Scene Understanding

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

Francois BREMOND

PULSAR project-team,
INRIA Sophia Antipolis, FRANCE

Francois.Bremond@sophia.inria.fr
http://www-sop.inria.fr/pulsar/

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

Video Understanding

Objective:

Designing systems for

Real time recognition of human activities observed by sensors

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

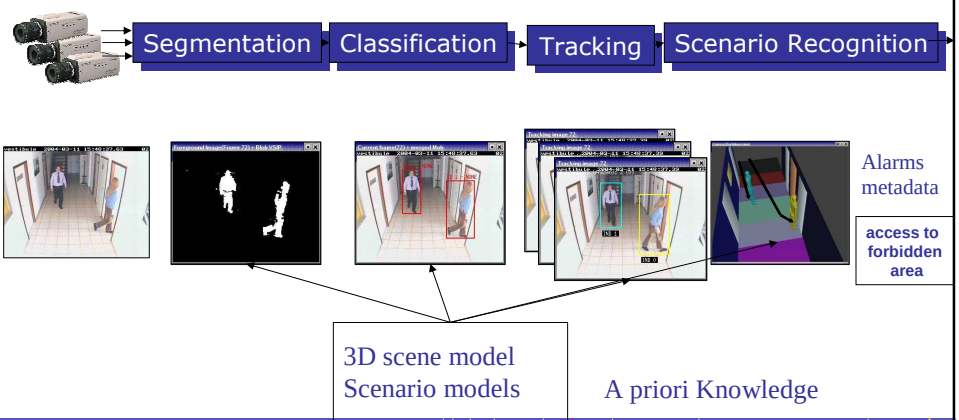
3 parts:

1. perception, detection, classification, tracking and multi-sensor fusion,
2. spatio-temporal reasoning and activity recognition,
3. evaluation, designing systems, autonomous systems, activity learning and clustering.

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

Video Understanding

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

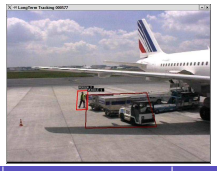


Video Understanding Applications

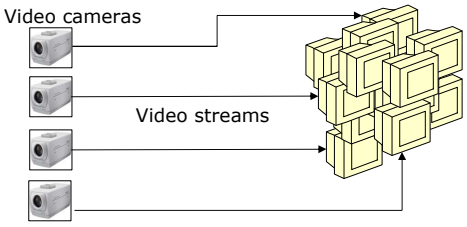
- Strong impact for visual surveillance in transportation (metro station, trains, airports, aircraft, harbors)
- Control access, intrusion detection and Video surveillance in building
- Traffic monitoring (parking, vehicle counting, street monitoring, driver assistance)
- Bank agency monitoring
- Risk 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 (20 persons): http://www.keeneo.com/

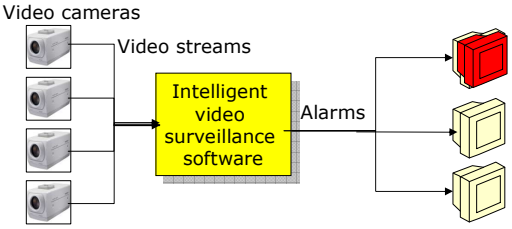


Intelligent Visual Surveillance



➢ **Huge information flow**
➢ **Few pertinent information**



➢ **Selection of the information**
➢ **Increase of the detection rate**

Video Understanding Application

Typical application-1:

European project **ADVISOR:**
(Annotated **D**igital **V**ideo for **I**ntelligent **S**urveillance and **O**ptimised Retrieval, 2000 - 2003)

- Intelligent system of video surveillance in metros
- Problem : 1000 cameras but few human operators
- Automatic selection in real time of the cameras viewing abnormal behaviours
- Automatic annotation of recognised behaviors in a video data base using XML

Video Understanding Application

Typical application-2 :
industrial project Cassiopée

Objectives :

- To build a Video Surveillance platform for automatic monitoring of bank agencies
- To detect suspicious behaviours leading to a risk
- Enabling a feedback to human operators for checking alarms
- To be ready for next aggression type

Video Understanding: Domains

- Smart Sensors: Acquisition (dedicated hardware), thermal, omni-directional, PTZ, cmos, IP, tri CCD, FPGA, DSP, GPU.
- 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

Video sequence categorization :

V1) Acquisition information:

- V1.1) Camera configuration: mono or multi cameras,
- V1.2) Camera type: CCD, CMOS, large field of view, colour, thermal cameras (infrared),
- V1.3) Compression ratio: no compression up to high compression,
- V1.4) Camera motion: static, oscillations (e.g., camera on a pillar agitated by the wind), relative motion (e.g., camera looking outside a train), vibrations (e.g., camera looking inside a train),
- V1.5) Camera position: top view, side view, close view, far view,
- V1.6) Camera frame rate: from 25 down to 1 frame per second,
- V1.7) Image resolution: from low to high resolution,

V2) Scene information:

- V2.1) Classes of physical objects of interest: people, vehicles, crowd, mix of people and vehicles,
- V2.2) Scene type: indoor, outdoor or both,
- V2.3) Scene location: parking, tarmac of airport, office, road, bus, a park,
- V2.4) Weather conditions: night, sun, clouds, rain (falling and settled), fog, snow, sunset, sunrise,
- V2.5) Clutter: empty scenes up to scenes containing many contextual objects (e.g., desk, chair),
- V2.6) Illumination conditions: artificial versus natural light, both artificial and natural light,
- V2.7) Illumination strength: from dark to bright scenes,

Video Understanding: Issues

Video sequence categorization :

V3) Technical issues:

- V3.1) Illumination changes: none, slow or fast variations,
- V3.2) Reflections: reflections due to windows, reflections in pools of standing water, reflections,
- V3.3) Shadows: scenes containing weak shadows up to scenes containing contrasted shadows (with textured or coloured background),
- V3.4) Moving Contextual objects: displacement of a chair, escalator management, oscillation of trees and bushes, curtains,
- V3.5) Static occlusion: no occlusion up to partial and full occlusion due to contextual objects,
- V3.6) Dynamic occlusion: none up to a person occluded by a car, by another person,
- V3.7) Crossings of physical objects: none up to high frequency of crossings and high number of implied objects,
- V3.8) Distance between the camera and physical objects of interest: close up to far,
- V3.9) Speed of physical objects of interest: stopped, slow or fast objects,
- V3.10) Posture/orientation of physical objects of interest: lying, crouching, sitting, standing,
- V3.11) Calibration issues: little or large perspective distortion,

Video Understanding: Issues

Video sequence categorization :

V4) Application type:

- V4.1) Tool box : primitive events, enter/exit zone, change zone, running, following someone, getting close,
- V4.2) Intrusion detection: person in a sterile perimeter zone, car in no parking zones,
- V4.3) Suspicious behaviour: violence, fraud, tagging, loitering, vandalism, stealing, abandoned bag,
- V4.4) Monitoring: traffic jam detection, counter flow detection, activity optimization, homecare,
- V4.5) Statistical estimation: people counting, car speed estimation, data mining, video retrieval,
- V4.6) Simulation: risk management,
- V4.7) Biometry and object classification: fingerprint, face, iris, gait, soft biometry, license plate, pedestrian.
- V4.8) Interaction and 3D animation: 3D motion sensor (Kinect), action recognition, serious games.

Video Understanding: Issues

Successful application: right balance between

- Structured scene: constant lighting, low people density, repetitive behaviours,
- Simple technology: robust, low energy consumption, easy to set up, to maintain,
- Strong motivation: fast payback investment, regulation,
- Cheap solution: 120 to 3000 euros per smart camera.

Commercial products:

- Intrusion detection: ObjectVideo, Keeneo, Evitech, FoxStream, IOimage, Acic,…
- Traffic monitoring: Citilog, Traficon,…
- Swimming pool surveillance: Poseidon,…
- Parking monitoring: Ivisiotec,…
- Abandoned Luggage: Ipsotek,…
- Biometry: Sagem, Sarnof,…
- Integrators: Honeywell, Thales, IBM, Siemens, GE, …
- Camera providers: Bosh, Sony, Panasonic, Axis, …

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

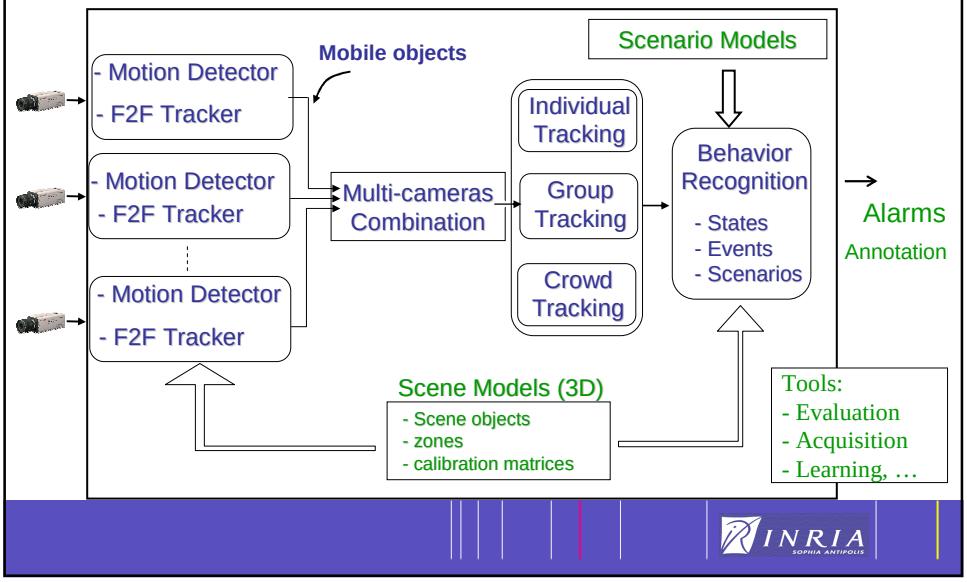
while focusing on the easy generation of dedicated systems based on

