Video Understanding for Activity Recognition

Francois BREMOND
STARS project-team,
INRIA Sophia Antipolis, FRANCE

Francois.Bremond@inria.fr

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

Objective: Designing systems for Real time recognition of human activities observed by 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 Applications

• Strong impact for visual surveillance in transportation (metro station, trains, airports, aircraft, harbors)

• Control access, intrusion detection and Video surveillance in building

• Traffic monitoring (parking, vehicle counting, street monitoring, driver assistance)

• Bank agency monitoring

• Risk management (3D virtual reality simulation for crisis management)

• Video communication (Mediaspace)

• Sports monitoring (Tennis, Soccer, F1, Swimming pool monitoring)

• New application domains: Aware House, Health (HomeCare), Teaching, Biology, Animal Behaviors, ...

➢ Creation of a start-up Keeneo July 2005 (20 persons):  http://www.keeneo.com/
Video Understanding: Issues

Practical issues

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

- Tiny objects
- Strong perspective
- Shadows
- Lighting conditions
- Close view
- Clutter
Video Understanding

Objective: *Real-time Interpretation of videos from pixels to events*
Video Understanding

Objective: *Real-time Interpretation of videos from pixels to events*

\[ [a_1, \ldots, a_m] \]

\[ V_t \rightarrow \text{Segmentation} \rightarrow \text{Classification} \rightarrow \text{Tracking} \rightarrow \text{Scenario Recognition} \rightarrow \text{Alarms} \]

\[ f_1 \rightarrow f_2 \rightarrow f_3 \rightarrow f_n \]

Inria
Global framework for video understanding

Video processing

Sensing data, signal

Interpretation at time $t$:
moving objects, metadata, events, …

$$f_n \circ f_{1[a_1, \ldots, a_m]} (V_{[t-1,t]}, I_{t-1}) = I_t$$

Processing Parameters: thresholds, reference image,…
Contextual Information: sensor, static scene model,…
Knowledge: physical object models, scenario models,…
Video Understanding: Approach

5 Challenges in video understanding

1) Robustness of Video Processing depending on data domains

2) spatio-temporal reasoning, uncertainty and semantics

3) Evaluation, Ground-truth, Metrics, videos

4) Knowledge representation, learning

5) Configuration, optimisation, system generation from specification

Video processing

Sensing data, signal

Interpretation at time t:
moving objects, metadata, events, …

\[ f_n \circ f_1[a_1, \ldots a_m] (V_{t-1:t}, I_{t-1}) = I_t \]

Processing Parameters: thresholds, reference image,…
Contextual Information: sensor, static scene model,…
Knowledge: physical object models, scenario models,…

Video processing

Processing Parameters: thresholds, reference image,…
Contextual Information: sensor, static scene model,…
Knowledge: physical object models, scenario models,…
Video Understanding

Outline:

• Knowledge Representation : Scene Model
• Input of the Scenario Recognition process:
  • Object Detection, Object Tracking, Action Recognition
• Event/Scenario Representation
• Bag of Words
• Graphical models
• Temporal Scenario Recognition
  • Scenario representation
  • Recognition process
• Applications: recognition of several scenarios
• Learning Scenario Models
Knowledge Representation
Knowledge Representation

**A priori knowledge**

- Sensor information
- Mobile object classes
- Tracked object types
- 3D Scene Model
- Descriptions of event recognition routines
- Scenario library

**Video streams**

- Moving region detection
- Mobile object tracking

**Recognition of primitive states**

- Recognition of scenario 1
- Recognition of scenario 2
- ... (ellipses)
- Recognition of scenario n

**Scenario recognition module**

- Recognised scenario
Knowledge Representation: 3D Scene Model - Context

**Definition:** a priori knowledge of the observed empty scene

- **Cameras:** 3D position of the sensor, calibration matrix, field of view,...
- **3D Geometry of physical objects** (bench, trash, door, walls) and interesting zones (entrance zone) with position, shape and volume
- **Semantic information:** type (object, zone), characteristics (yellow, fragile) and its function (seat)

**Role:**

- to keep the interpretation independent from the sensors and the sites: many sensors, one 3D referential
- to provide additional knowledge for behavior recognition
Knowledge Representation: 3D Scene Model

3D Model of 2 bank agencies

Les Hauts de Lagny

Villeparisis
Object (People) detection

Estimation of Motion

• Need of textured objects
  • Optical Flow: Estimation of apparent motion (pixel intensity between 2 frames)
  • Local descriptors (patches, tracklets, gradients (SURF, HOG), color histograms, moments over a neighborhood)

Object model

• Need of mobile object model
  • 2D appearance model (shape, color, pixel template)
  • 3D articulate model

Reference image subtraction

• Need of static cameras
  • Most robust approach (model of background image)
  • Most common approach even in case of PTZ, mobile cameras
Issues in Local Descriptor People Detection:

- **Features:**
  - HOG, LBP, Covariance Matrix, Haar, SIFT, Granules

- **Learning paradigm:**
  - Adaboost, Hierarchical trees, SVM

- **Training / testing databases:**
  - Camera view point, distortion, resolution,
  - Occlusion, pose,
  - Background samples

- **Processing time:**
  - Training (best feature selection)
  - Detection (scanning window sampling rate, multi-resolution)

- **Filtering:**
  - Overlapping scanning window, candidate selection
  - 3D constraint, motion segmentation,

- **Body parts:**
  - Global detection
  - Model based association
  - E.g. head, torso, legs...
Complex Scenes: People detection

Body part combination

- **Body parts** combination:
  - Detected body parts (HOG detector trained on manually selected areas of the person)
  - Example below in TrecVid camera 1

Example of detected with corresponding HOG cells

Detection examples
Head detection and tracking results

Training head database: selection of 32x32 head images from publicly available MIT, INRIA and NLDR datasets. A total of 3710 images were used.

Training background dataset: selection of 20 background images of TrecVid and 5 background images of Torino ‘Biglietattrice.

Speed: Once integral images are computed, the algorithm reaches ~ 1fps for 640x480 pixels.

Left: head detection examples and right: tracking examples in Torino underground.
Posture Recognition: Set of Specific Postures

Hierarchical representation of postures
Posture Recognition: silhouette comparison

Real world

Virtual world

Generated silhouettes
Posture Recognition: results
Event/Scenario Recognition

Outline:

• Event/Scenario Representation
• Bag of Words
• Graphical models
• Temporal Scenario Recognition
  • Scenario representation
  • recognition process
• Applications: recognition of several scenarios
• Learning Scenario Models
Event Representation: Video Event Ontology

Definition:

• Video Event Ontology: a set of concepts and relations is used as a reference between all the actors of the domain to describe knowledge

Properties:

• Enable experts to describe video events of interest (e.g. composite event) and to structure the knowledge: ontology of the application domain.
• Share knowledge between developers: ontology of visual concepts (e.g. a stopped mobile object)
• Ease communication between developers and end users and enable performance evaluation: ontology of the video understanding process (what should be detected: mobile object (a parked car), object of interest (a door), visible object (occluded person))
• Architecture interoperability: separation between specification and knowledge description
Event Representation

Several entities are involved in the scene understanding process:

• Moving region: any intensity change between images.

