Scene Understanding

perception, multi-sensor fusion, spatio-temporal reasoning and activity recognition.

Francois BREMOND

Orion project-team,
INRIA Sophia Antipolis, FRANCE

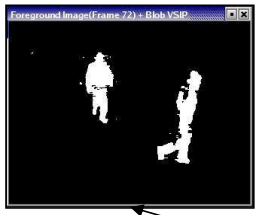
Francois.Bremond@sophia.inria.fr
http://www-sop.inria.fr/orion/orion-eng.html

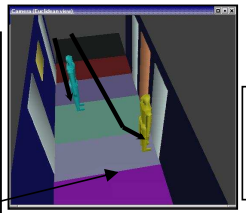
Key words: Artificial intelligence, knowledge-based systems, cognitive vision, human behavior representation, scenario recognition

Video Understanding

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

Segmentation → Classification → Tracking → Scenario Recognition

Alarms

access to forbidden area

3D scene model
Scenario models

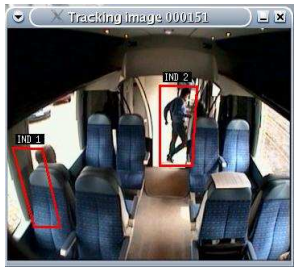
A priori Knowledge

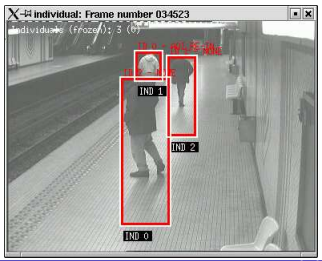
INRIA SOPHIA ANTIPOLIS

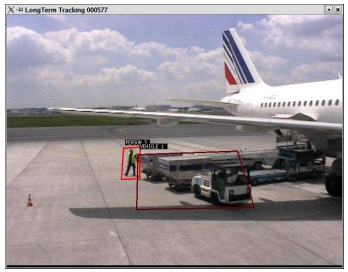
Video Understanding Applications

- 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 (simulation)
- 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 (15 persons): http://www.keeneo.com/







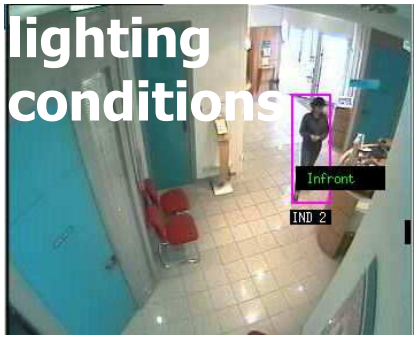
Video Understanding: Domains

- Smart Sensors: Acquisition (dedicated hardware), thermal, omni-directional, PTZ, cmos, IP, tri CCD, FPGA.
- Networking: UDP, scalable compression, secure transmission, indexing and storage.
- Computer Vision: 2D object detection (Wei Yun I2R Singapore), active vision, tracking of people using 3D geometric approaches (T. Ellis Kingston University UK)
- Multi-Sensor Information Fusion: cameras (overlapping, distant) + microphones, contact sensors, physiological sensors, optical cells, RFID (GL Foresti Udine Univ I)
- Event Recognition: Probabilistic approaches HMM, DBN (A Bobick Georgia Tech USA, H Buxton Univ Sussex UK), logics, symbolic constraint networks
- Reusable Systems: Real-time distributed dependable platform for video surveillance (Multitel, Be), OSGI, adaptable systems, Machine learning
- Visualization: 3D animation, ergonomic, video abstraction, annotation, simulation, HCI, interactive surface.

Video Understanding: Issues

Practical issues

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



Video Understanding: 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, models and context awareness
  - Knowledge acquisition versus ((none, semi)-supervised, incremental) learning techniques
  - Formalization, modeling, ontology, standardization

