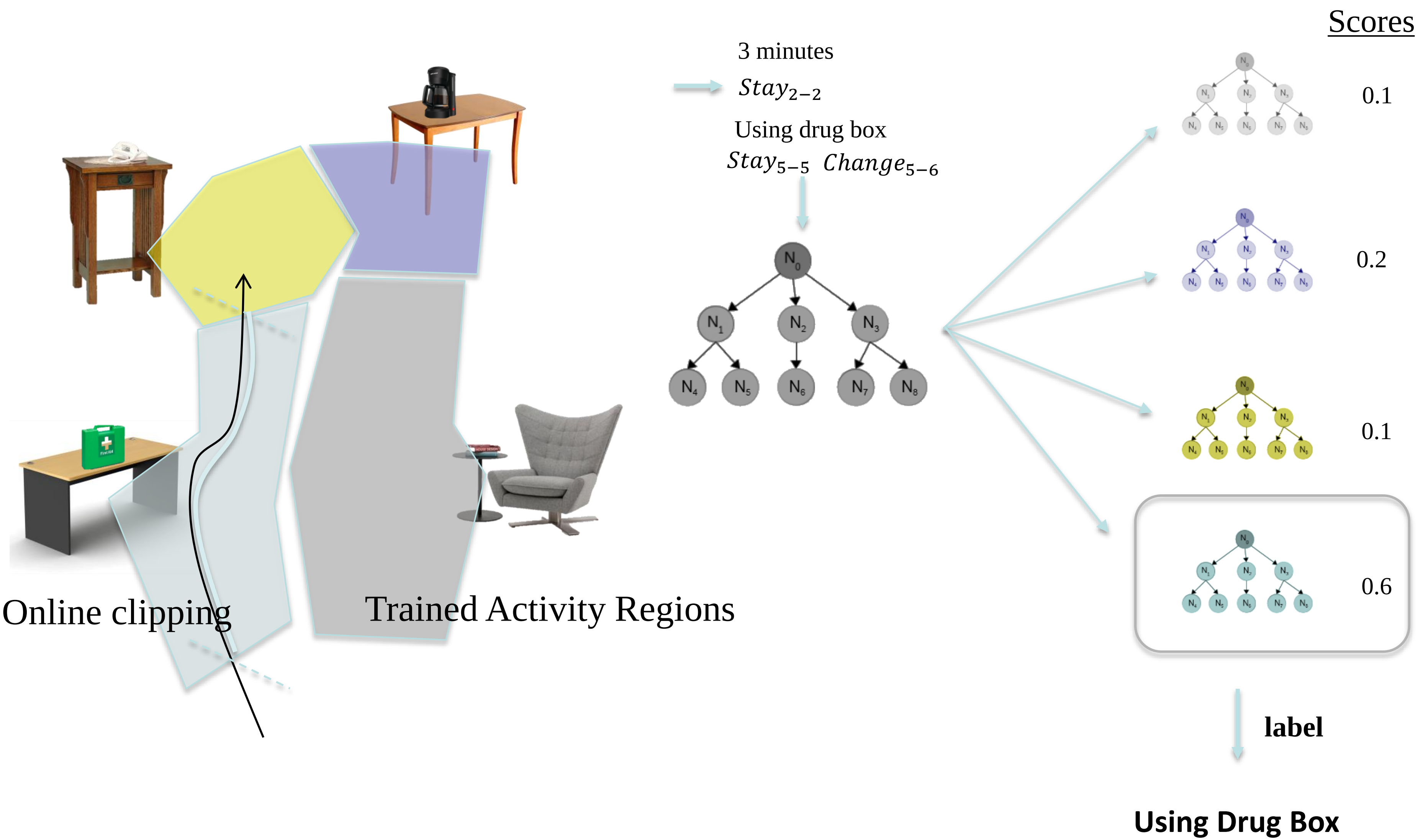
Sub-events

label

Construction of tree structure for
Activity of the
region



Testing (Online Recognition)



Discovering Activities - RESULTS

CHU

	Supervised (Manually Clipped) of [20]		Online Version of [20]		Unsupervised Using Global Motion [7]		Proposed Approach	
ADLs	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)
Answering Phone	57	78	**100**	**86**	**100**	60	**100**	81.82
P. Tea + W. Plant	89	**86.5**	76	38	84.21	80	**94.73**	81.81
Using Phar. Basket	**100**	83	**100**	43	90	**100**	**100**	**100**
Reading	35	**100**	92	36	81.82	**100**	**100**	91.67
Using Bus Map	90	**90**	**100**	50	**100**	54.54	**100**	83.34
AVERAGE	74.2	87.5	93.6	50.6	91.2	78.9	**98.94**	**87.72**