• Context object: predefined static object of the scene environment (entrance zone, wall, equipment, door...).

• Physical object: any moving region which has been tracked and classified (person, group of persons, vehicle, … etc).

• Physical object of interest: meaningful object, but depending on applications (person/ door, parked vehicle, … etc).
Event Representation

Actions, States, Streams, Threads, Events, Situations, chronicles, behaviours, activities and scenarios...: a large variety

- more or less composed of sub-events (running/fighting).
- involving few/many actors (football game).
- general (standing)/sensor and application/view (sit down, stop) dependent.
- spatial granularity: the view observed by one camera/the whole site.
- temporal granularity: instantaneous/long term with complex relationships (synchronize).

3 levels of complexity depending on the complexity of temporal relations and on the number of physical objects:

- non-temporal constraint relative to one physical object (sitting). Intuitive combination of feature probabilities to get better precision.
- temporal sequence of sub-scenarios relative to one physical object (open the door, go toward the chair then sit down). Filtering noisy input, versus meaningful changes.
- complex temporal constraints relative to several physical objects (A meets B at the coffee machine then C gets up and leaves). Need of logic reasoning (declarative, expressive) but sensitive to vision errors.
Event Representation

Video events: real world notions corresponding to short actions (coherent unit of motion) up to activities.

- **Primitive State:** a spatio-temporal property linked to vision routines involving one or several actors, valid at a given time point or stable on a time interval
  
  Ex: « close», « walking», « sitting»

- **Composite State:** a combination of primitive states

- **Primitive Event:** significant change of states
  
  Ex: « enters», « stands up», « leaves »

- **Composite Event:** a combination of states and events. Corresponds to a long term (symbolic, application dependent) activity.
  
  Ex: « fighting», « vandalism»
Event Recognition

Several formalisms can be used:

**Event representation:**
- n-ary tree, frame, aggregate (structure).
- finite state automaton, sequence (evolution).
- graph, set of constraints.

**Event recognition:**
- Feature based routine.
- Classification, Bayesian, neural network, SVM, clustering, BoW.
- DBN, HMM, Petri net.
- Stochastic grammar, Prolog.
- Constraint propagation, verification of temporal constraints.
Event Recognition: Issues

Performance: robustness of real-time (vision) algorithms

Bridging the gaps at different abstraction levels:
  • From sensors to image processing
  • From image processing to 4D (3D + time) analysis
  • From 4D analysis to semantics

Uncertainty management:
  • uncertainty management of noisy data (imprecise, incomplete, missing, corrupted)
  • formalization of the expertise (fuzzy, subjective, incoherent, implicit knowledge)

Independence of the models/methods versus:
  • Sensors (position, type), scenes, low level processing and target applications
  • several spatio-temporal scales

Knowledge management:
  • Bottom-up versus top-down, focus of attention
  • Regularities, invariants, generic models and context awareness
  • Knowledge acquisition versus (none, semi)-supervised, incremental learning techniques
  • Formalization, modeling, ontology, standardization
# Action Recognition (MB. Kaaniche, P. Bilinski)

<table>
<thead>
<tr>
<th>Walking</th>
<th>Jogging</th>
<th>Running</th>
<th>Boxing</th>
<th>Hand waving</th>
<th>Hand clapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image](Image615x2 to 718x39)</td>
<td>![Image](Image64x101 to 692x418)</td>
<td><img src="Image696x517" alt="Image" /></td>
<td><img src="Image28" alt="Image" /></td>
<td><img src="Image153x76" alt="Image" /></td>
<td></td>
</tr>
</tbody>
</table>

Type of gestures and actions to recognize
Action Recognition Algorithms

Videos → Point detector → Point descriptor → All feature vectors → Codebook generation → BOW model
Bag-of-words model

Offline Learning:

All training feature vectors → Codebook generation (different sizes) → Database

Online recognition:

All testing feature vectors → Assignment to the closest codeword → Histogram of codewords → Non-linear SVM
ADL Dataset
## ADL - Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Size 1000</td>
<td>85.33%</td>
<td>90.00%</td>
<td>94.67%</td>
<td>92.00%</td>
</tr>
<tr>
<td>Size 2000</td>
<td>88.67%</td>
<td>90.00%</td>
<td>92.67%</td>
<td>91.33%</td>
</tr>
<tr>
<td>Size 3000</td>
<td>83.33%</td>
<td>89.33%</td>
<td>94.00%</td>
<td>90.67%</td>
</tr>
<tr>
<td>Size 4000</td>
<td>86.67%</td>
<td>89.33%</td>
<td>94.00%</td>
<td>85.00%</td>
</tr>
<tr>
<td>Best</td>
<td>88.67% (4)</td>
<td>90.00% (3)</td>
<td>94.67% (1)</td>
<td>92.00% (2)</td>
</tr>
</tbody>
</table>

Best (7% diff)

SOA: 96% Wang [CVPR11]
Issues in Action Recognition

- Different detectors (Hessian, Dense sampling, STIP...)
- Different parameters of descriptors (grid size, ...)
- Different classifiers (k-NN, linear-SVM, ...)
- Different clustering algorithms (Bossa Nova, Fisher Kernels,...)
- Different resolutions of videos
- Generic to other datasets (IXMAS, UCF Sports, Hollywood, Hollywood2, YouTube, ...)
- Finer actions, more discriminative, without context...
Issues in Action Recognition

- Finer actions, more discriminative
Event Recognition: Specific Routines
Advisor project: F. Cupillard, A. Avanzi,…
Event Recognition: Specific Routines
Results in metro station

Scenario Recognition:
Running

Scenario: Running
-> ALARM
State: walking
State: stopped
Event Recognition: Specific Routines
Results in metro station

Scenario: Agitated Behaviour
- > ALARM

State: Lively
Event Recognition: automaton

The scenario “A Group of people blocks an Exit” is based on a Finite state automaton.
Event Recognition: Brussels and Barcelona Metros

- **Group behavior**
  - Blocking

- **Crowd behavior**
  - Overcrowding

- **Group behavior**
  - Fighting

- **Individual behavior**
  - Jumping over barrier
Event Recognition using Posture
Event Recognition : automaton


• Tested on 50 metro sequences (10 hours) and one week live recognition

• True positive per sequence: 70% (“Fighting”) to 95% (“Blocking”)

• False positive per sequence: 5% (“Fighting”, “Jumping over barrier”) to 0% (others)

However:
• Sensitive to noise
• Difficulties to tune to get best performance
Scenario Recognition: Temporal Constraints
Work done in collaboration with T. Vu
Event Representation

Representation Language to describe Temporal Events of interest.