Video Understanding: Approach

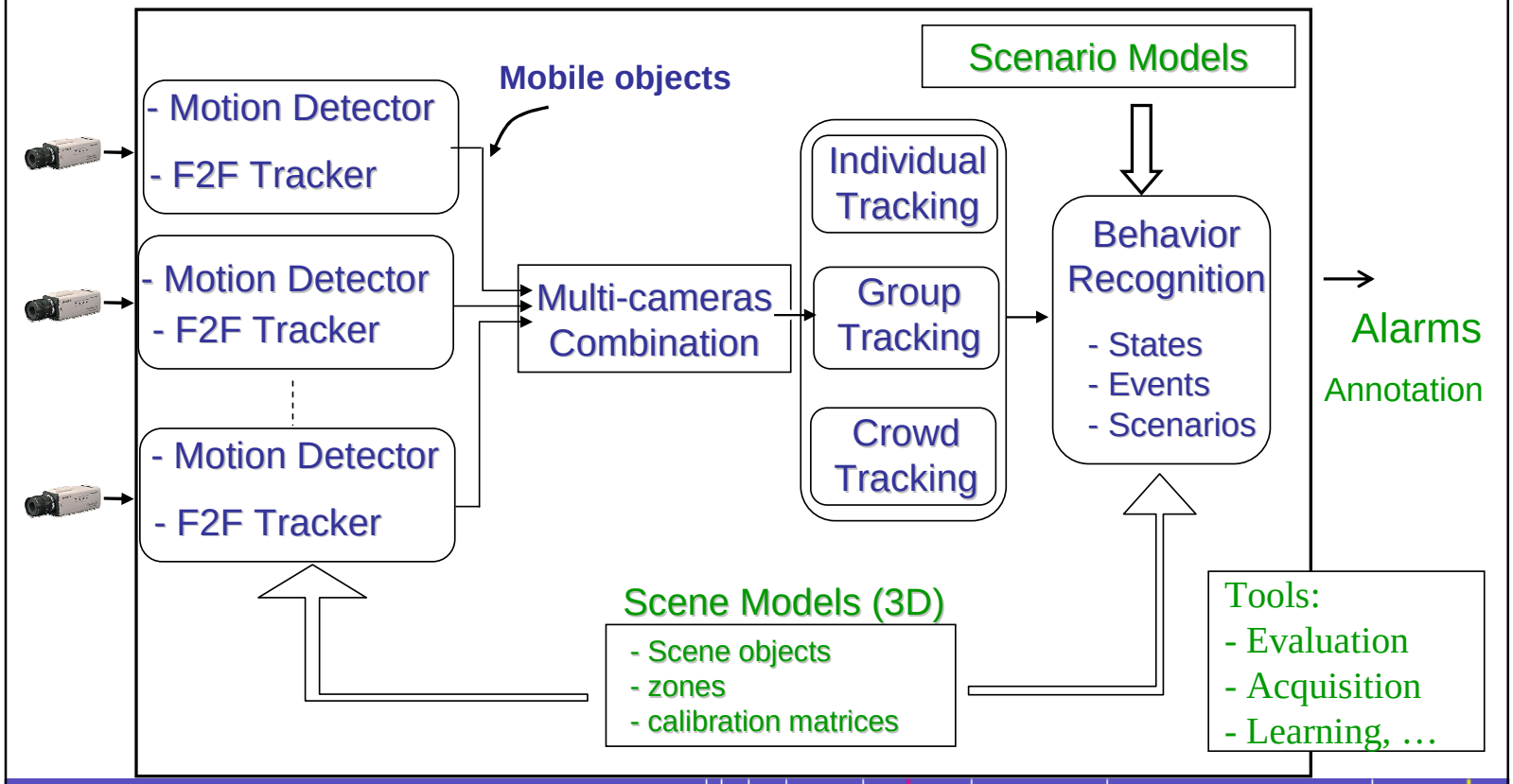
Global approach integrating all video understanding functionalities,
while focusing on the easy generation of dedicated systems based on

- 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)

Video Understanding: platform



Outline (1/2)

- Introduction on Video Understanding
- Knowledge Representation [WSCG02]
- Perception
  - People detection [IDSS03a]
  - Posture recognition [VSPETS03], [PRLetter06]
  - Coherent Motion Regions
- 4D coherency
  - People tracking [IDSS03b], [CVDP02]
  - Multi cameras combination [ACV02], [ICDP06a]
  - People lateral shape recognition [AVSS05a]
- Event representation [KES02], [ECAI02]

Outline (2/2)

- Event recognition:
  - State of the art
  - finite state automata [ICNSC04]
  - Bayesian network [ICVS03b]
  - CSP
  - Temporal constraints [AVSS05b], [IJCAI03], [ICVS03a], [PhDTV04], [ICDP06]
- Autonomous systems:
  - performance evaluation [VSPETS05], [PETS05], [IDSS04], [ICVIIP03], [WMVC07], [AVSS07]
  - program supervision [ICVS06c], [ICVIIP04], [MVA06a]
  - parameter learning [PhDBG06]
  - knowledge discovery [ICDP06], [VIE07]
  - learning scenario models [ICVS06a], [ICDP06b]
- Results and demonstrations: metro, bank, train, airport, trichogramma monitoring, Homecare [ICVS06b], [AJCAI06], [ICVW06], [ITSC05], [BR06], [MVA06b], [SETIT07]