- 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)*

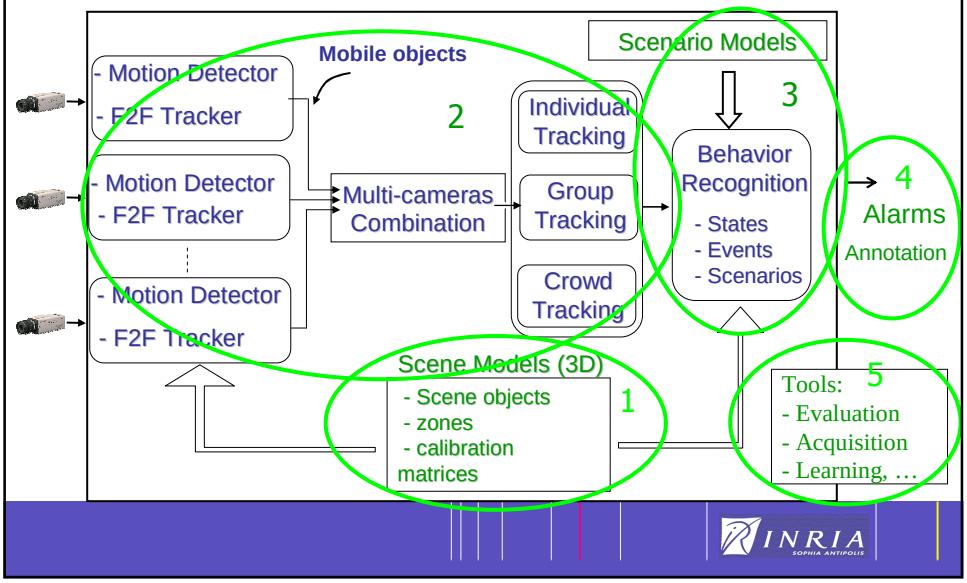
⇨ **Extract and structure knowledge (invariants & models) for**

- 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



Video Understanding



Outline (1/2)

Knowledge Representation [WSCG02], [Springer-Verlag11]

Perception

- People detection [IDSS03a], [ICDP09], [IJPRAI09]
- Posture recognition [VSPETS03], [PRLetter06], [AVSS10]
- Coherent Motion Regions [ACVIS08], [PETS09]
- Action Recognition [CVPR10]

4D coherency

- People tracking [IDSS03b], [CVDP02], [VISAP08], [ICDP09], [Neurocomputing11], [InTech11]
- Multi sensor combination [ACV02], [ICDP06a], [SFTAG09]
- People recognition [AVSS05a], [ICDP09], [IJPRAI09]

Outline (2/2)

Event representation **[KES02], [ECAI02]**

Event recognition:

- Finite state automata [ICNSC04]
- Bayesian network [ICVS03b]
- Temporal constraints [AVSS05b], [IJCAI03], [ICVS03a], [PhDTV04], [ICDP06], [ICDP09]

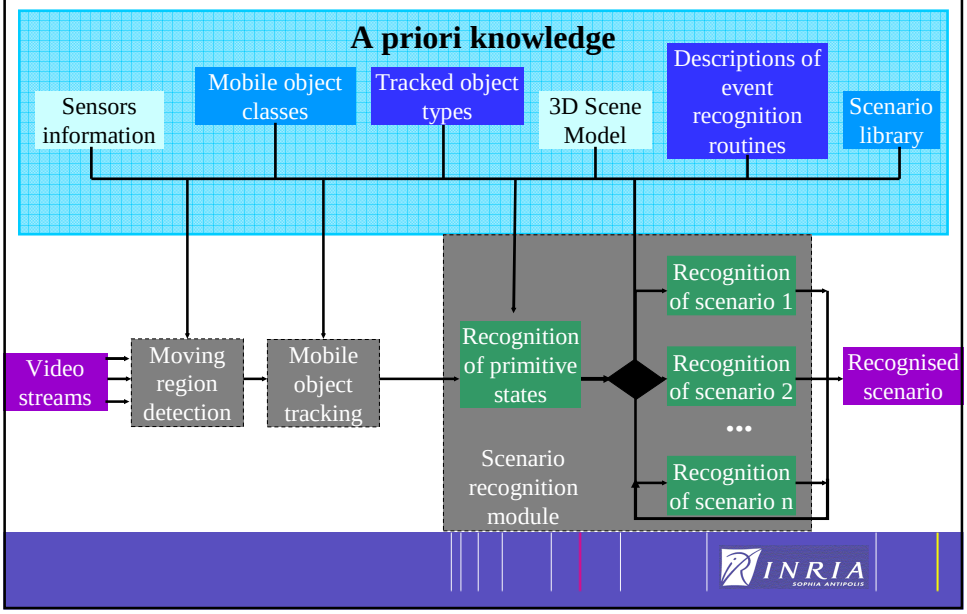
Autonomous systems:

- performance evaluation [VSPETS05], [PETS05], [IDSS04], [ICVIIP03], [WMVC07], [AVSS10]
- program supervision [ICVS06c], [ICVIIP04], [MVA06a]
- parameter learning [PhDBG06]
- knowledge discovery [ICDP06], [VIE07], [Springer-Verlag11]
- learning scenario models [ICVS06a], [ICDP06b], [CV08]

Results and demonstrations: metro, bank, train, airport.

Knowledge Representation

Knowledge Representation



Knowledge Representation: 3D Scene Model

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)

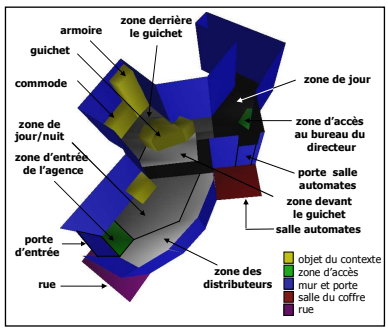
Role:

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

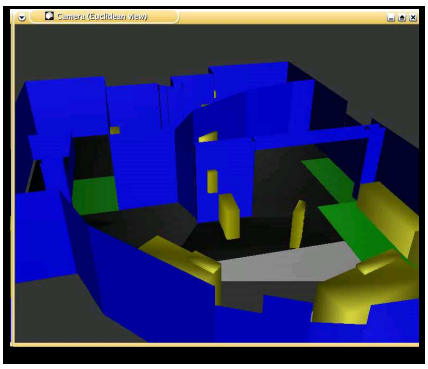
Knowledge Representation : 3D Scene Model

3D Model of 2 bank agencies

Les Hauts de Lagny

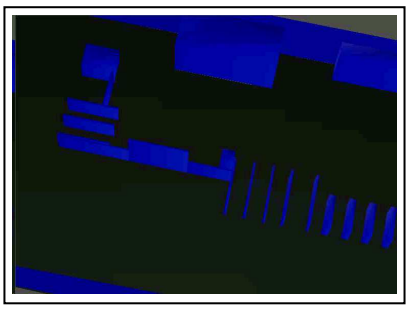


Villeparisis

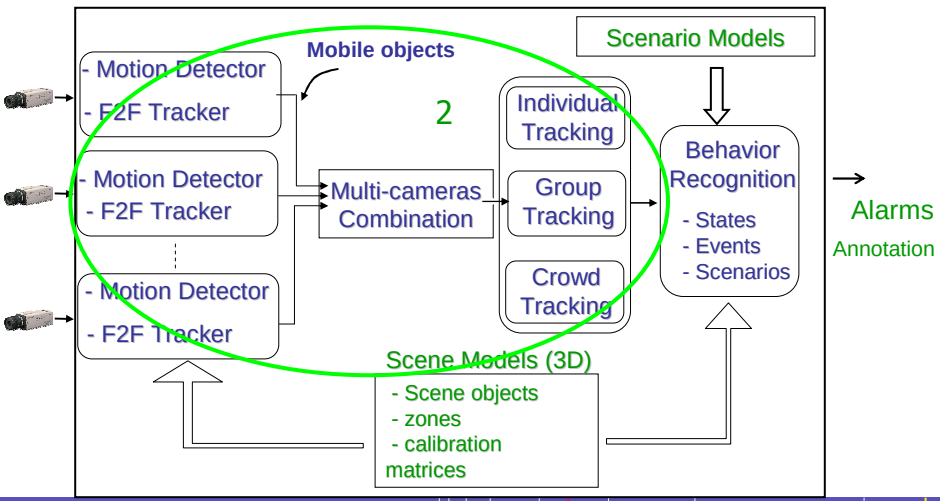


Knowledge Representation: 3D Scene Model

Barcelona Metro Station Sagrada Famiglia mezzanine
(cameras C10, C11 and C12)



Video Understanding



People detection

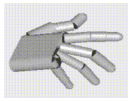
Estimation of Optical Flow

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

Object detection



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

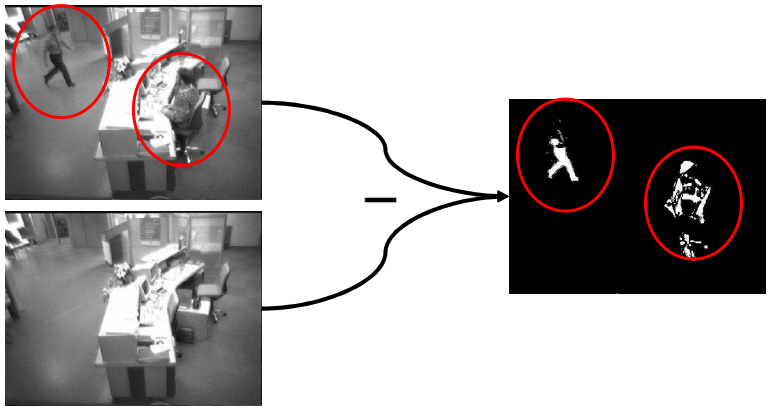




Reference image subtraction

- Need of static cameras
  - Most robust approach (model of background image)
  - Most common approach even in case of PTZ, mobile cameras

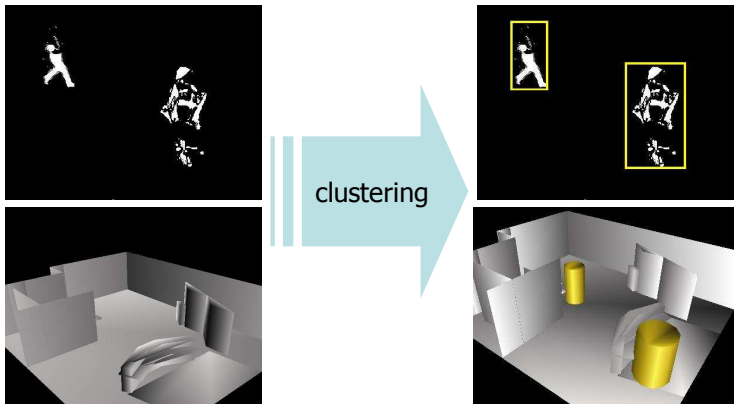
People detection

Difference between the current image and a reference image (computed) of the empty scene



People detection

Approach: Group the moving pixels together to obtain a moving region matching a mobile object model



People detection: Reference Image

Reference image representation:

- Non parametric model (set of images)
- K Multi-Gaussians (means and variances)
- Code Book (min, max)

Update of reference image

- Take into account slow illumination change
- Managing sudden and strong illumination change
- Managing large object appearance wrt camera gain control

Issues:

• Integration of noise (opened door, shadows, reflection, parked car, fountain, trees) in the reference image, of shadows.