A video 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
Event Representation

Representation Language to describe Temporal Events of interest.
Example: a “Bank_Attack” scenario model

composite-event (Bank_attack,
   physical-objects ((employee : Person), (robber : Person))
   components(
      (e1 : primitive-state inside_zone (employee, "Back"))
      (e2 : primitive-event changes_zone (robber, "Entrance", "Infront"))
      (e3 : primitive-state inside_zone (employee, "Safe"))
      (e4 : primitive-state inside_zone (robber, "Safe"))
   )
   constraints ((e2 during e1)
      (e2 before e3)
      (e1 before e3)
      (e2 before e4)
      (e4 during e3) )
   action ("Bank attack!!!") )
Scenario Representation

A “Bank attack” scenario instance

(4) Both of them arrive at the safe door.
Scenario Recognition: Temporal Constraints

- Overview of the recognition process
- Recognition of elementary scenarios
- Scenario compilation
- Recognition of composed scenarios
- Prediction and uncertainty
- Example of the recognition of a “Bank attack” scenario and more…
Scenario Recognition: Temporal Constraints (T. Vu)

- **Scenario** (algorithmic notion): any type of video events

- Two types of scenarios:
  - **elementary** (primitive states)
  - **composed** (composite states and events).

- Algorithm in **two steps**.

- **A priori Knowledge**
  - Scenario knowledge base
  - 3D geometric & semantic information of the observed environment

**Diagram**

1. Recognize all **Elementary Scenario models**
2. Trigger the recognition of selected **Composed Scenario models**

1. Recognize all **triggered Composed Scenario models**
2. Trigger the recognition of other **Composed Scenario models**

Recognized Scenarios
Elementary Scenario Recognition
Example: a scenario model & an observed environment

Scenario(Working at Machine, physical-objects(p : Person, e : Machine, z : Zone)
constraints(
(1) \((\text{height of } p \leq 170)\)
(2) \(((p \text{ in } z) \& (\text{name of } z = \text{"Machine zone"}))\)
(3) \((\text{distance}(p, e) \leq \text{close_distance})\)
)

zone: Entrance zone \(z_1\)
zone: Waiting zone \(z_2\)
zone: Machine zone \(z_3\)
machine: \(m\)
Elementary Scenario Recognition

Example: a situation

Scenario \((\text{Working}_\text{at}_\text{Machine}, \text{physical-objects}(p : \text{Person, } e : \text{Machine, } z : \text{Zone})\)

- Problem: [Rota, 2001] attempts all combinations of physical objects \(\Rightarrow\) combinatorial explosion.
- Solution: reorganize the knowledge represented in an elementary scenario model \(\Rightarrow\) elementary scenario model compilation.

Recognized scenario: \(\text{Working}_\text{at}_\text{Machine}(p_4, m, z_3)\)
Scenario Recognition: Elementary Scenario

- Object of a domain
- Satisfied constraint
- Unsatisfied constraint
- Recognized scenario

**Recognized scenario:**
*Working_at_Machine(p₄, m, z₃)*
Scenario Recognition: Elementary Scenario

- The recognition of an elementary scenario model $m_e$ consists of a loop:

  1. Choosing a physical object for each physical-object variable

  2. Verifying all constraints linked to this variable

$m_e$ is recognized if all the physical-object variables are assigned a value and all the linked constraints are satisfied.
Scenario Recognition: Composed Scenario

• Problem:

  given a scenario model \( m_c = (m_1 \text{ before } m_2 \text{ before } m_3) \);
  if a scenario instance \( i_3 \) of \( m_3 \) has been recognized
  then the main scenario model \( m_c \) may be recognized.

  However, the classical algorithms will try all combinations of scenario instances (already recognised) of \( m_1 \) and of \( m_2 \) with \( i_3 \)

  \( \Rightarrow \) a combinatorial explosion in the past.

• Solution:

  decompose the composed scenario models into simpler scenario models
  in an initial (compilation) stage such as each composed scenario model is
  composed of two components: \( m_c = (m_4 \text{ before } m_3) \)

  \( \Rightarrow \) a linear search in the past.
Scenario Recognition: Composed Scenario

Example: original “Bank_attack” scenario model

\(\text{composite-event}(\text{Bank_attack},\)
\(\text{physical-objects}((\text{employee : Person}), (\text{robber : Person}))\)
\(\text{components}(\)
   \(1\) (\(e_1 : \text{primitive-state } \text{inside\_zone}(\text{employee, "Back"})\))
   \(2\) (\(e_2 : \text{primitive-event } \text{changes\_zone}(\text{robber, "Entrance", "Infront"})\))
   \(3\) (\(e_3 : \text{primitive-state } \text{inside\_zone}(\text{employee, "Safe"})\))
   \(4\) (\(e_4 : \text{primitive-state } \text{inside\_zone}(\text{robber, "Safe"})\))\)
\(\text{constraints}((e_2 \text{ during } e_1)\)
   (\(e_2 \text{ before } e_3)\)
   (\(e_1 \text{ before } e_3)\)
   (\(e_2 \text{ before } e_4)\)
   (\(e_4 \text{ during } e_3)\)
\(\text{alert(“Bank attack!!!”) })\)
Scenario Recognition: Composed Scenario

Compilation: Original scenario model is decomposed into 3 new scenarios

composite-event(\textit{Bank\_attack\_1},
    physical-objects((employee : Person), (robber : Person))
    components(
        (1)  (e1 : primitive-state \textit{inside\_zone} (employee, "Back"))
        (2)  (e2 : primitive-event \textit{changes\_zone} (robber, "Entrance", "Infront"))
    constraints((e1 during e2) )
)

composite-event(\textit{Bank\_attack\_2},
    physical-objects((employee : Person), (robber : Person))
    components(
        (3)  (e3 : primitive-state \textit{inside\_zone} (employee, "Safe"))
        (4)  (e4 : primitive-state \textit{inside\_zone} (robber, "Safe"))
    constraints((e3 during e4) )
)

composite-event(\textit{Bank\_attack\_3},
    physical-objects((employee : Person), (robber : Person))
    components(
        (att\_1 : composite-event \textit{Bank\_attack\_1} (employee, robber))
        (att\_2 : composite-event \textit{Bank\_attack\_2} (employee, robber))
    constraints(((termination of att\_1) before (start of att\_2)) )

alert(“Bank attack!!!” )

Scenario Recognition: Composed Scenario

- A compiled scenario model $m_c$ is composed of two components: start and termination.
- To start the recognition of $m_c$, its termination needs to be already instantiated.