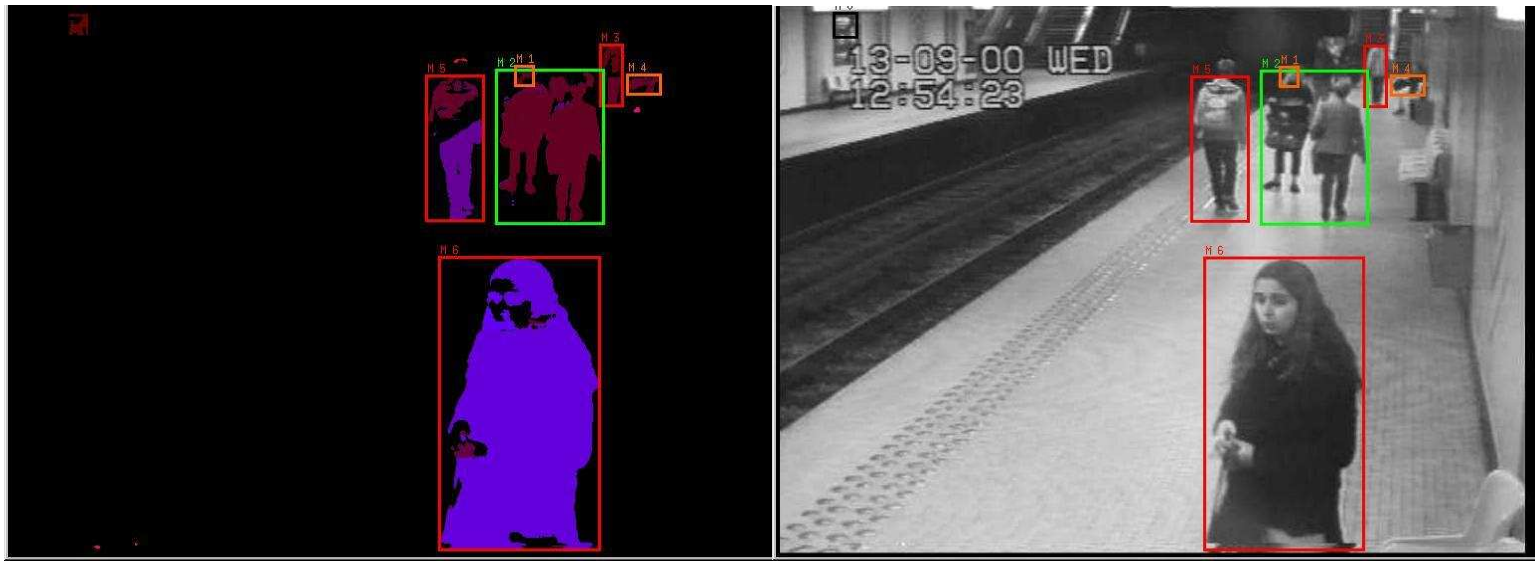
People detection

- 4 levels of people detection
  - 3D ratio height/width
  - 3D parallelepiped
  - 3D articulate human model
  - Coherent 2D motion regions

People detection

Classification into more than 8 classes (e.g. Person, Groupe, Train)
based on 2D and 3D descriptors (position, **3D ratio height/width**, …)

Example of 4 classes: Person, Group, Noise, Unknown



INRIA SOPHIA ANTIPOLIS

People detection (M. Zuniga)

Classification into 3 people classes : 1Person, 2Persons, 3Persons, Unknown, …, based on **3D parallelepiped**

Posture Recognition (B. Boulay)

- Recognition of human body postures :
  - with only one static camera
  - in real time

- Existing approaches can be classified :
  - 2D approaches : depend on camera view point
  - 3D approaches : markers or time expensive

- Approach: combining
  - 2D techniques (eg. Horizontal & Vertical projections of moving pixels)
  - **3D articulate human** model (10 joints and 20 body parts)

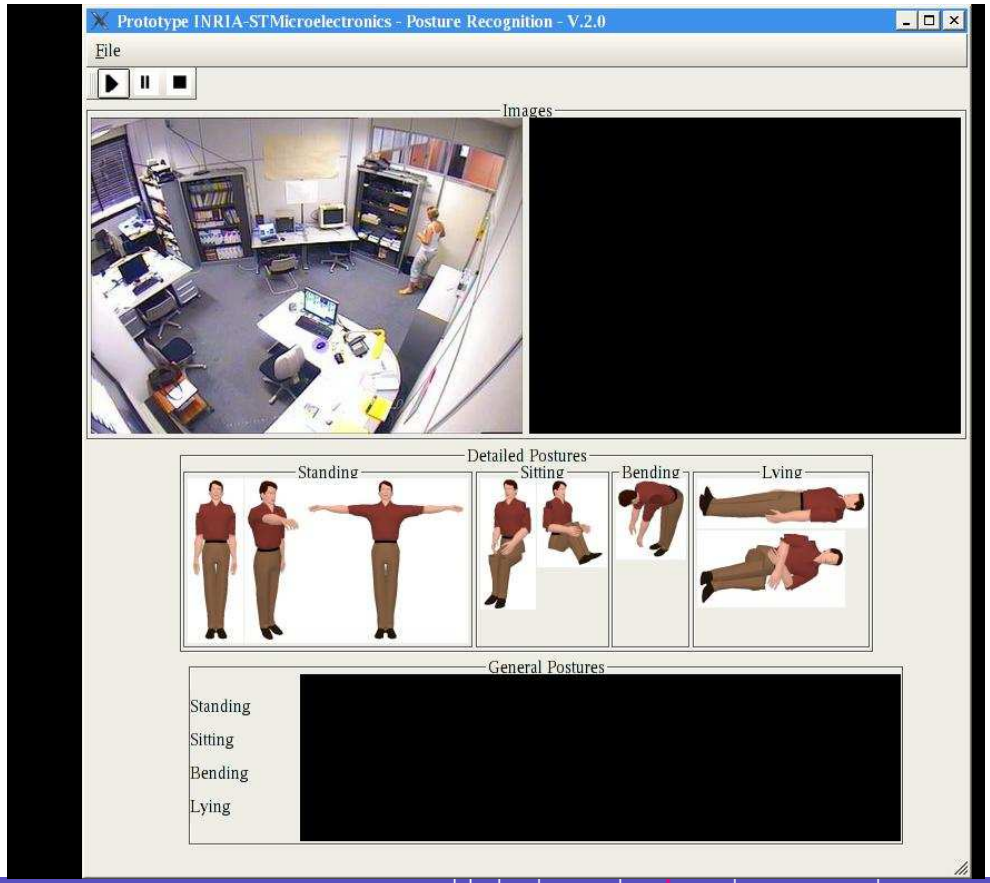
Posture Recognition : silhouette comparison