• Ghost detection, multi-layer background,

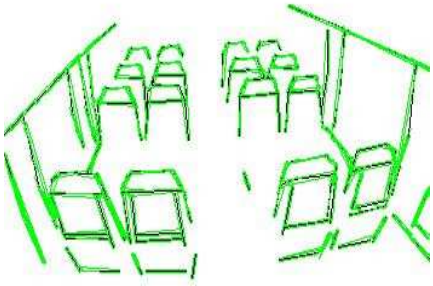
• Compensate for Ego-Motion of moving camera, handling parallax.

People detection: Reference Image Issues

Reference image representation using characteristic points

People detection: Reference Image issues

Reference image representation using characteristic contours



People detection

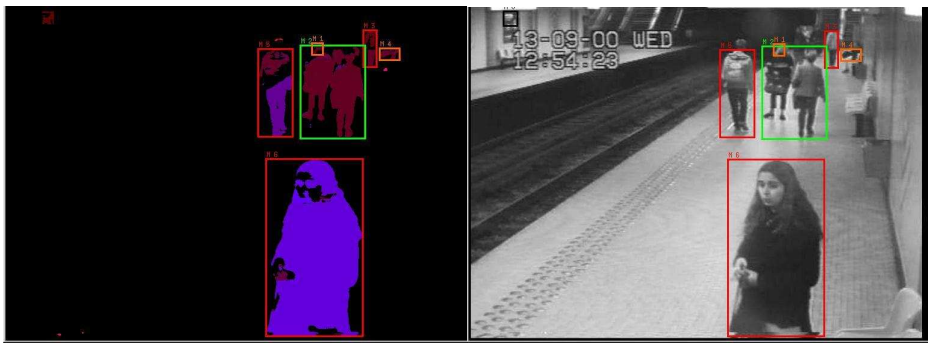
5 levels of people classification

- 3D ratio height/width
- 3D parallelepiped
- 3D articulate human model
- People classifier based on local descriptors
- 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



People detection

Utilization of the 3D geometric scene model

People detection

People counting in bank agency

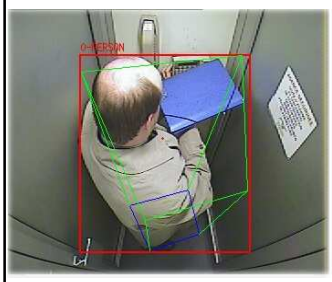
People detection

People counting in bank agency

People Counting scenario 2

People detection (M. Zuniga)

Classification into 3 people classes : 1Person, 2Persons, 3Persons, Unknown



People detection

Proposed Approach

- calculation of 3D parallelepiped model MO
- Given a 2D blob

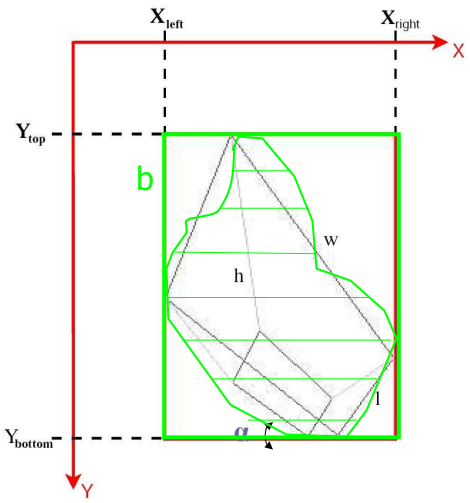
b = (Xleft, Ybottom, Xright, Ytop).

the problem becomes:

$$MO = F(\alpha, h \mid b)$$

- Solve the linear system:

– 8 unknowns.
– 4 equations from 2D borders.
– 4 equations from perpendicularity between base segments.



People detection (M. Zuniga)

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

Posture Recognition

Posture Recognition (B. Boulay)

Recognition of human body postures :

- with only one static camera
- in real time

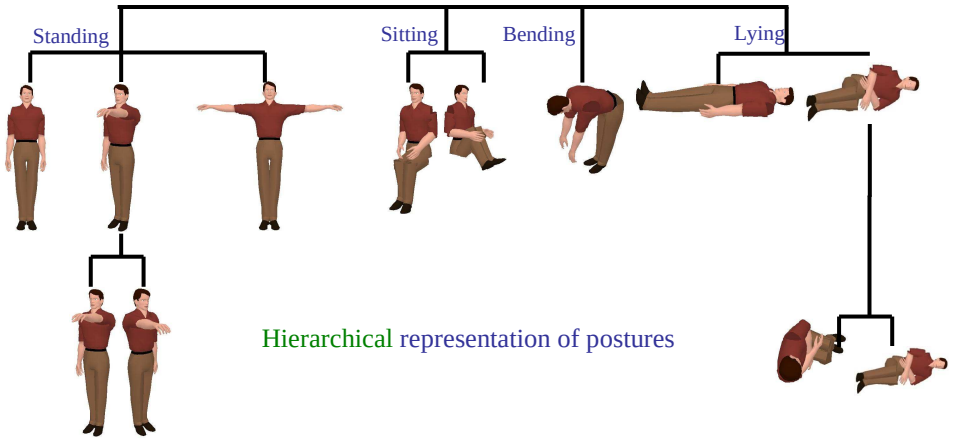
Existing approaches can be classified :

- 2D approaches : depend on camera view point
- 3D approaches : markers or time expensive

Approach: combining

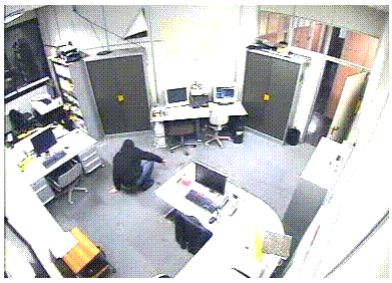
- 2D techniques (eg. Horizontal & Vertical projections of moving pixels)
- **3D articulate human** model (10 joints and 20 body parts)

Posture Recognition : Set of Specific Postures



Hierarchical representation of postures

Posture Recognition : silhouette comparison

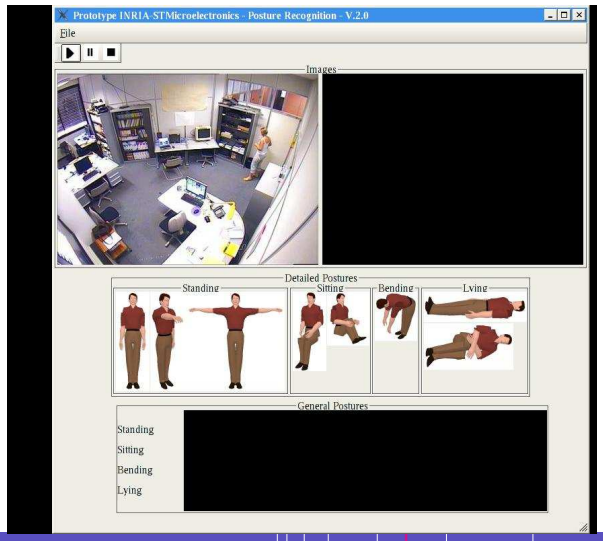


Real world

Virtual world

Generated silhouettes

Posture Recognition : results



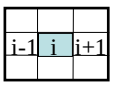
Posture Recognition : results

Complex Scenes: interest point descriptors

- To characterize people: Computation of interest point descriptors
  - Point Detection (e.g. Corners)
    - Find salient points
      - characterizing people (well contrasted) and
      - where the motion can be easily tracked.
    - Ensure uniform distribution of feature through the body.
  - Descriptors : Extraction of 2D Histogram of Oriented Gradients (HOG) or SIFT, SURF, OF3D,…
    - For each interest point compute a 2D HOG descriptor.

Complex Scenes: interest point descriptors

- Corners detection:



  - Shi-Tomasi features:

Given an image I and its gradients g_x and g_y respectively through the x axis and the y axis.

The Harris matrix for an image pixel in a window of size (u,v) is:

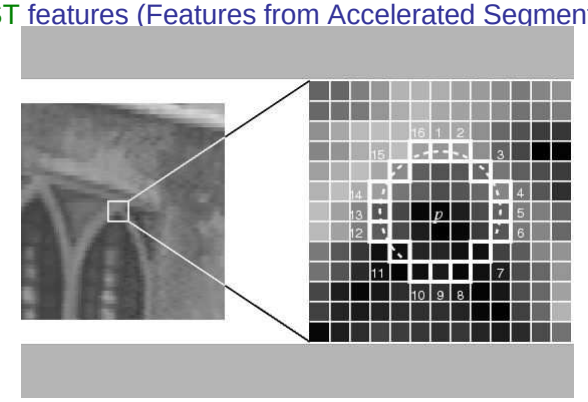
$$H = \sum_u \sum_v \begin{bmatrix} g_x^2 & g_x.g_y \\ g_x.g_y & g_y^2 \end{bmatrix}$$

[Shi 1994] prove that $\min(\lambda_1, \lambda_2)$ is a good measure of corner strength. Where λ_1 and λ_2 are the Eigen values of the Harris matrix.