- The recognition of a compiled scenario model $m_c$ consists of a loop:
  1. Choosing a scenario instance for the start of $m_c$,
  2. Verifying the temporal constraints of $m_c$,
  3. Instantiating the physical-objects of $m_c$ with physical-objects of the start and of the termination of $m_c$,
  4. Verifying the non-temporal constraints of $m_c$.
  5. Verifying forbidden constraints
**Scenario Recognition: Composed Scenario**

<table>
<thead>
<tr>
<th></th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$t_3$</th>
<th>$t_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>inside_zone</td>
<td>(P1, BC)</td>
<td>(P2, EZ)</td>
<td>(P2, IC)</td>
<td>(P1, Safe)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(P2, Safe)</td>
</tr>
<tr>
<td>changes_zone</td>
<td></td>
<td></td>
<td></td>
<td>(P1, BC, Safe)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(P2, IC, Safe)</td>
</tr>
<tr>
<td>Bank_attack_1</td>
<td></td>
<td></td>
<td>(P1, P2)</td>
<td></td>
</tr>
<tr>
<td>Bank_attack_2</td>
<td></td>
<td></td>
<td></td>
<td>(P1, P2)</td>
</tr>
<tr>
<td>Bank_attack_3</td>
<td></td>
<td></td>
<td></td>
<td>(P1, P2)</td>
</tr>
<tr>
<td>(Bank_attack)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **BC**: Back_Counter
- **IC**: Infront_Counter
- **EZ**: Entrance_Zone

- : the scenario instance that triggers the recognition of a composed scenario instance ended by it.
- the start of a composed scenario instance.
Scenario Recognition: Temporal Constraints

- The resolution of temporal constraints is improved by structuring the search domain of already recognized states, events and scenarios.

- The path (Person 1 → Inside_zone → Zone 1) shows the list of time intervals while Person 1 is inside Zone 1.
Scenario recognition: capacity of prediction

- **Issue:** in the bank monitoring application, an alert “Bank attack!!!” is triggered when a scenario “Bank_attack” is completely recognized. However, it can be too late for security agents to cope with the situation.

- **Requirement:** is the temporal scenario recognition method able to predict scenarios that may occur in the near future?

- **Answer:**
  - Yes, with some probabilities.
  - The recognition algorithm can predict scenarios that may occur by adding automatically alerts (during the compilation) to some generated partial scenario models. This task can be specified in the scenario models.
Scenario recognition : uncertainty

• Temporal precision
  • Issue: several scenario models are defined with too precise temporal constraints ⇒ they cannot be recognized with real videos.
  • Solution: we defined a temporal tolerance $\Delta t$ as an integer, then all temporal comparisons are estimated using an approximation of $\Delta t$.

• Incorrect mobile object tracking
  • Issue: the vision algorithms may loose the track of several detected mobile objects ⇒ the system cannot recognize correctly scenario occurrences in several videos.
  • Solution1: experts describe different scenario models representing various situations corresponding to several combinations of physical objects.
Uncertainty Representation

Solution 2: management of the vision uncertainty (likelihood):

- within predefined event models (off-line)
  - coefficients (on mobile objects and components) are provided by default.
  - Several notions of uncertainty (data, model, process) and utility.

- propagated (on-line) through the event instances
  1. mobile objects: computed by vision algorithms.
  2. primitive states (elementary):
     - a coefficient to each physical object for representing the likelihood relation between the state and each involved mobile object.
  3. events and composite states (composed):
     - a coefficient to each component for representing the likelihood relation between the event and each component.
       - defining a threshold into each state/event model for specifying at which likelihood level the given state/event should be recognized.
Uncertainty Representation

Combination of detection likelihood / confidence and utility:

PrimitiveState (**Person_Close_To_Vehicle**,
  Physical Objects ( (p : Person, 0.7), (v : Vehicle, 0.3) )
  Constraints ((p distance v ≤ close_distance)
    (recognized if likelihood > 0.8)) )

CompositeEvent (**Crowd_Splits**,
  Physical Objects ( (c1: Crowd, 0.5), (c2 : Crowd, 0.5), (z1: Zone) )
  Components ( (s1 : CompositeState Move_toward (c1, z1), 0.3)
    (e2 : CompositeEvent Move_away (c2, c1), 0.7) )
  Constraints ( (e2 during s1)
    (c2's Size > Threshold)
    (recognized if likelihood > 0.8)) )
Scenario recognition: Results

Evaluation: the experts of 20 projects in video interpretation have realized three types of tests.

- on recorded videos: to verify whether the recognition algorithm can recognize effectively scenario occurrences (correct detections).
- on live videos: to verify whether the recognition algorithm can work on a longtime interval (no false alarms).
- on recorded/simulated videos: to estimate the processing time and efficiently of the recognition algorithm.
Scenario recognition: Results
Experiment 1: recorded videos

- many sites: 2 bank agencies, several metro stations, a train and an airport...
- Bank: 27 recorded positive videos and many negative videos.
- 40 original scenario models (before the compilation): “inside_zone”, “Bank_attack”, “Vandalism”,...
- The algorithm fails to recognize some scenario occurrences only when the vision module fails to detect the mobile objects in the scene.
- No false alarm has been reported during all the experiments.

<table>
<thead>
<tr>
<th></th>
<th>Number of tested sequences</th>
<th>Average number of persons/frame</th>
<th>Recognition rate (%)</th>
<th>Number of false alarms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank 1</td>
<td>10</td>
<td>4</td>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>Bank 2</td>
<td>1</td>
<td>2</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Metros</td>
<td>4</td>
<td>2</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Apron</td>
<td>10</td>
<td>1</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Train</td>
<td>2</td>
<td>4</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>
Scenario recognition: Results
Experiment 2: live-videos

- 4 sites: 2 bank agencies, two offices, a parking and a metro station.
- 40 original scenario models (before decomposition): “inside_zone”, “Bank_attack”, “Vandalism”,...
- Results:
  - in a bank (5 days),
  - in an office (4h),
  - one week in a metro station of Barcelona,
  - in a parking (1 day)
  - the scenarios were most of the time (95%) correctly recognized (as in the first experiment) ➔ the recognition algorithm can work reliably and robustly in real-time and in a continuous mode.
Scenario recognition: Results

Experiment 3: checking the processing time

60 scenario models defined with 2 to 10 physical object variables and 2 to 10 components. The algorithms are tested on simulated videos containing up to 240 persons in the scene.

The composed scenario recognition algorithm is able to process up to 240 persons in the scene.
Scenario Recognition: Temporal Constraints

Results

- A generic formalism to help experts model intuitively states, events and scenarios.

- Recognition algorithm processes temporal operators in an efficient way.
  - Linear search in the past.