Real world

Virtual world

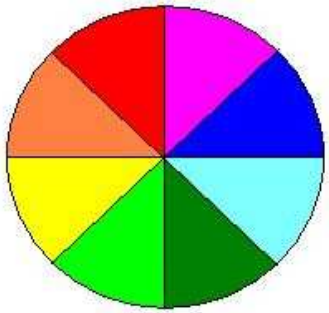
Generated silhouettes

Posture Recognition : results



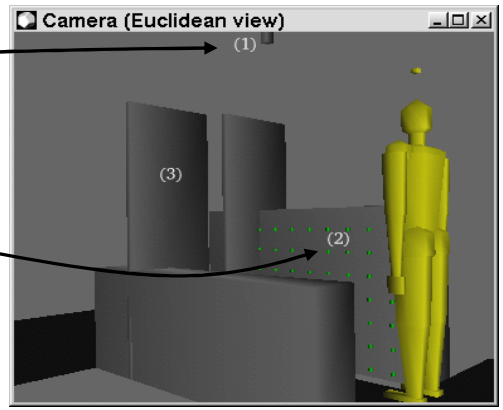
Coherent Motion Regions (MB. Kaaniche)

Approach: Track and Cluster KLT (Kanade-Lucas-Tomasi) feature points.



Multi sensors information fusion: Lateral Shape Recognition (B. Bui)

- Objective: access control in subway, bank,…
- Approach: real-time recognition of lateral shapes such as “adult”, “child”, “suitcase”
  - based on naive Bayesian classifiers
  - combining video and multi-sensor system (leds, optical cells).



A fixed camera at the height of 2.5m observes the mobile objects from the top.

Lateral sensors (leds, 5 cameras, optical cells) on the side.

Lateral Shape Recognition: Mobile Object Model

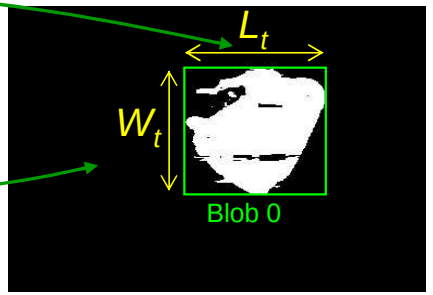
- Shape Model composed of 13 characteristics

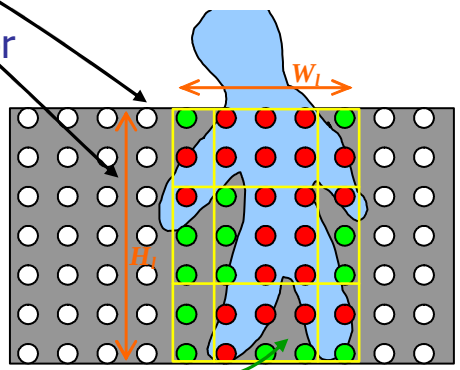
  ✓ 3D length L_t and 3D width W_t

  ✓ 3D width W_l and the 3D height H_l of the occluded zone.

  ✓ We divide the occluded zone into 9 sub-zones and for each sub-zone i, we use the density S_i (i=1..9) of the occluded sensors.

- Model of a mobile object = (L_t, W_t, W_l, H_l, S_1,..., S_9) combine with a Bayesian formalism.





Lateral Shape Recognition: Experimental Results

- Recognition of “adult with child”

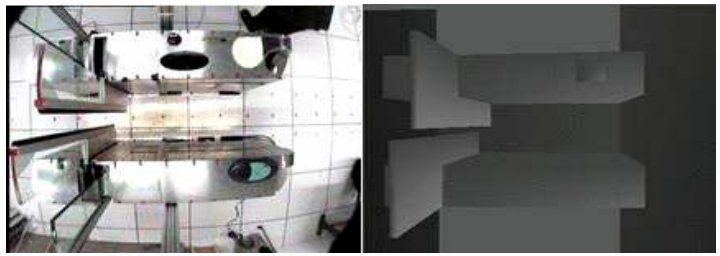
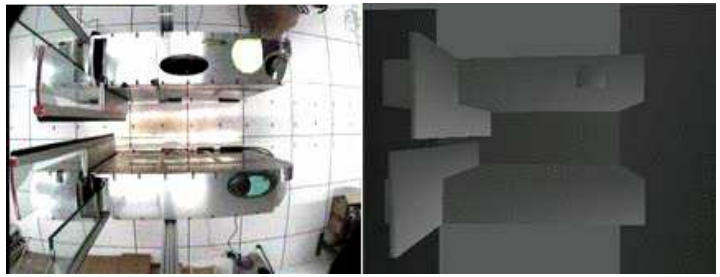