Complex Scenes: interest point descriptors

- Corners detection (cont'd):

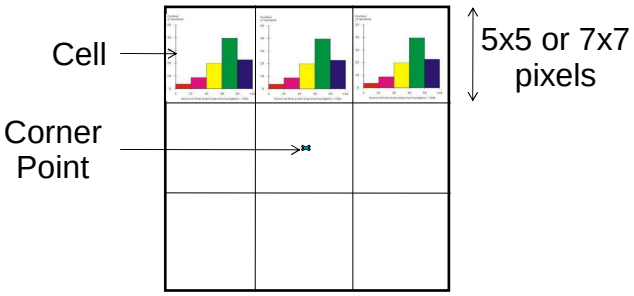
  FAST features (Features from Accelerated Segment Test) :



Complex Scenes: interest point descriptors

- HOG : 2D Histogram of Oriented Gradient Descriptor

  Descriptor bloc (3x3 cells):



Complex Scenes: interest point descriptors

- 2D HOG Descriptor (cont'd):
- For each pixel in the descriptor bloc we compute:

$$g(u,v)=\sqrt{g_x(u,v)^2+g_y(u,v)^2} \quad \text{and} \quad \theta(u,v) = \tan^{-1}\left(\frac{g_y(u,v)}{g_x(u,v)}\right)$$

- For each cell c_{ij} in the descriptor bloc we compute:

$$f_{ij} = [\, f_{ij}(\beta)]_{\beta \in [1..K]}$$

where K=8 is the number of orientation bins and :

$$f_{ij}(\beta) = \sum_{(u,v)\in c_{ij}} g(u,v).\delta[bin(u,v) - \beta]$$

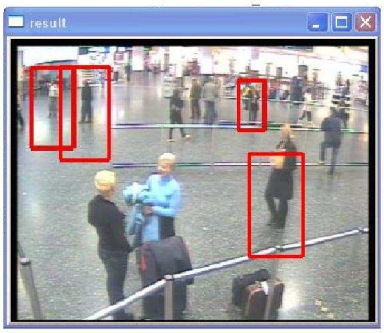
Complex Scenes: People detection

Introduction

- **HOG descriptors** are widely used for people detection
  - e.g Dalal & Triggs 05 (Implemented in OpenCv library ~1 fps)
- Main issue: complex people appearances:
  - Clothing (e.g. long coat, hat)
  - **Occlusion** issue (e.g. caused by another person, a carried luggage)
  - **Postures** (e.g. running, slightly bent)
  - Camera's viewpoint
- Drawbacks
  - **Noisy** detection
  - Database dependency
  - Viewpoint restriction of DB: camera facing up-right people
  - Requires time consuming **training** phase
  - Feature information not available (during detection)

Complex Scenes: People detection

- People classifier based on HOG features and Adaboost cascade at Gatwick airport (Trecvid 2008)



Complex Scenes: People detection

People detector training - **Find the most dominant HOG orientation**

Input sample

48x96

Sobel convolution

8 Bins voting

8 Integral Images

8x8 cell scanning every 4 pixels = 230 cells

Cell

$c=\{1:230\}$

230 cells of 8 bins HOG

HOG

$h_c(b)$

$Argmax_b\ h_c(b)$

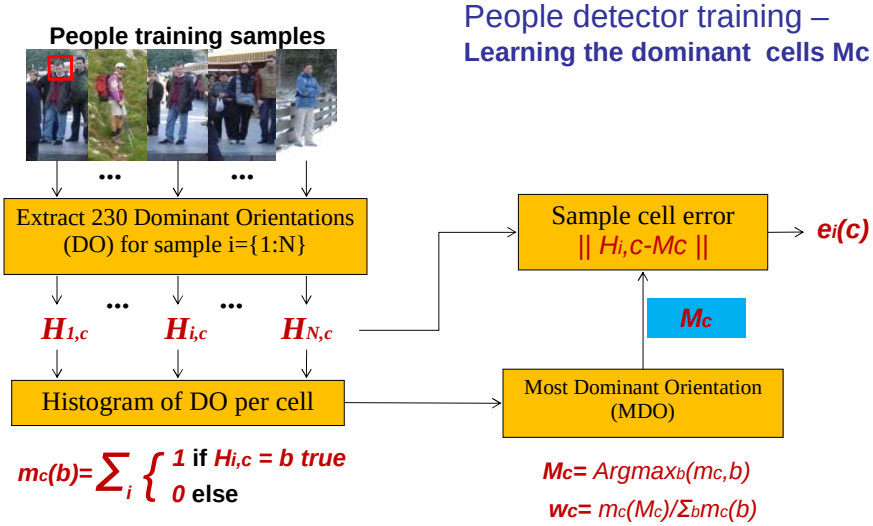
Dominant Orientation

230 Dominant Orientations

$H_c,\ c=\{1:230\}$

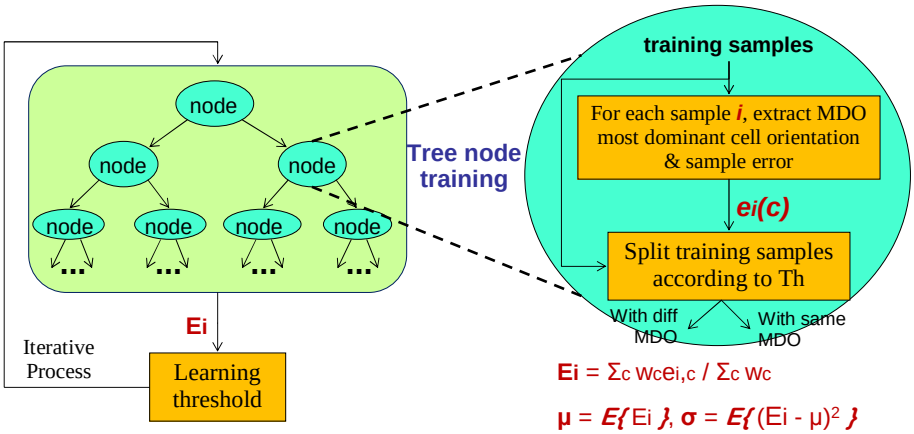
Complex Scenes: People detection

People detector training – **Learning the dominant cells Mc**



Complex Scenes: People detection

People detector training - **Hierarchical trees training**



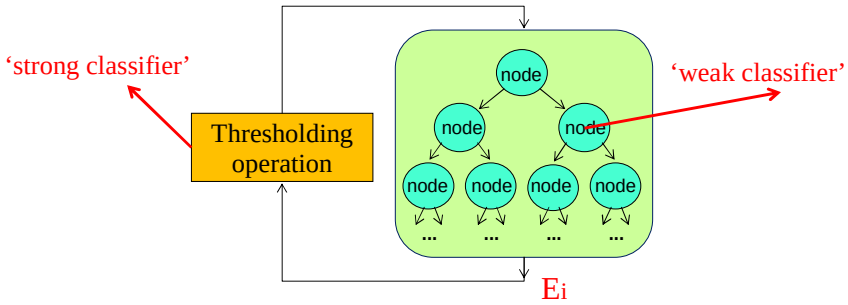
Ei = Σc wcei,c / Σc wc

μ = E{ Ei }, σ = E{ (Ei - μ)² }

Threshold = 2.8 * | Ei – μ | / σ

Complex Scenes: People detection

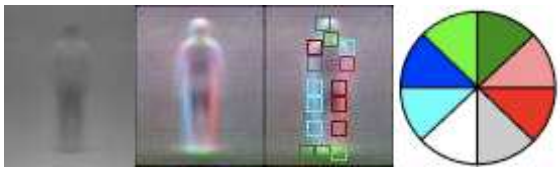
People detector training - **Hierarchical trees training**

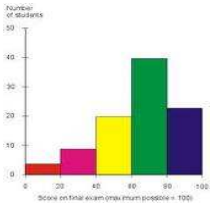


- Iteration process involves **several trees** training and classification: '**strong classifiers**'
- After several iterations, Ei **converges**.
- Sample errors Ei assumed **normally** distributed
- Trees are constructed with maximum 6 levels of weak classifiers and maximum 10 strong classifiers

Complex Scenes: People detection

- HOG descriptors as visual signature
  - HOG extracted in cells of size 8x8 pixels
  - During training (2000 + and – image samples):
    - Automatic selection of the 15 cells i.e. giving the strongest mean edge magnitude

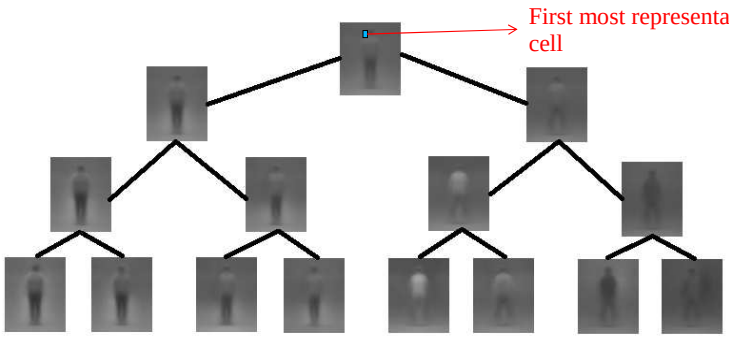




mean edge magnitude
over the + training image samples

Complex Scenes: People detection

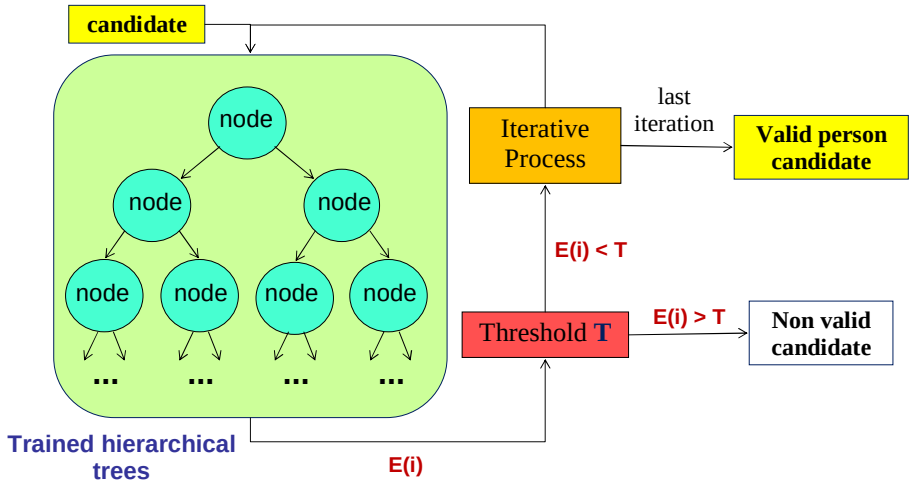
- Tree of **People Samples** organized along the strongest mean edge magnitude HOG:
  - Postures defines human global visual signatures
  - Best cells location and content vary from one posture to another
  - Postures categorized in a hierarchical tree