- The recognition of complex scenarios (large number of actors) becomes real time.
  - Indexed Trees to structure and access the already recognized scenarios

- However,
  - uncertainty needs to be taken care
  - Scenario modeling is not always easy
Scenario recognition: Results

Bank agency monitoring in Paris (M. Maziere)
Scenario recognition: Results

Vandalism scenario example (temporal constraints):

\[\text{Scenario}(\text{vandalism\_against\_ticket\_machine},\]
\[\quad \text{Physical\_objects}((p:\text{Person}), (eq:\text{Equipment}, \text{Name}="\text{Ticket\_Machine}") )\]
\[\quad \text{Components} ((\text{event} \ s1: p \ \text{moves\_close\_to} \ eq)
\quad \quad \ (\text{state} \ s2: p \ \text{stays\_at} \ eq)
\quad \quad \ (\text{event} \ s3: p \ \text{moves\_away\_from} \ eq)
\quad \quad \ (\text{event} \ s4: p \ \text{moves\_close\_to} \ eq)
\quad \quad \ (\text{state} \ s5: p \ \text{stays\_at} \ eq))\]
\[\quad \text{Constraints} ((s1 \neq s4) (s2 \neq s5)
\quad \quad \ (s1 \ \text{before} \ s2) (s2 \ \text{before} \ s3)
\quad \quad \ (s3 \ \text{before} \ s4) (s4 \ \text{before} \ s5))\)\)
Scenario Recognition: Results

Vandalism in metro in Nuremberg

close_to
stay_at
vandalism
Scenario recognition: Results

Example: a “Vandalism against a ticket machine” scenario
Group Scenario Detection – Paris subway
Waiting example - Erratic group example

```
PrimitiveState(in.WaitingZone2,
    PhysicalObjects((g1 : Group), (z1 : Zone))
    Constraints ((g1->Position in z1->Vertices)
        (z1->Name = WaitingZone2)
    )
    Alarm ((Level : NOTURGENT))
)
```

```
CompositeState(Erratic_Group,
    PhysicalObjects((g1 : Group))
    Components((c1 : PrimitiveState Erratic(g1)))
    Constraints((duration(c1) >= MIN_TIME_ERRATIC) )
    Alarm((Level : URGENT))
)
```
Scenario recognition:
Results Example: “Unloading Front Operation” event

• Example of the **Unloading Front Operation** (global)

  CompositeEvent (UnLoading_Front_GlobalOperation,
  PhysicalObjects ( (v1 : Vehicle), (v2 : Vehicle),
   (z1 : Zone), (z2 : Zone), (z3 : Zone))
  Components ( (c1 : CompositeEvent Loader_Arrival(v1, z1, z2))
    (c2 : CompositeEvent Transporter_Arrival(v2, z1, z3))
  Constraints ( (v1->SubType = LOADER)
    (v2->SubType = TRANSPORTER)
    (z1->Name = ERA)
    (z2->Name = RF_DoorC_Access)
    (z3->Name = LOADER_BackZone)
    (c1 before c2)))
Scenario recognition: Results
Example: “Unloading Global Operation” event

• “Unloading Global Operation”
Example of the **Unloading Front Operation** (detailed)

**CompositeEvent** *(UnLoading_Front_Detailed_Operation*,

**PhysicalObjects** *( (p1 : Person), (v1 : Vehicle), (v2 : Vehicle), (v3 : Vehicle),

( z1 : Zone), (z2 : Zone), (z3 : Zone), (z4 : Zone))

**Components** *( (c1 : CompositeEvent Loader_Arrival(v1, z1, z2))

(c2 : CompositeEvent Transporter_Arrival(v2, z1, z3))

(c3 : CompositeState Worker_Manipulating_Container(p1, v3, v2, z3, z4)))

**Constraints** *( (v1->SubType = LOADER)

(v2->SubType = TRANSPORTER)

(z1->Name = ERA) (z2->Name = RF_DoorC_Access)

(z3->Name = LOADER_BackZone)

(z4->Name = Behind_RF_DoorC_Access)

(c1 before c2)

(c2 before c3))
Scenario recognition: Results
Parked aircraft monitoring in Toulouse (F Fusier)

• “Unloading Front Operation”

PHYSICAL OBJECTS:
VEHICLES: {Loader, Transporter}
PERSONS: {Worker}
STATIC ZONES: {ERA}
AIRCRAFT ZONES: {Front_Unloading_Area, Baggages_Unloading_Area}
DYNAMIC ZONES: {Transporter_Parking_Area}

VIDEO EVENTS:
Loader_Arrival
Transporter_Arrival
Worker_Arrived
Worker_Manipulating_Container
Scenario recognition: Results
Example: “Aircraft Arrival Preparation” event

• Aircraft Arrival Preparation (involving the GPU)
Scenario recognition: Results

Example: “Tow Tractor Arrival” event

• Tow Tractor Arrival
Scenario recognition: Results
Example: “vandalism_against_window” event

CompositeEvent( vandalism_against_window,

PhysicalObjects( (vandal : Person) ), (w : Equipment))

Components( (vandalism_against_window_VIDEO : 
CompositeEvent vandal_close_to_window(vandal, w))
(vandalism_against_window_AUDIO : 
CompositeEvent tag_detected_close_to_person(vandal)))

Constraints( (vandalism_against_window_VIDEO during 
vandalism_against_window_AUDIO) )

Alarm( AText("Vandalism against window")
ATYPE("URGENT") )
Scenario recognition: Results

Example: “Scratch & theft in a train” scenarios
Scenario recognition: Results
Example: a “Disturbing people in a train” scenario
1st experiment: Multi-sensor Scenario recognition

Example of “Taking meal” event model

CompositeEvent (M_TakingMeal,

PhysicalObjects ((p : Person), (z1 : Zone), (z2 : Zone), (eq1 : Equipment))

Components ((c1 : PrimitiveState V_InLivingroom(p, z1))
  (c2 : PrimitiveState V_CloseToTable(p, eq1))
  (c3 : CompositeState M_PersonSittingAtDinningtable(p, z2)))

Constraints ((z1’s Name = Livingroom),
  (z2’s Name = Dinningtable)
  (eq1’s Name = table),
  (c2 Duration >= threshold1),
  (c2 During c1),
  (c3 During c2),
  (c3 Duration >= threshold2))

Alert (“Person is taking a meal”, “NOTURGENT”)
Multi-sensor Scenario recognition

- Language combining multi-sensor information

Activity (Use Fridge,

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

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

(c2: Close_to (p, Fridge))

(c3: Bending (p))

(c4: Opening (Fridge))

(c5: Closing (Fridge)) )

Constraints ((c1 before c2 )

(c3 during c2 )

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

Detected by video camera

Detected by contact sensor
Recognition of the “Prepare meal” event

- The person is recognized with the posture "standing with one arm up", “located in the kitchen” and “using the microwave”.

Visualization of a recognized event in the Gerhome laboratory
Recognition of the “Resting in living-room” event

- The person is recognized with the posture “sitting in the armchair” and “located in the living-room”.