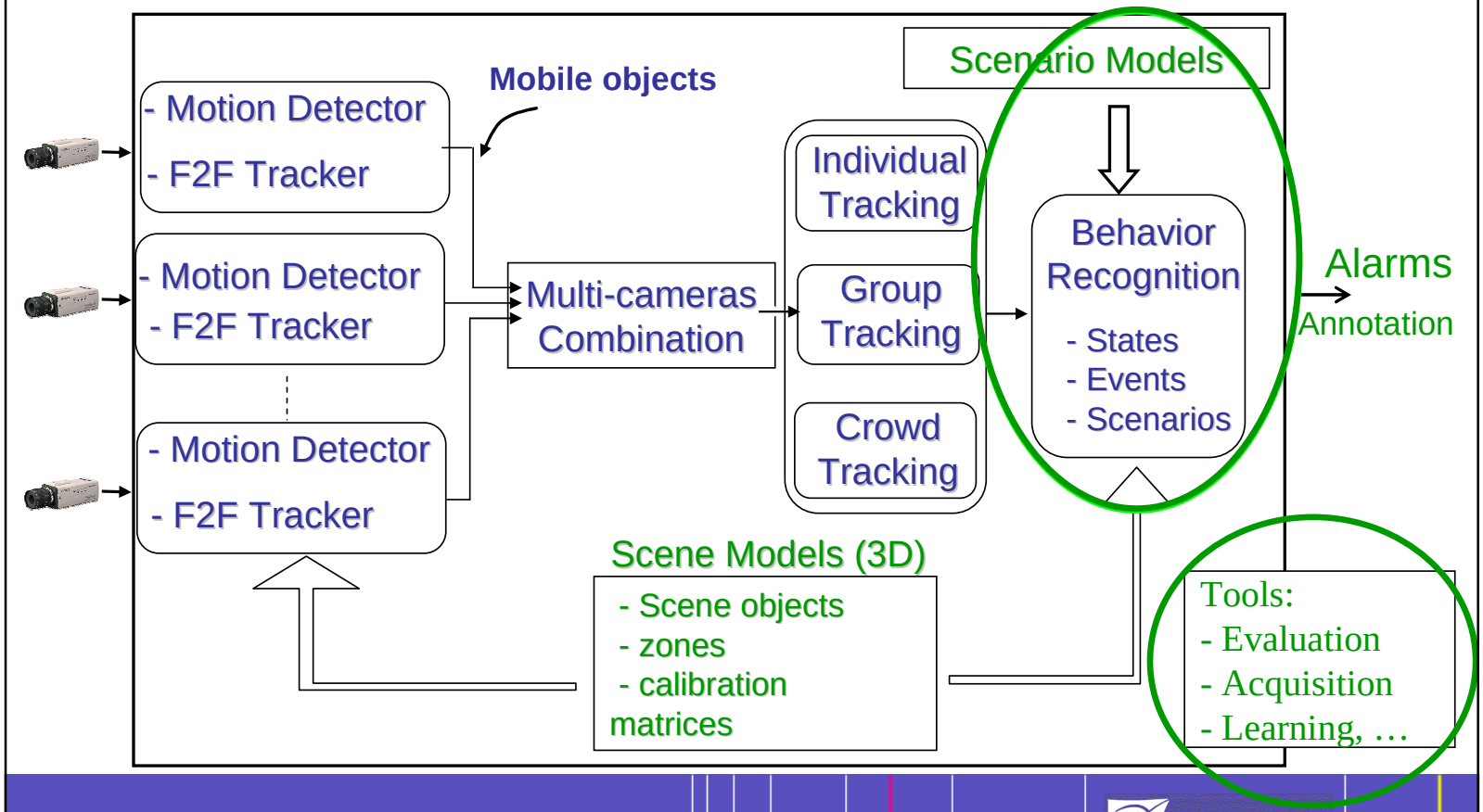
Image from the top camera

3D synthetic view of the scene

- Recognition of “two overlapping adults”



Video Understanding



Event Representation

Video events: real world notion corresponding to short actions 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 Representation

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

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!!!”))

Uncertainty Representation

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)))

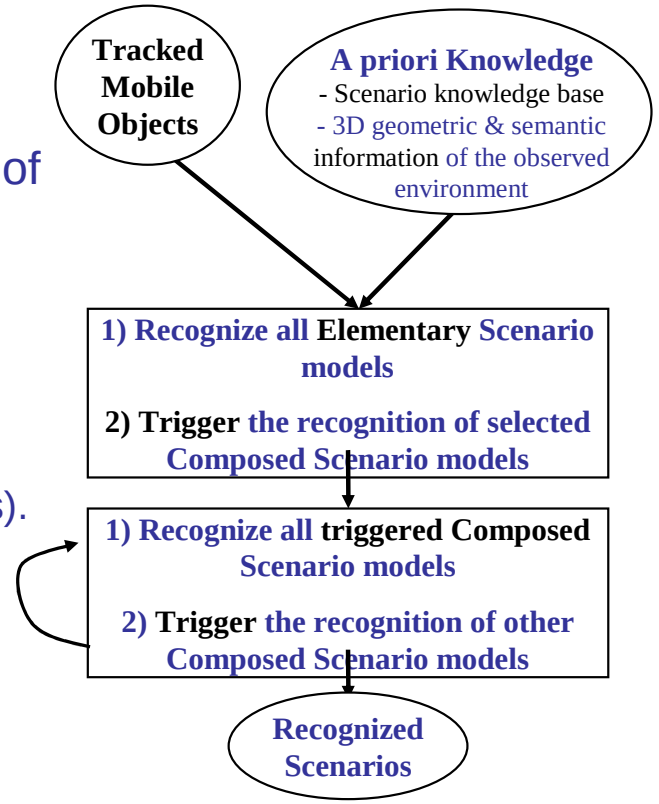
CompositeEvent (**Crowd_Splits**,
Physical Objects ((c1: Crowd, 0.5), (c2 : Crowd, 0.5), (z1: Zone))
Components ((s1 : CompositeState Move_toward (c1, z1), 0.3)
(e2 : CompositeEvent Move_away (c2, c1), 0.7))
Constraints ((e2 during s1)
(c2's Size > Threshold)
(recognized if likelihood > 0.8)))

Scenario Recognition: 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.