GAADRD

	Supervised (Manually Clipped) Approach [20]		Online Version of [20]		Classification by detection using SSBD [2]		Unsupervised Using Global Motion [7]		Proposed Approach	
ADLs	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)
Answering Phone	**100**	88	**100**	70	96	34.29	**100**	**100**	**100**	88
Establish Acc. Bal.	67	**100**	**100**	29	41.67	41.67	**100**	86	67	**100**
Preparing Drink	**100**	69	**100**	69	96	80	78	**100**	**100**	82
Prepare Drug Box	58.33	**100**	11	20	**86.96**	51.28	33.34	**100**	22.0	**100**
Watering Plant	54.54	**100**	0	0	**86.36**	86.36	44.45	57	44.45	**80**
Reading	**100**	**100**	88	37	**100**	31.88	**100**	**100**	**100**	**100**
Turn On Radio	60	86	**100**	75	96.55	19.86	89	89	89	89
AVERAGE	77.12	**91.85**	71.29	42.86	**86.22**	49.33	77.71	90.29	74.57	91.29

- Our approach always performs equally or better than online supervised approach. And even most of the time it outperforms totally supervised approach (manually clipped)
- Our recognition mechanism helps each element to correct others, i.e. if the classifier predicts a wrong label for a test instance, duration score or scores from sub-activities could be more informative and then turn over the final decision

[20] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In CVPR 2011.
[2] K. Avgerinakis, A. Briassouli, and I. Kompatsiaris. Activity detection using sequential statistical boundary detection (SSBD). In CVIU, 2015
[7] S. Elloumi, S. Cosar, G. Pusiol, F. Bremond, and M. Thonnat. Unsupervised discovery of human activities from long-time videos. In IET Computer Vision, 2014.

Conclusion - video understanding

A **global framework** for building real-time video understanding systems:

- 3 types of **Activity Monitoring Systems** to measure levels of everyday activities: from hand-craft to (un)supervised learned models of activity
- Robust for **long term** video monitoring
- Online and real-time recognition with limited user interaction during training

Perspectives:

- Generate totally unsupervised models
- Use finer features as input for the algorithm (head, posture, emotions, intentions…)
- Generating language description for the models (learning the semantics)
- Generic activity models (cross scenes), Adaptive learning

Conclusion for Assistive Living

Key advance : ICT software performance still needs to be measured

- Bracelets (wandering), fall detectors, serious games, low techs…
- Activity monitoring systems to measure levels of everyday activities.

Key perspectives : diagnosis, protection, engagement, empowerment

- Medical research, education : complete knowledge on AD, ageing through behavioural studies.
- Assessment : to understand behavioural disorders (sleeping disorders, apathy), frailty, disease burden. Reasons for going to institutions? (un-adapted environment)
- Tools for personalised coaching, care : links between behavioural disorders and their causes: corrective actions, carer training.
- Engagement : social interaction, initiate activities, stimulation (serious games).

Limitations:

- User-center systems : large variety of people, environment…
- ICT software : reliable, accurate, autonomous
- Local companies : Installation and maintenance of large variety of sensors

Are we addressing End-user needs?

There are several end-users in homecare:

- Doctors (gerontologists, clinicians):
  - Frailty measurement (depression, …)
  - Alarm detection (falls, gas, dementia, …).
- Caregivers and nursing home:
  - Cost reduction: no false alarm and reduction employee involvement.
  - Employee protection.
- Persons with special needs, including young children, disabled and older people:
  - Feeling safe at home.
  - Autonomy: at night, lighting up the way to bathroom.
  - Improving life: smart mirror, summary of user day, week, month, in terms of walking distance, TV, water consumption.
- Family members and relatives:
  - Older people safety and protection.
  - Social connectivity.

Practical Problems and Solutions

Problems	Solutions
Privacy confidentiality and ethics: video (and other data) recording, processing and transmission.	No video recording and transmission, only textual alarms.
Acceptability for older people	User empowerment.
Usability	Easy ergonomic interface (no keyboard, large screen), friendly usage of the system.
Cost effectiveness	The right service for the right price, large variety of solutions.
Legal issues, no certification	Robustness, benchmarking, on site evaluation
Installation, maintenance, training, interoperability with other home devices	Adaptability, X-Box integration, wireless, open standards (OSGI, …)
Research financing	Insurances, Companies or Governments : France (lobbies), Europe (not organized), US, Asia.

informatics mathematics

Monitoring of Activities of Daily Living

- Studies of older people behaviors (CoBTeK, CHU Nice, CG06...)

  - Objectif1**: living autonomously**
    - Detection of critical situations (e.g. falls, gas),
    - Objective and functional assessment of older people frailty (measurement of ADLs),
    - Detecting the deviations of a behavioral profile (missing activities, disorder, interruptions, repetitions, inactivity).
    - Building a video library of reference behaviors characterizing people frailty.

  - Objectif2: studies of behavioral disorders of **Alzheimer** patients:
    - Early diagnosis of the AD : correlation with gold standard scale,
      - Assessment scale : Alzheimer patient versus healthy older people, versus MCI…
    - Delay the admittance into the institution,
    - Monitor and assess the degree of dementia (impact of drugs, therapies).

  - Objectif3: design **sensor-based** systems : video, RGBD cameras
    - Ambient sensors : pressure, contact, RFID, environmental…
    - Wearable sensors : video cameras, accelerometers, physiological,…
    - microphones

  - Objectif4: **evaluation** platform for geron-technologies,
    - Ecological and clinical experimentations
      - in laboratory, at Hospital, Nursery Home and at regular Home
      - Over extensive duration (months).