Visualization of a recognized event in the Gerhome laboratory
Event recognition results

- **14 elderly volunteers** have been monitored during 4 hours (total: more than 56 hours).
- Recognition of the “Prepare meal” event for a 65 old man
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 $5 \times 4 = 20$ hours

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

- 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)

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

- Errors occur at the border between living-room and kitchen
- Mixed postures such as bending and sitting due to segmentation errors
Table 2: Monitored activities, their frequency (n1 & n2), mean duration (m1 & m2) and total duration for 2 volunteers staying in the GERHOME laboratory for 4 hours; NDA=Normalized Difference of mean durations of Activities=|m1-m2|/ (m1+m2); NDI=Normalized Difference of Instances number=|n1-n2|/(n1+n2); possible differences in behavior of the 2 volunteers are signified in bold
## Recognition of a set of activities comparing two elderly people

<table>
<thead>
<tr>
<th>Activity</th>
<th>Used sensor (s)</th>
<th>Elderly people 1 (64 years)</th>
<th>Elderly people 2 (85 years)</th>
<th>Normalized Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Activity duration (min:sec)</td>
<td>Nb inst (n1)</td>
<td>Activity duration (min:sec)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (m1)</td>
<td>Total</td>
<td>Mean (m2)</td>
</tr>
<tr>
<td>Use fridge</td>
<td>Video + contact</td>
<td>0:12</td>
<td>2:50</td>
<td>14</td>
</tr>
<tr>
<td>Use stove</td>
<td>Video + power</td>
<td>0:08</td>
<td>4:52</td>
<td>35</td>
</tr>
<tr>
<td>Use upper-cupboard</td>
<td>Video + contact</td>
<td>0:51</td>
<td>21:34</td>
<td>25</td>
</tr>
<tr>
<td>Sitting on chair</td>
<td>Video + pressure</td>
<td>6:07</td>
<td>73:27</td>
<td>12</td>
</tr>
<tr>
<td>Entering the living-room</td>
<td>Video</td>
<td>1:25</td>
<td>25:00</td>
<td>20</td>
</tr>
<tr>
<td>Standing</td>
<td>Video</td>
<td>0:09</td>
<td>30:00</td>
<td>200</td>
</tr>
<tr>
<td>Bending</td>
<td>Video</td>
<td>0:04</td>
<td>2:00</td>
<td>30</td>
</tr>
</tbody>
</table>

**Table 2:** Monitored activities, their frequency (n1 & n2), mean duration (m1 & m2) and total duration for 2 volunteers staying in the GERHOME laboratory for 4 hours; NDA=Normalized Difference of mean durations of Activities=|m1-m2|/(m1+m2); NDI=Normalized Difference of Instances number=|n1-n2|/(n1+n2); possible differences in behavior of the 2 volunteers are signified in bold
### Recognition of a set of activities comparing two elderly people

<table>
<thead>
<tr>
<th>Activity</th>
<th>Used sensor (s)</th>
<th>Elderly people 1 (64 years)</th>
<th>Elderly people 2 (85 years)</th>
<th>Normalized Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Activity duration (min:sec)</td>
<td>Nb inst (n1)</td>
<td>Activity duration (min:sec)</td>
<td>Nb inst (n2)</td>
</tr>
<tr>
<td></td>
<td>Mean (m1)</td>
<td>Total</td>
<td>Mean (m2)</td>
<td>Total</td>
</tr>
<tr>
<td>Use fridge</td>
<td>Video + contact</td>
<td>0:12</td>
<td>2:50</td>
<td>14</td>
</tr>
<tr>
<td>Use stove</td>
<td>Video + power</td>
<td>0:08</td>
<td>4:52</td>
<td>35</td>
</tr>
<tr>
<td>Use upper-cupboard</td>
<td>Video + contact</td>
<td>0:51</td>
<td>21:34</td>
<td>25</td>
</tr>
<tr>
<td>Sitting on chair</td>
<td>Video + pressure</td>
<td>6:07</td>
<td>73:27</td>
<td>12</td>
</tr>
<tr>
<td>Entering the living-room</td>
<td>Video</td>
<td>1:25</td>
<td>25:00</td>
<td>20</td>
</tr>
<tr>
<td>Standing</td>
<td>Video</td>
<td>0:09</td>
<td>30:00</td>
<td>200</td>
</tr>
<tr>
<td>Bending</td>
<td>Video</td>
<td>0:04</td>
<td>2:00</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 2: Monitored activities, their frequency (n1 & n2), mean duration (m1 & m2) and total duration for 2 volunteers staying in the GERHOME laboratory for 4 hours; NDA=Normalized Difference of mean durations of Activities=|m1-m2|/ (m1+m2); NDI=Normalized Difference of Instances number=|n1-n2|/(n1+n2); possible differences in behavior of the 2 volunteers are signified in bold.
Evaluation and results

Duration of 6 activities for 9 observed elderly people

Duration (time unit is hh:mm:ss)

Person: P1 P2 P3 P4 P5 P6 P7 P8 P9
2nd experiment : CMRR in Nice Hospital Screening of AD patients

- **Medical staff & healthy younger**
  - 22 people (more female than male)
  - Age: ~ 25-35 years
  - Medical staff
  - 1 video camera, Actiwach

- **Older persons**
  - 20 (woman & man)
  - Age: ~ 60-85 years
  - 2 video cameras
  - Actiwach/ motionPod

- **Alzheimer patients:**
  - 21 AD people (woman & man)
  - 19 MCI (mild cognitive impairment) and mixed
  - Age: ~ 60-85 years
  - 2 video cameras
  - Actiwach/ motionPod
Reconnaissance de l’activité “stop and go” et «sit down» en utilisant le capteur vidéo au CM2R.

Reconnaissance d’un protocole au CM2R - CoBTek
Centre Mémoire de Ressources et de Recherche du CHU Nice
Activity monitoring in Nice Hospital with AD patients

Recognition of the “stand-up” activity.
Activity monitoring in Nice Hospital with AD patients

Recognition of the “stand-up & walking” activity.
Learning Scenario Models: scene model (G. Pusiol)

**Localization** of the person during 4 observation hours

**Stationery** positions of the person

Walked distance = 3.71 km
Learning Scenario Models: scene model

Topologies are important because is where the reasoning is

The Scene Model = 3 Topologies: Multi-Resolution.
Primitive Event: global object motion between 2 zones.

Advantage:
The topology regions and primitive events are semantically understandable.
1. Initialize sparse KLT points
2. Track the points during the whole PFC - pyramidal KLT - [Bouguet 2000]
3. Filter with the global tracker
4. Re-initialize for a new PFC (means short errors)
Learning Scenario Models: Local tracklets

**Goal:** Get semantic describable **main motion** of the body parts from perceptual information. (*i.e. complement the global spatial description*)

\[ \text{Local Dynamics}_{PE} = \text{Clustering} \text{ (Mean Shift) the Pixel tracklets} \]

**Trick:** Adapt the bandwidth dynamically to the walked distance.

\[ h = \| PFC_{\text{Departure}} \cdot \mu - PFC_{\text{Arrival}} \cdot \mu \| \ast C \]

*b* WE GET and USE THE GREEN LINES
Learning Scenario Models: Local tracklets

EXAMPLE

SURF & SIFT: slower to compute
Primitive Events Results:

Each PE is colored by its type

SIMILAR COLOR IS SIMILAR ACTIVITY

EATING

COOKING

Video Frame

Person A

Eating

Cooking

Person D
**Activity Discovery**: Find start/end of interesting activity and classify them

**Input: Sequences of PE**

**RESOLUTIONS**

- Easy to understand
- Non parametric and Deterministic
- The basic patterns can describe complex ones

**Stay**

\[ Stay_{A\rightarrow A} = (A \rightarrow A)^+ \]

**Change**

\[ Change_{A\rightarrow B} = (A \rightarrow B), A \neq B \]