Scenario Recognition: Elementary Scenario

- The recognition of a compiled 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 $m_c = (m_1\ \textit{before}\ m_2\ \textit{before}\ m_3)$;

  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 of m_1 and of m_2 with i_3

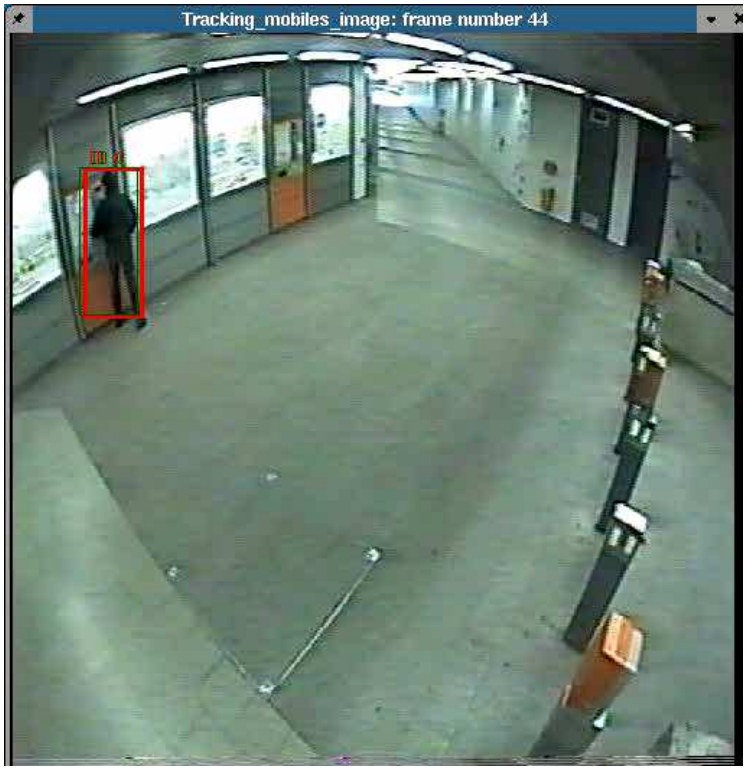
  ➔ a combinatorial explosion.

- **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: $m_c = (m_4\ \textit{before}\ m_3)$

  ➔ a linear search.

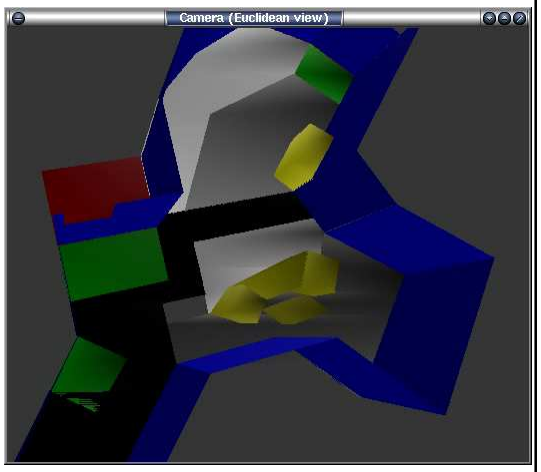
Scenario Recognition: Results
Vandalism in metro in Nuremberg



INRIA SOPHIA ANTIPOLIS

Scenario recognition: Results

Bank agency monitoring : Paris (M. Maziere)



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

- **"Unloading Front Operation"**

SCENARIO UNLOADING_DETAILED_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 HealthCare Monitoring (N. Zouba)

Approach :

- Multi-sensor analysis of **elderly activities**
- Detect in real-time any alarming situation
- Identify a person profile from the global trends of life parameters

Examples:

- Use_foodcupboard
- Use_microwave

Video Understanding:
Performance Evaluation (V. Valentin, R. Ma)

- **ETISEO**: French initiative for algorithm validation and knowledge acquisition: http://www-sop.inria.fr/orion/ETISEO/

- **Approach**: 3 critical evaluation concepts
  - Selection of test video sequences
    - Follow a specified characterization of problems
    - Study one problem at a time, several levels of difficulty
    - Collect long sequences for significance
  - Ground truth definition
    - Up to the event level
    - Give clear and precise instructions to the annotator
      - E.g., annotate both visible and occluded part of objects
  - Metric definition
    - Set of metrics for each video processing task
    - Performance indicators: sensitivity and precision

Evaluation : current approach
(AT. NGHIEM)

- ETISEO limitations:
  - Selection of video sequence according to difficulty levels is subjective
  - Generalization of evaluation results is subjective.
  - One video sequence may contain several video processing problems at many difficulty levels

- Approach: treat each video processing problem separately
  - Define a measure to compute difficulty levels of input data (e.g. video sequences)
  - Select video sequences containing only the current problems at various difficulty levels
  - For each algorithm, determine the highest difficulty level for which this algorithm still has acceptable performance.