Average training samples. Using MIT dataset on a 3 levels tree

Complex Scenes: People detection

Detection process - **HOG classification**



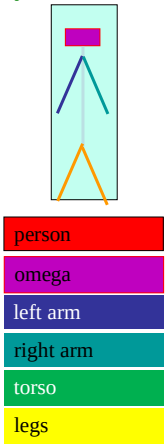
Complex Scenes: People detection

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

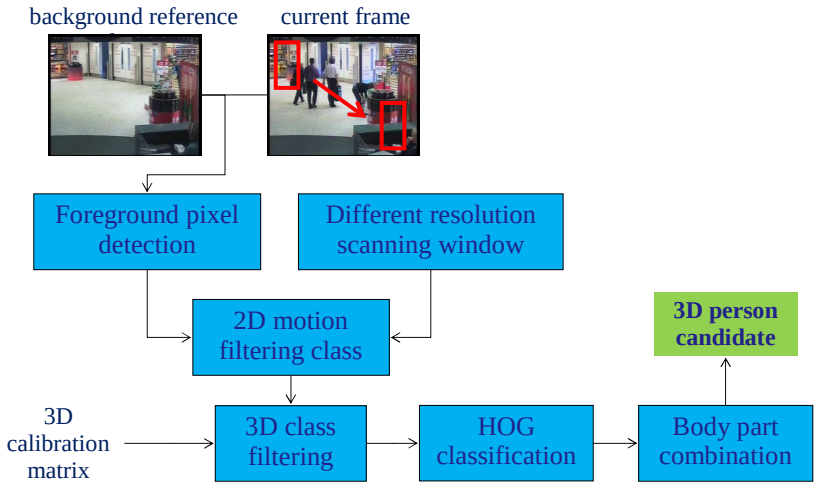
Example of detected with corresponding HOG cells

Detection examples



Complex Scenes: People detection

Algorithm overview



Complex Scenes: People detection

Object detection - Filtering

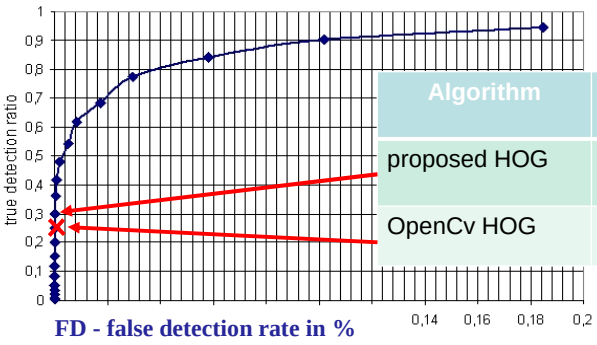
- 2D Foreground filtering:
  - Foreground pixels thresholded from **background** pixels
  - Foreground objects and **body parts** are discarded if they contain less than 50% of foreground pixels (use of Integral Image to rapidly calculate this percentage)
- 3D filtering:

  Use of Tsai calibrated camera and a 3D person model (size) to filter out non 3D person candidates
- Body part combination:
  - People must be associated with at least N body parts
- Overlapping filtering:
  - Multi resolution scanning gives rises to overlapping candidates
  - Averaging operation performed to fuse locally overlapping candidates

Complex Scenes: People detection

Evaluation of people detection on a testing database

Input: NICTA database: 424 positive and 5000 negatives



Algorithm	TD rate %	FD rate %
proposed HOG	29.76	0
OpenCv HOG	25.01	0

Complex Scenes: People detection

Evaluation of people detection in video sequences

Method: Comparison with and without filtering scheme

Input: Caviar sequence 'cwbs1' and 5 sequences of TrecVid camera1

algorithm	False alarm rate: FA	Missed Detection rate: MD
OpenCv HOG	0.68	1.42
Our HOG without filtering	0.22	1.57
Our HOG with filtering	0.19	1.61

FA – Number of false alarms per frame
MD – number of missed detected ground truth per frame

Complex Scenes: People detection

Examples of HOG People in TSP

Complex Scenes: People detection

- Head and face detection
  - Head detected using same people detection approach.
    - Head training DB:
      - 1000 Manually cropped TrecVid heads plus
      - 419 TUD images
    - Speed increased when detecting in top part of people
  - Face detected using LBP (Local Binary Pattern) features
    - Face training DB:
      - MIT face database (2429 samples)
    - Training performed by Adaboost
    - Speed increased when detecting within head areas
  - Tracking is performed independently for each object class

Complex Scenes: People detection

Face classifier based on HOG and Adaboost cascade at Gatwick airport (Trecvid 2008)

Training based on CMU database:
http://vasc.ri.cmu.edu/idb/html/face/frontal_images/index.html

Training based on CMU database and reference image

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

INRIA SOPHIA ANTIPOLIS

Complex Scenes: Coherent Motion Regions

Based on KLT (Kanade-Lucas-Tomasi) tracking

Computation of 'interesting' feature points (corner points with strong gradients) and tracking them (i.e. extract motion-clues)

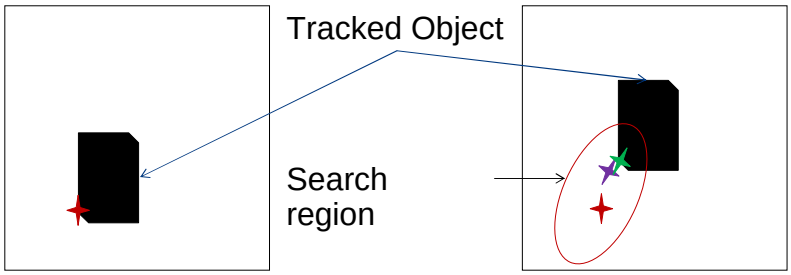
Cluster motion-clues of same directions on spatial locality

- define 8 principal directions of motion
- Clues with almost same directions are grouped together
- Coherent Motion Regions: clusters based on spatial locations

INRIA SOPHIA ANTIPOLIS

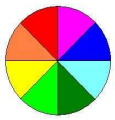
Complex Scenes: feature point tracking

- Feature point tracker:



- Previous position
- Predicted position using Kalman filter
- Corrected position

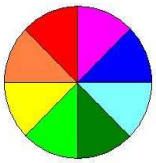
Results : Crowd Detection and Tracking



Results : Crowd Detection and Tracking

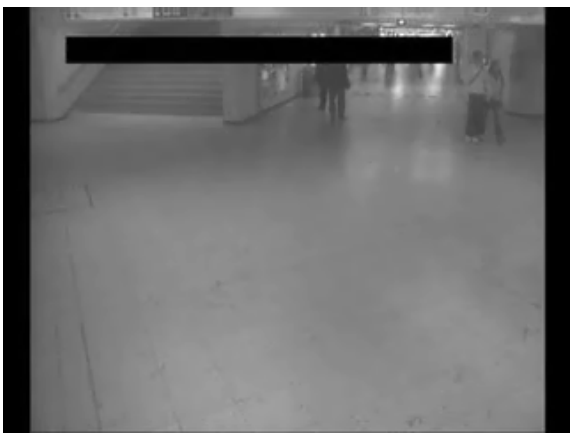
Results : Crowd Detection and Tracking

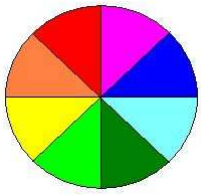
Results : Crowd Detection and Tracking



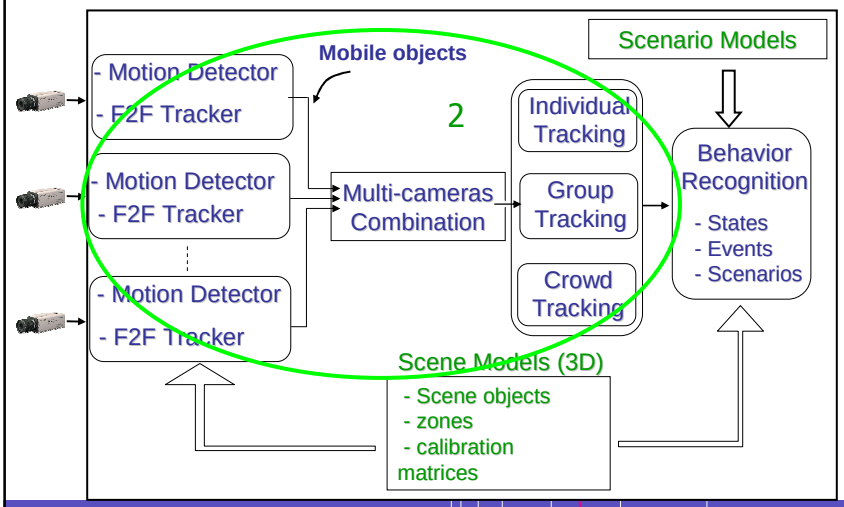
Coherent Motion Regions (MB. Kaaniche)

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





Video Understanding



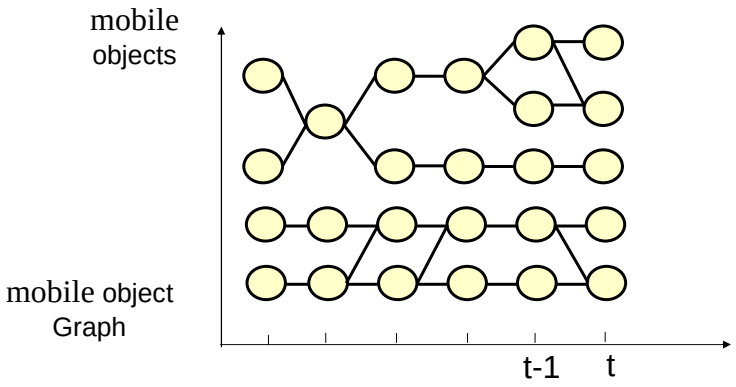
People tracking

People tracking

- Optical Flow and Local Feature tracking (texture, color, edge, point)
- 2D Region tracking based on
  - overlapping regions
  - 2D signature (dominant colors)
  - Contour tracking (Snakes, B-Splines, shape models)
- Object tracking based on 3D models

People tracking

Frame to frame tracking: For each image all newly detected moving regions are associated to the old ones through a graph



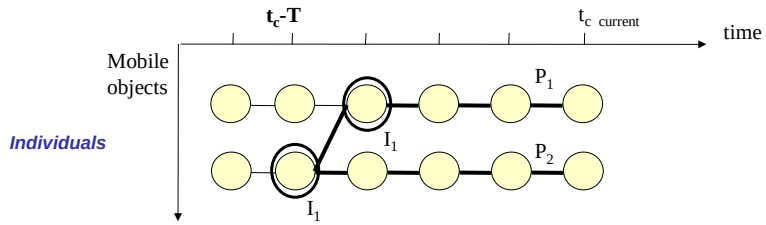
People tracking: individual tracking

Goal : To track isolated individual on a long time period

Method: Analysing of the mobile object graph

- Model of individual
- Model of individual trajectory

- Utilisation of a time delay to increase robustness



People tracking: individual tracking

mobile object: person

tracked INDIVIDUAL

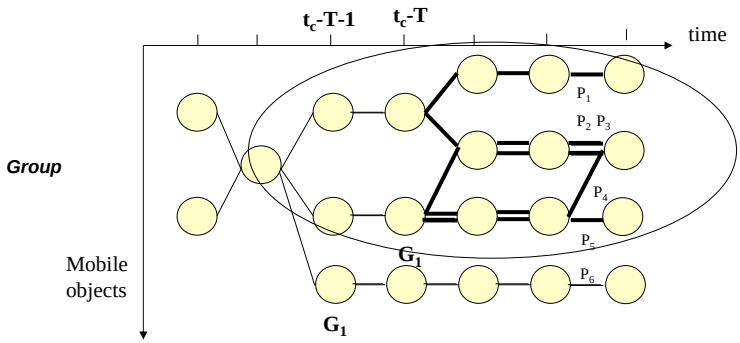
Limitations : Mixed of individuals in difficult situations (e.g. static and dynamic occlusion, long crossing)

People tracking: group tracking

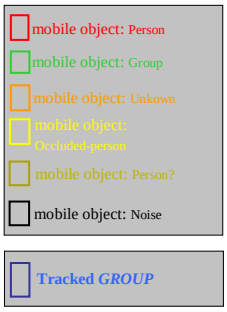
Goal : To track globally people over a long time period

Method: Analysis of the mobile object graph based on

Group model, Model of trajectories of people inside a group, time delay



People tracking: group tracking



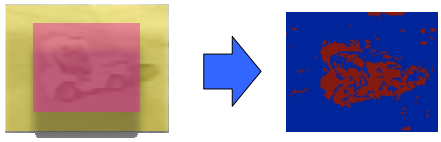
Limitations : - Imperfect estimation of the group size and location when there are shadows or reflections strongly contrasted.