**DA = Discovered Activity**

(color = DA type)
Activity Discovery

Discovery Results:

Similar color is similar Discovered Activity

Multi-resolution sequence of discovered activities

Level 11
Level 6
Level 4

a b c d f g h j k m r q n o p

4 hours

Sitting in the armchair
Prep. Meal
In the bathroom

Eating

a
b
c
f
h
k
j

106
Activity Models *Histograms of Multi-resolutions (HM)*

Is a set of 3 histograms. Each histogram has 2 dimensions. Containing global and local descriptions of the DAs.

\[
H_i(type_{PE}, \theta) = \sum LDD_N(\theta) : \{\forall N | type_N = type_{PE}\}
\]
Activity Models: Hierarchical Activity Models (HAM)

Building Nodes

A node is composed of two elements

1. Attributes
2. Sub-attributes

The Hierarchical Activity Model node structure of A1 and A2

A node N is a set of discovered activities {DA1, DA2..., DAn} where all DAs are at the same resolution level and are of the same type (color)

Input Training Neighborhoods of a target activity

Tree of Nodes

“Coding at Chair”
Results

5 targeted activities to be recognized
“Sitting in the armchair”
“Cooking”
“Eating at position A”
“Sitting at Position B”
“Going from the kitchen to the bathroom”.

4 Test Persons

“Cooking”

“Eating at Position A”

“Sitting in the Armchair”

Scene logical Regions
Evaluation

Results: RGB-D **Multiples Persons**

(i) Sitting.far (red)  (a) Sitting.near  (a) Preparing Meal, Interaction.TrashCan

<table>
<thead>
<tr>
<th>Tracked object:</th>
<th>silhouette center</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
</tr>
<tr>
<td>Preparing.Meal</td>
<td>10</td>
</tr>
<tr>
<td>Interacting.Trashcan</td>
<td>2</td>
</tr>
<tr>
<td>Sitting.near</td>
<td>3</td>
</tr>
<tr>
<td>Sitting.Far</td>
<td>3</td>
</tr>
</tbody>
</table>
Video Understanding: Knowledge Discovery (E. Corvee, JL. Patino_Vilchis)

• CARETAKER/VANAHEIM: European projects to provide an efficient tool for the management of large multimedia collections.

Processing units

On-line

Multiple Audio/Video sensors

Acquisition
- Audio
- Video

Detection
- Object detection
- Object tracking
- Event detection

Knowledge Modelling

Knowledge Discovery
- Trajectory characterization (Agglomerative clustering)
- Object statistics
- Activity discovery (Relational analysis)

Off-line

Data storage

Raw Data

Simple Events

Complex Events
Online learning of zones: Point Clustering

\[ Z_{\text{cl}}(x, y) = \exp\left(-\|x - x_{Li}\|^2 T^2\right) \exp\left(-\|y - y_{Li}\|^2 T^2\right) \]
Trajectory feature weight optimization: Results

Before: two close clusters

After: merge of the clusters

Before: a too large cluster

After: split of the cluster
Results: Trajectory Clustering

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of objects</th>
<th>Object types</th>
<th>Start time (min)</th>
<th>Duration (sec)</th>
<th>Trajectory types</th>
<th>Significant event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 38</td>
<td>385</td>
<td>{'Crowd'}</td>
<td>[0.1533, 48.4633]</td>
<td>[0.04, 128.24]</td>
<td>{'4' '3' '7'}</td>
<td>{'void '}</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>15</td>
<td>{'Person'}</td>
<td>[28.09, 46.79]</td>
<td>[2.04, 75.24]</td>
<td>{'13' '12' '19'}</td>
<td>{'inside_zone_Platform '}</td>
</tr>
</tbody>
</table>
Trajectory Clustering: rare events in Roma subway
Online learning of zones

Learned zones are stable after processing long term data
Online learning of events through time

<table>
<thead>
<tr>
<th>Rank</th>
<th>(%)</th>
<th>Event</th>
<th>Rank</th>
<th>(%)</th>
<th>Event</th>
<th>Rank</th>
<th>(%)</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31.46</td>
<td>at zone Turnstiles</td>
<td>1</td>
<td>29.74</td>
<td>at zone Turnstiles</td>
<td>1</td>
<td>28.33</td>
<td>at zone Turnstiles</td>
</tr>
<tr>
<td>2</td>
<td>9.79</td>
<td>at zone Entrance2</td>
<td>2</td>
<td>9.86</td>
<td>at zone Entrance2</td>
<td>2</td>
<td>10.08</td>
<td>at zone Entrance2</td>
</tr>
<tr>
<td>3</td>
<td>7.86</td>
<td>zone Entrance2 to zone Turnstiles</td>
<td>3</td>
<td>8.61</td>
<td>zone Entrance2 to zone Turnstiles</td>
<td>3</td>
<td>7.85</td>
<td>zone Entrance2 to zone Turnstiles</td>
</tr>
<tr>
<td>4</td>
<td>4.89</td>
<td>zone Turnstiles to zone Entrance2</td>
<td>4</td>
<td>4.64</td>
<td>zone Turnstiles to zone Entrance2</td>
<td>4</td>
<td>5.47</td>
<td>zone Turnstiles to zone Entrance2</td>
</tr>
<tr>
<td>5</td>
<td>4.83</td>
<td>zone Turnstiles to zone Centre hall</td>
<td>5</td>
<td>5.04</td>
<td>at zone Centre hall</td>
<td>5</td>
<td>4.55</td>
<td>zone Entrance1 to zone Turnstiles</td>
</tr>
<tr>
<td>6</td>
<td>3.72</td>
<td>zone Centre hall to zone Turnstiles</td>
<td>6</td>
<td>3.91</td>
<td>zone Entrance1 to zone Turnstiles</td>
<td>6</td>
<td>3.84</td>
<td>zone Centre hall to zone Turnstiles</td>
</tr>
<tr>
<td>7</td>
<td>3.45</td>
<td>at zone Centre hall</td>
<td>7</td>
<td>4.15</td>
<td>zone Turnstiles to zone Centre hall</td>
<td>7</td>
<td>4.69</td>
<td>at zone Centre hall</td>
</tr>
<tr>
<td>8</td>
<td>3.31</td>
<td>zone Entrance1 to zone Turnstiles</td>
<td>8</td>
<td>3.75</td>
<td>zone Centre hall to zone Turnstiles</td>
<td>8</td>
<td>3.77</td>
<td>zone Turnstiles to zone Centre hall</td>
</tr>
<tr>
<td>9</td>
<td>2.48</td>
<td>zone Turnstiles to zone Entrance1</td>
<td>9</td>
<td>2.60</td>
<td>zone Turnstiles to zone Entrance1</td>
<td>9</td>
<td>2.41</td>
<td>zone Turnstiles to zone Entrance1</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td>11</td>
<td></td>
<td></td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1.79</td>
<td>at zone Vending machine1</td>
<td>12</td>
<td>1.57</td>
<td>zone Vending machine1 to zone Turnstiles</td>
<td>12</td>
<td>1.65</td>
<td>at zone Vending machine1</td>
</tr>
<tr>
<td>13</td>
<td>1.51</td>
<td>at zone Vending machine2</td>
<td>13</td>
<td>1.31</td>
<td>zone Vending machine1 to zone Centre hall</td>
<td>13</td>
<td>1.55</td>
<td>zone Vending machine1 to zone Turnstiles</td>
</tr>
<tr>
<td>14</td>
<td>1.51</td>
<td>zone Vending machine1 to zone Centre hall</td>
<td>14</td>
<td>1.43</td>
<td>at zone Vending machine1</td>
<td>14</td>
<td>1.51</td>
<td>zone Entrance2 to zone Centre hall</td>
</tr>
</tbody>
</table>