- Approach validation : applied to two problems
  - Detect weakly contrasted objects
  - Detect objects mixed with shadows

Video Understanding:
Learning Parameters (B.Georis)

- **Objective**: a learning tool to automatically tune algorithm parameters with experimental data

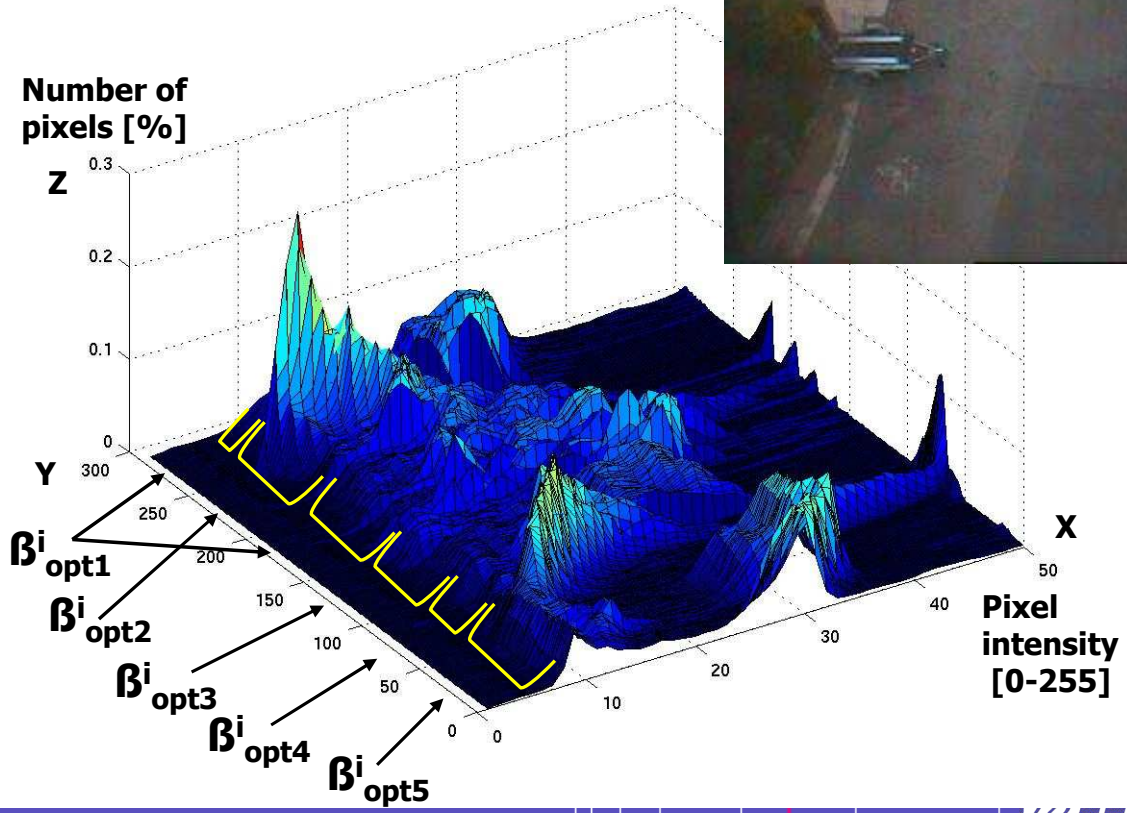
- Used for learning the segmentation parameters with respect to the illumination conditions

- **Method**
  - Identify a set of parameters of a task
    - 18 segmentation thresholds
  - depending on environment characteristics
    - Image intensity histogram
  - Study the variability of the characteristic
    - Histogram clustering -> 5 clusters
  - Determine optimal parameters for each cluster
    - Optimization of the 18 segmentation thresholds

Video Understanding:
Learning Parameters

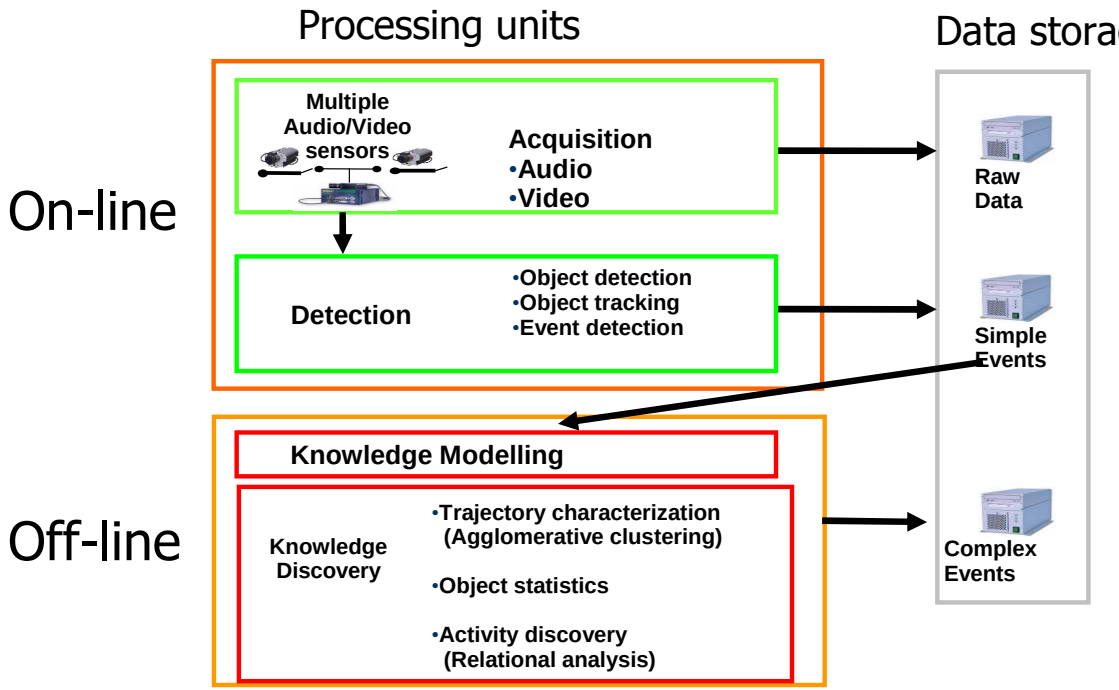
Camera View

Learning Parameters
Clustering the Image Histogram



Video Understanding : Knowledge Discovery (E. Corvee, JL. Patino_Vilchis)

- CARETAKER: An European initiative to provide an efficient tool for the **management of large multimedia** collections.