- Imperfect estimation of the number of persons in the group when the persons are occluded, overlapping each others or in case of miss detection.

Online Adaptive Neural Classifier for Robust Tracking

Object/Background Separation:

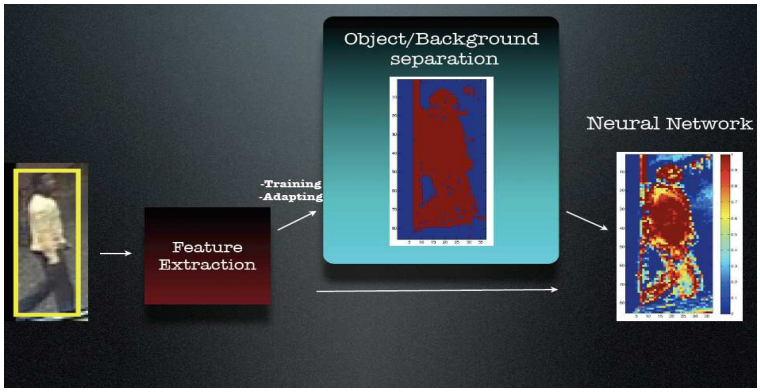
- To build an object model, an object/background separation scheme is used to identify the object/background pixels
- If the log-likelihood of object at frame is greater than threshold value then the pixel belongs to object class, otherwise not.



$$L_i = \log \frac{\max\{h_o(i), \varepsilon\}}{\max\{h_b(i), \varepsilon\}}$$

Online Adaptive Neural Classifier for Robust Tracking

The neural classifier is used to differentiate the feature vector of the object (inside) / from local background (outside)



Online Adaptive Neural Classifier for Robust Tracking

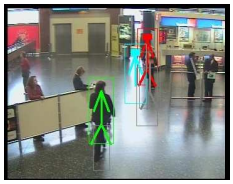
People detection and tracking

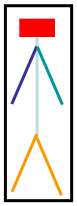
Complex Scenes:
People detection and tracking

People detection and tracking

People detection using HOG :

- Algorithm trained on different parts of the people database
  - omega (head+shoulder),
  - torso,
  - left arm,
  - right arm
  - legs
- Combination with body parts detectors
  - Gives more details about the detected persons



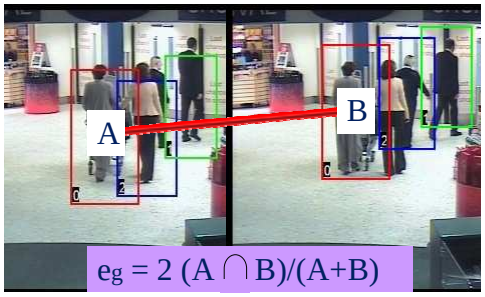


person
omega
right arm
left arm
torso
legs

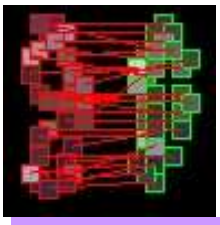
People detection and tracking

Frame to frame tracking: creating links between two objects in two frames based on:

- Geometric overlap: e_g = dice coefficient
- HOG map dissimilarity: e_h = Average of closest cells HOG differences



$e_g = 2\,(A \cap B)/(A+B)$



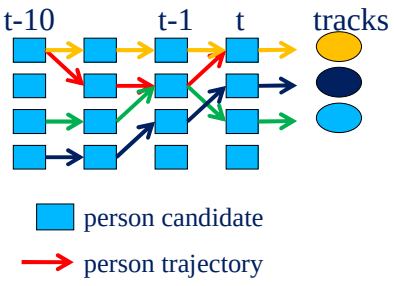
e_h =Somme{|□□|}

F2F link error = e_g . e_h

People detection and tracking

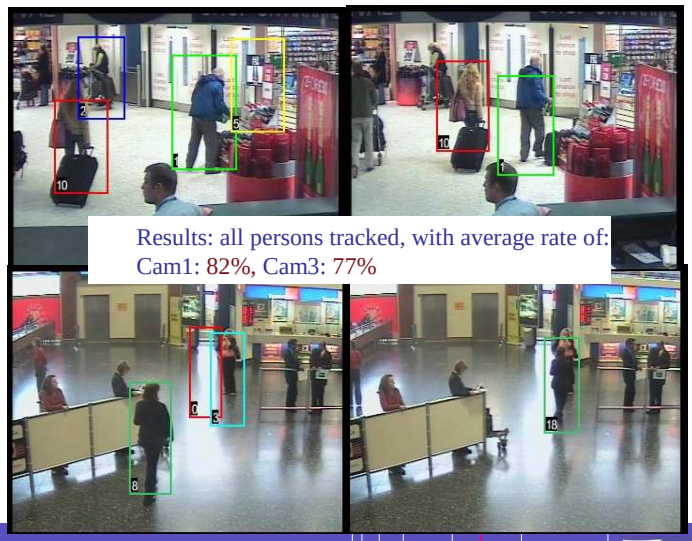
People tracking:

- Graph based long term tracking
  - Links between successive persons established
    - Based on best 2D/3D, descriptor similarities
    - Recorded in an array: history of the e.g. 10 last frames
  - Possible paths constructed and updated with new links
  - Best path leads to a person track



People detection and tracking

- Detection and tracking : results with TrecVid



Results: all persons tracked, with average rate of:
Cam1: 82%, Cam3: 77%

Gatwick cam1

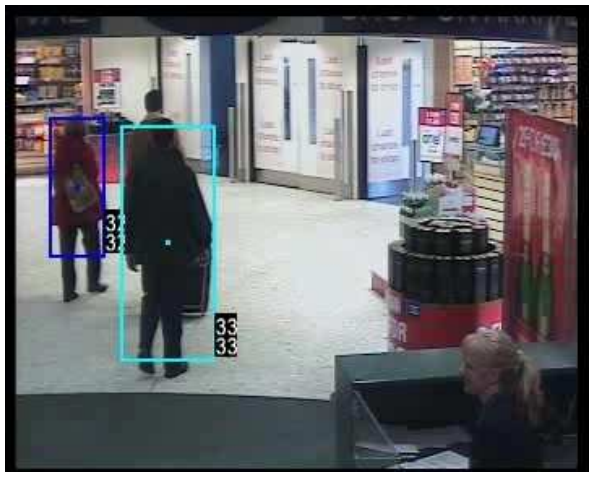
Gatwick cam3

People detection and tracking

- Detection and tracking : results with CAVIAR

People detection and tracking

Results



a background updating scheme was used for some sequences

People detection and tracking

Evaluation

algorithm	MF	MLT %	MTT %
A Tracker Geo	3.33	52.2	72.1
B Tracker HOG	3.27	56.7	73.5
C Combined tracker	2.88	57.3	73.4
Rank	C,B,A	C,B,A	B,C,A

Inputs:

5 sequences from TrecVid camera 1

Tracker Geo: Frame to frame F2F link calculated solely from 2D overlap factor (e_g)

Tracker HOG: Frame to frame F2F link calculated solely from HOG map dissimilarity (e_h)

MF: Mean Fragmentation rate (mean number of detected tracks per GT ID track)

MLT: Mean Longest Track lifetime (mean of the longest fragment for each GT track)

MTT: Mean Total Track lifetime (mean of all fragment total lifetime for each GT track)

People detection and tracking

Results: tracked people in red, head in green and faces in cyan

People detection and tracking

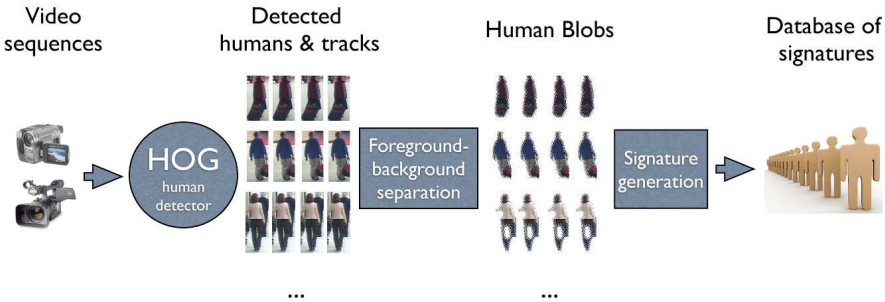
Results: tracked faces in higher resolution

People re-identification

- Re-identification:
  - The objective is to determine whether a given person of interest has already been observed over a network of cameras