Four simple events are the most frequently occurring; The frequency of occurrence of other events changes slightly.
Online learning of activities through time

<table>
<thead>
<tr>
<th>set001</th>
<th>set002</th>
<th>set008</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.13</td>
<td>at zone Turnstiles</td>
</tr>
<tr>
<td>2</td>
<td>11.21</td>
<td>at zone Entrance2</td>
</tr>
<tr>
<td>3</td>
<td>5.98</td>
<td>zone Entrance2 to zone Turnstiles</td>
</tr>
<tr>
<td>4</td>
<td>4.13</td>
<td>at zone Turnstiles; at zone Turnstiles</td>
</tr>
<tr>
<td>5</td>
<td>3.16</td>
<td>zone Turnstiles to zone Entrance2</td>
</tr>
<tr>
<td>6</td>
<td>2.61</td>
<td>zone Entrance1 to zone Turnstiles</td>
</tr>
<tr>
<td>7</td>
<td>2.29</td>
<td>at zone Entrance1</td>
</tr>
<tr>
<td>8</td>
<td>2.18</td>
<td>zone Turnstiles to zone Centre hall</td>
</tr>
<tr>
<td>9</td>
<td>1.74</td>
<td>zone Centre hall to zone Turnstiles</td>
</tr>
<tr>
<td>10</td>
<td>1.52</td>
<td>zone Turnstiles to zone Entrance1</td>
</tr>
<tr>
<td>11</td>
<td>1.41</td>
<td>zone Turnstiles to zone Entrance1; at zone Turnstiles</td>
</tr>
<tr>
<td>12</td>
<td>1.31</td>
<td>at zone Centre hall</td>
</tr>
<tr>
<td>13</td>
<td>1.31</td>
<td>zone Vending machine1 to zone Turnstiles</td>
</tr>
<tr>
<td>14</td>
<td>1.20</td>
<td>at zone Turnstiles; zone Turnstiles to zone Entrance2</td>
</tr>
</tbody>
</table>

The most frequently occurring activities correspond to three simple events.
Online learning: Most common activities

at Turnstiles

zone South Entry to zone Turnstiles

at zone South Entry

zone North Entry to zone Turnstiles
Online learning: Most common activities

- Zone Turnstiles to Zone South Entry
- Zone Vending machine1 to Zone Turnstiles
- Zone Vending machine2 to Turnstiles
Online learning: Rare activities

Loitering (from tracking error)
at zone Turnstiles; at zone Turnstiles; at zone Turnstiles

Loitering: people talking then entering the station
at zone Centre hall; at zone Centre hall; at zone Centre hall; at zone Turnstiles

VM not working (from tracking error)
at zone Centre hall; zone Centre hall to zone Vending machine2; zone Vending machine2 to zone Vending machine1

VM not working
zone South Entry to zone Centre hall; zone Centre hall to zone Vending machine2; zone Vending machine2 to zone Vending machine1; zone Vending machine1 to zone Turnstile
Online learning: Rare activities

going through the station
zone North Entry to zone South Entry

tickets do not work; need new tickets
zone Turnstiles to zone Centre hall; zone Centre hall to zone Vending machine2; zone Vending machine2 to zone Centre hall; zone Centre hall to zone Turnstiles
Online learning : Loitering activities

Look for object O
where

\[ O.\text{Zone}.\text{avg}\_\text{speed} < (M.\text{global}\_\text{avg}\_\text{speed} - M.\text{global}\_\text{stddev}\_\text{speed}/2) \]
and
\[ O.\text{path}\_\text{length} > (M.\text{avg}\_\text{path}\_\text{length} + M.\text{stddev}\_\text{path}\_\text{length}) \]
and
\[ O.\text{walked}\_\text{zones}\_\text{nb} > 8 \]

M : average on objects tracked on 8 hours

<table>
<thead>
<tr>
<th>Loitering</th>
<th>GT #</th>
<th>TP #</th>
<th>FP #</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011-01-29T18_00_01-1262318</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-01-28T20_00_01-1763039</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>0.45</td>
<td>1</td>
</tr>
<tr>
<td>2011-01-28T20_00_01-1467943</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-01-30T20_00_00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011-02-01T20_00_00</td>
<td>0</td>
<td>0</td>
<td>~1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Online learning: **Loitering** activities
Global approach integrating all video understanding functionalities while focusing on the easy generation of dedicated systems based on

- cognitive vision: 4D analysis (3D + temporal analysis)
- artificial intelligence: explicit knowledge (scenario, context, 3D environment)
- software engineering: reusable & adaptable platform (control, library of dedicated algorithms)

⇒ Extract and structure knowledge (invariants & models) for
  - Perception for video understanding (perceptual, visual world)
  - Maintenance of the 3D coherency throughout time (physical world of 3D spatio-temporal objects)
  - Event recognition (semantics world)
  - Evaluation, control and learning (systems world)
Conclusion

A **global framework** for building video understanding systems:

- **Hypotheses:**
  - mostly fixed cameras
  - 3D model of the empty scene
  - predefined behavior models

- **Results:**
  - Video understanding real-time **systems** for Individuals, Groups of People, Vehicles, Crowd, or Animals …

- **Perspectives:**
  - Finer **human shape** description: *gesture models, face detection*
  - Design of **learning** techniques to complement a priori knowledge:
    - visual concept learning
    - scenario model learning
  - Scaling issue: managing large network of **heterogeneous sensors** (cameras, PTZ, microphones, optical cells, radars….)
Event detection examples
Trajectory Clustering: two day analysis

Friday

Saturday

no queuing
Contextual Object Analysis

With an increase of people, there is an increase on the use of vending machines.
3.1 Results and Evaluation: 13 last learned zones (several days)

<table>
<thead>
<tr>
<th>13 Matching Zones</th>
<th>Missing Zones</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 GT Zns</td>
<td>1 GT Zn (Poster2 Zn)</td>
</tr>
</tbody>
</table>

*Matching between zones calculated from their intersection*
Video Understanding

Scene Models (3D)
- Scene objects
- zones
- calibration matrices

Mobile objects

Individual Tracking

Group Tracking

Crowd Tracking

Behavior Recognition
- States
- Events
- Scenarios

Tools:
- Evaluation
- Acquisition
- Learning, …