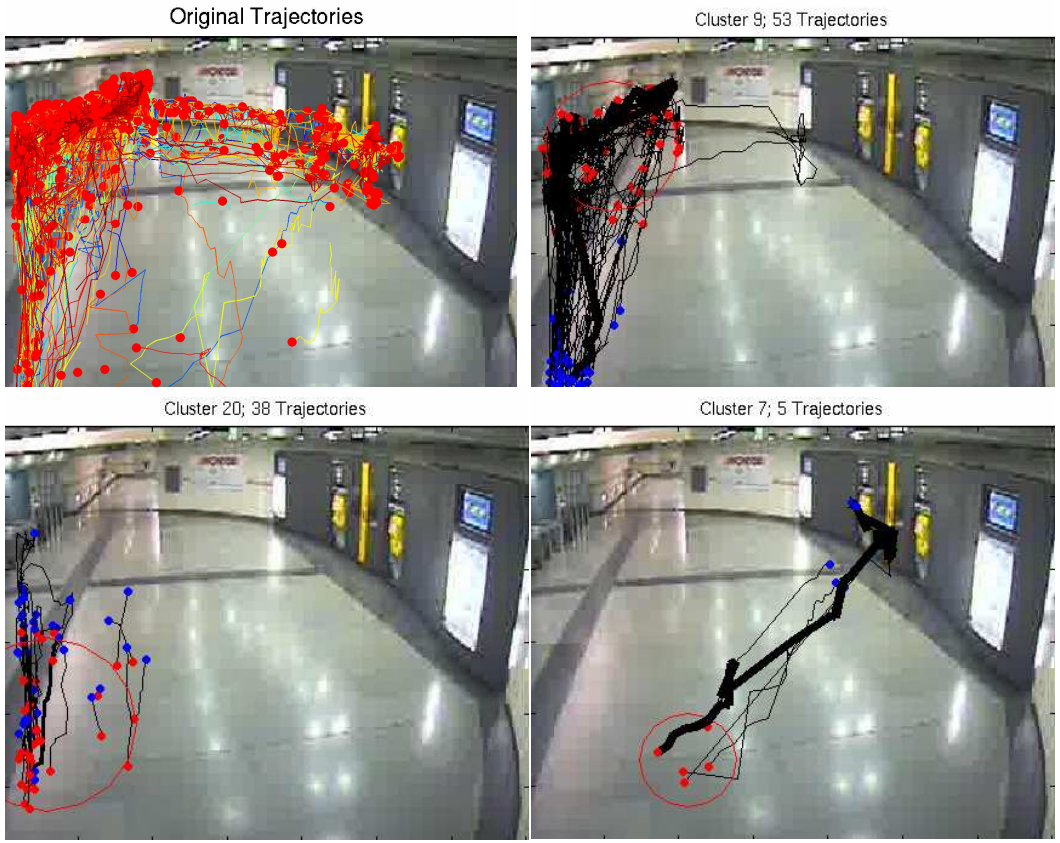


Knowledge Discovery: trajectory clustering

Objective: Clustering of trajectories into k groups to match people activities

- Feature set
  - Entry and exit points of an object
  - Direction, speed, duration, …
- Clustering techniques
  - Agglomerative Hierarchical Clustering.
  - K-means
  - Self-Organizing (Kohonen) Maps
- Evaluation of each cluster set

Results on Torino subway (45min), 2052 trajectories



Knowledge Discovery: achievements

- Computes on-line simple events and the interactions between moving objects and between contextual objects.

- Semantic knowledge is extracted by the off-line long term analysis of these interactions:
  - 70% of people are coming from north entrance
  - Most people spend 10 sec in the hall
  - 64% of people are going directly to the gates without stopping at the ticket machine
  - At rush hours people are 40% quicker to buy a ticket
  - …

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 …
  - **Knowledge** structured within the different abstraction levels (i.e. processing worlds)
    - Formal description of the empty scene
    - Structures for algorithm parameters
    - Structures for object detection rules, tracking rules, fusion rules, …
    - Operational **language** for **event** recognition (more than 60 states and events), video event ontology
  - Tools for **knowledge** management
    - Metrics, tools for performance evaluation, learning
    - Parsers, Formats for data exchange
    - …

Conclusion: perspectives

- Object and video event detection
  - Finer human shape description: *gesture models*
  - Video analysis robustness*: reliability computation*
- Knowledge Acquisition
  - Design of learning techniques to complement a priori knowledge:
    - visual concept learning
    - scenario model learning
- System Reusability
  - Use of program supervision techniques: *dynamic configuration of programs and parameters*
  - Scaling issue: managing large network of heterogeneous sensors (cameras, microphones, optical cells, radars….)