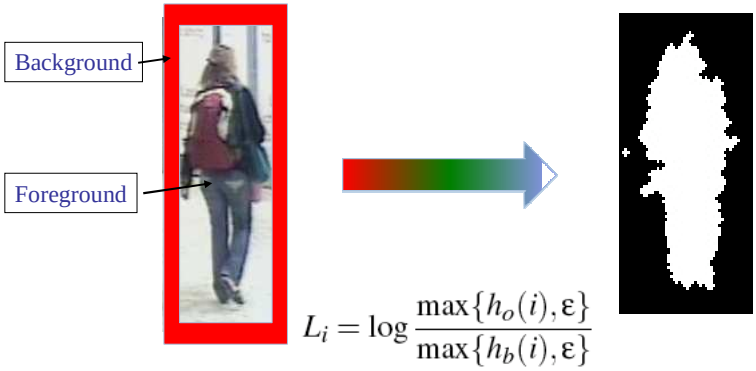
People re-identification

- The re-identification system



People re-identification

- Foreground-background separation



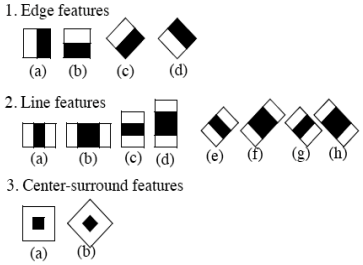
$$L_i = \log \frac{\max\{h_o(i), \varepsilon\}}{\max\{h_b(i), \varepsilon\}}$$

People re-identification

- Signature Computation
  - Find features which have a discriminative power (identification) concerning humans
- Co-variance matrices
- Haar-based signature : 20x40 x 14 = 11200 features

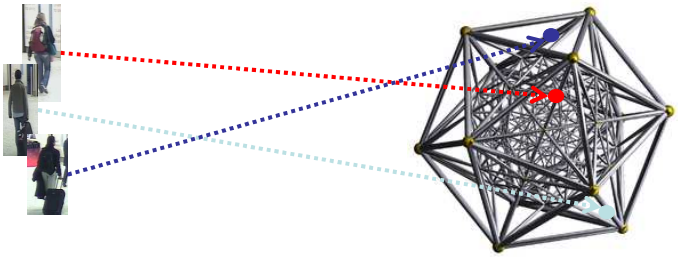
20x40 variables

(a) original image (b) mask (c) haar features



People re-identification

- Distance computation for haar-based signature



  - Distance definition

$$D(s_i, s_j) = 1 - \frac{\mathcal{V}_{s_i s_j}}{\min(\mathcal{V}_{s_i}, \mathcal{V}_{s_j})}.$$

People re-identification

- Dominant Color Descriptor (DCD) signature
  - DCD definition

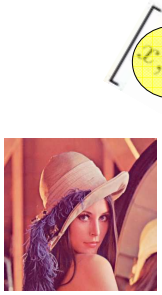
$$F = \{\{c_i, p_i\}, i = 1, \ldots, N\},$$

  - Signature

(a) original image (b) upper body part (c) lower body part

People re-identification

- Extraction of covariance matrices



$$\left[x, y, R_{xy}, G_{xy}, B_{xy}, \left\| \nabla R_{xy} \right\|, \theta R_{xy}, \left\| \nabla G_{xy} \right\|, \theta G_{xy}, \left\| \nabla B_{xy} \right\|, \theta B_{xy} \right]$$

LOCALIZATION

COLOR INFORMATION

GRADIENT INFORMATION

$$C_R = \frac{1}{n-1} \sum_{k=1}^{n} (f_k - \mu)(f_k - \mu)^T$$

$$C_R(i,j) = \frac{1}{n-1} \sum_{k=1}^{n} (f_k(i) - \mu(i))(f_k(j) - \mu(j))$$

O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In Proc. 9th European Conf. on Computer Vision, pages 589–600, 2006.

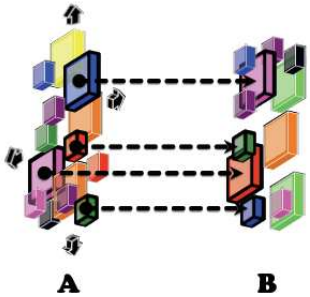
People re-identification

- Discriminate 2 signatures using Mean Covariance

$$\sigma_{i,j}^{c} = \frac{1}{n-1} \sum_{k=1;k\neq i}^{n} \rho^2(\mu_{i,j}^{c}, \mu_{k,j}^{c}).$$

People re-identification

- The distance between two human signatures



Every signature is a set of the covariance patches.

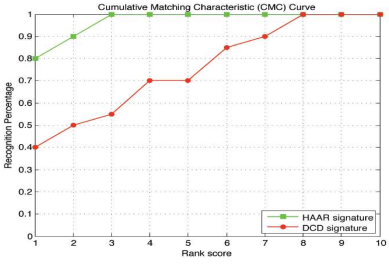
Signature A is shifted left/right/up and down to find out the best corresponding patches in second signature B (position of a patch determines matching).

Connections in the figure represent corresponding patches. Some connections are suppressed for clarity.

$$S(\mathfrak{s}_A, \mathfrak{s}_B) = \sum_{i \in K} \frac{\sigma_{A,i} + \sigma_{B,i}}{\rho(\mu_{A,i}, \mu_{B,i})} / \parallel K \parallel$$

People re-identification

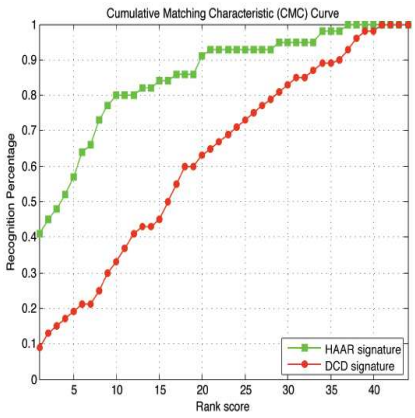
- Experimental results
  - 15 people from CAVIAR data



People re-identification

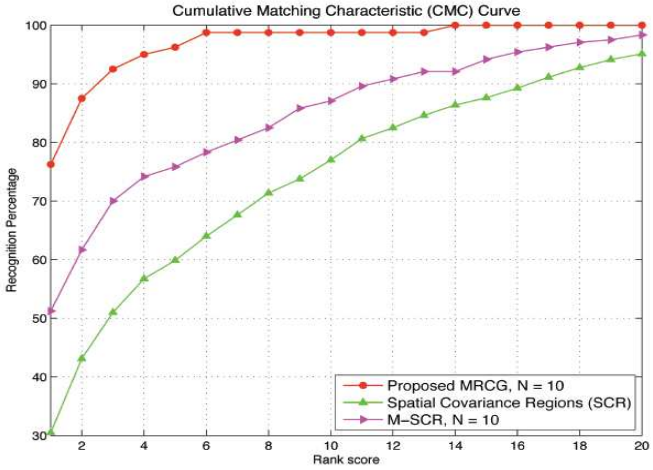
People re-identification

- Experimental results
  - 40 people from i-LIDS (TRECVID) data



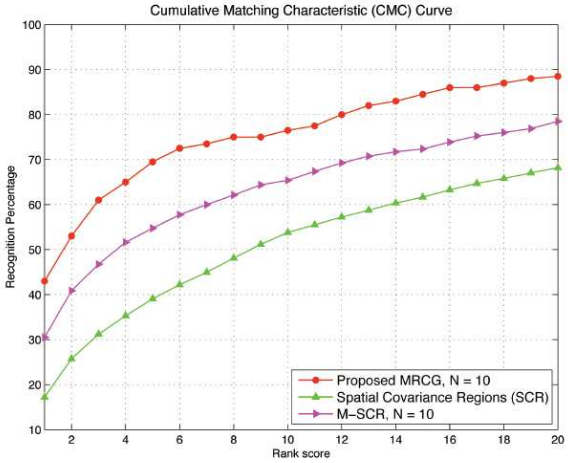
People re-identification

- Experimental results
  - i-LIDS data set (40 individuals) – manually detected



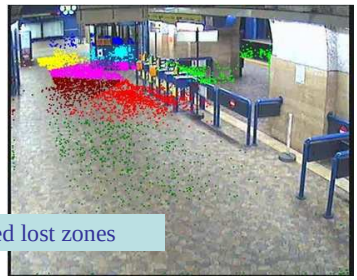
People re-identification

- Experimental results
  - i-LIDS data set (100 individuals) – automatically detected

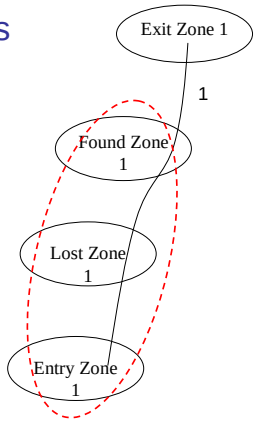


Global tracking: repairing lost trajectories

- Stitching 2 trajectories using zone triplets
  - Complete trajectories that pass through: ‘entry zone’, ‘lost zone’ and ‘found zone’, are used to construct the zone triplets.

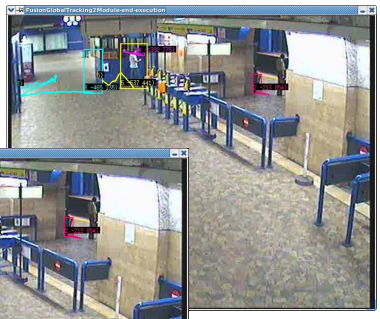


8 learned lost zones

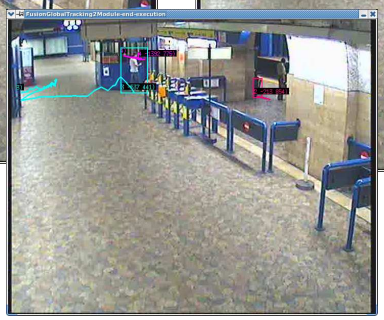


Global tracking: repairing lost trajectories

t = 709 s



t = 711s without the algorithm



t = 711s with the algorithm

Global tracking: repairing lost trajectories

t = 901 s

t = 903s without the algorithm

t = 903s with the algorithm

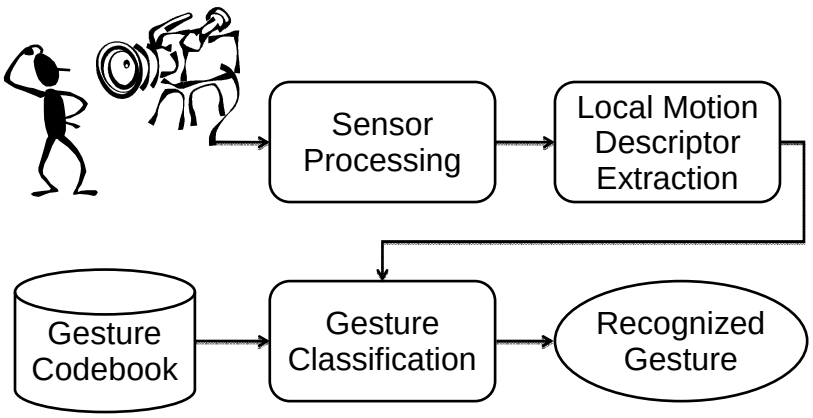
Action Recognition

Action Recognition (MB. Kaaniche)

Type of gestures and actions to recognize

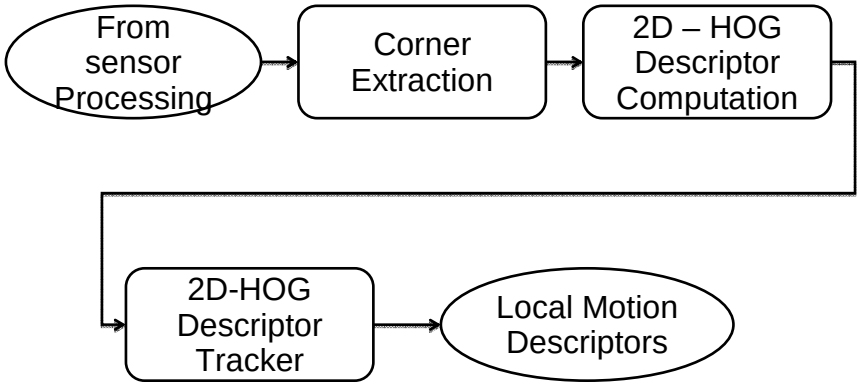
Action Recognition

- Method Overview



Action Recognition

- Local Motion Descriptor Extraction



Action Recognition

Local Motion Descriptor :

Let $[(x_1, y_1),...,(x_l, y_l)]^T$ the trajectory of a tracked HOG descriptor.

- The line trajectory is $[(w_1, h_1),...,(w_{l-1}, h_{l-1})]^T$ where:

$$\forall i \in [1..l-1]; w_i = x_{i+1} - x_i \wedge h_i = y_{i+1} - y_i$$

- The trajectory orientation vector is $[\theta_1,...,\theta_{l-2}]^T$ where:

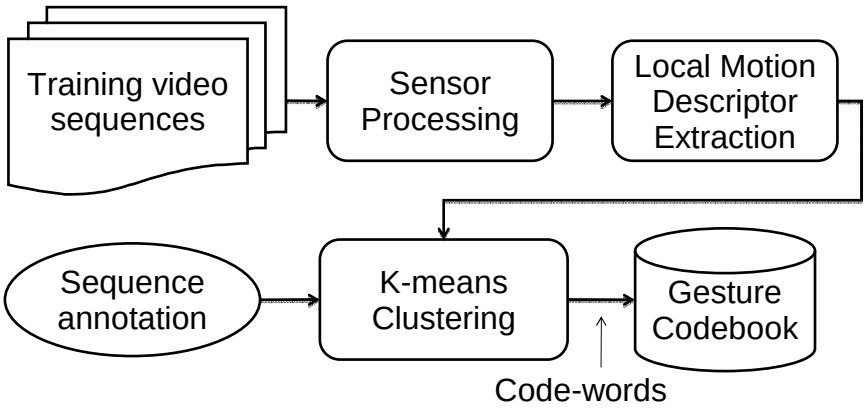
$$\forall i \in [1..l-2]; \theta_i = \arctan(h_{i+1}, w_{i+1}) - \arctan(h_i, w_i)$$

- The vector is normalized by dividing all its components by 2π.
- Using PCA, the vector is projected on the three principal axis.

$$lmd = \left[\hat{d} \quad \hat{\theta}_1 \quad \hat{\theta}_2 \quad \hat{\theta}_3\right]$$

Action Recognition

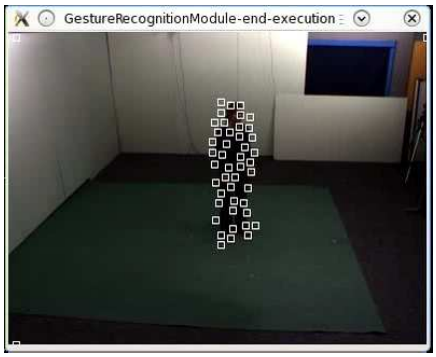
- Gesture Codebook Learning



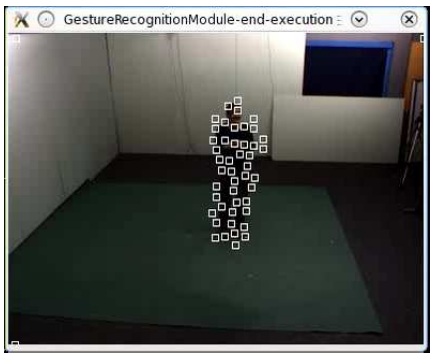
Gesture classifier based on motion descriptors

Approach :

- Track and Cluster KLT feature points using HOG descriptors.
- Extract gesture code-words for classification



Sit down



Kick

Multi sensor information fusion

Multi sensor information fusion

Three main rules for multi sensors information combination:

- Utilization of a 3D common scene representation (space, time and semantics) for combining heterogeneous information
- When the information is reliable the combination should be at the lowest level (signal): better precision
- When the information is uncertain or on heterogeneous objects, the combination should be at the highest level (semantics): prevent matching errors

Multiple Cameras Combination

Graphs Combination Approach:

Combine together all the mobile objects detected for two cameras using :

- a Combination Matrix computes correspondences between the Mobile objects detected for two cameras using a 3D position and a 3D size criteria.

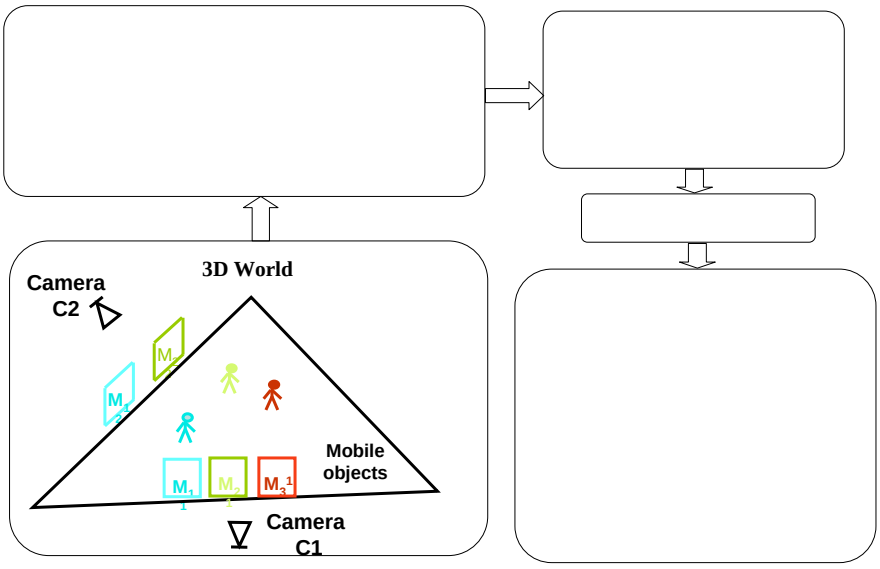
- a Set of rules: help to solve ambiguities

→ 3 types of combinations: Mobile objects from the two cameras can be either ***Fused***, ***Selected*** or ***Eliminated***

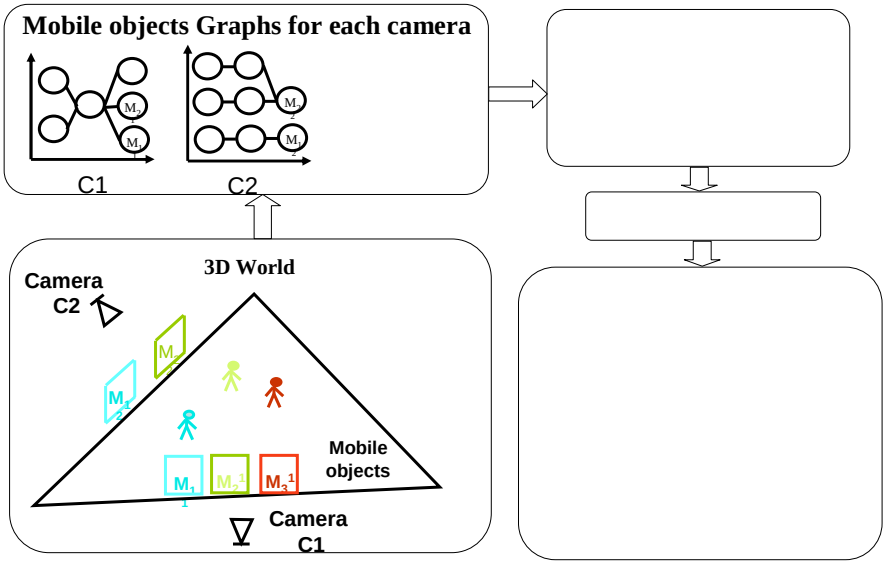
→ **The Combined Graph**

Mobile objects for N cameras are combined in an iterative way

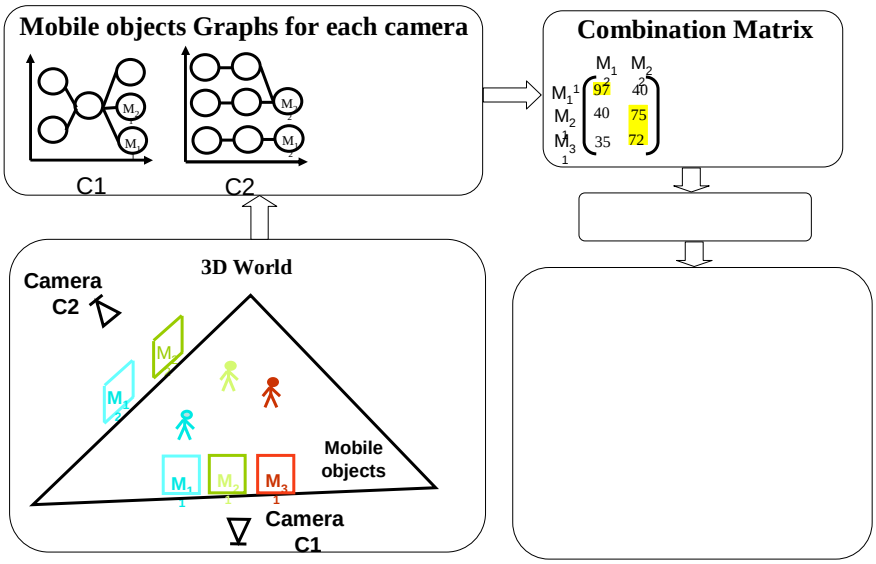
Multiple Cameras Combination



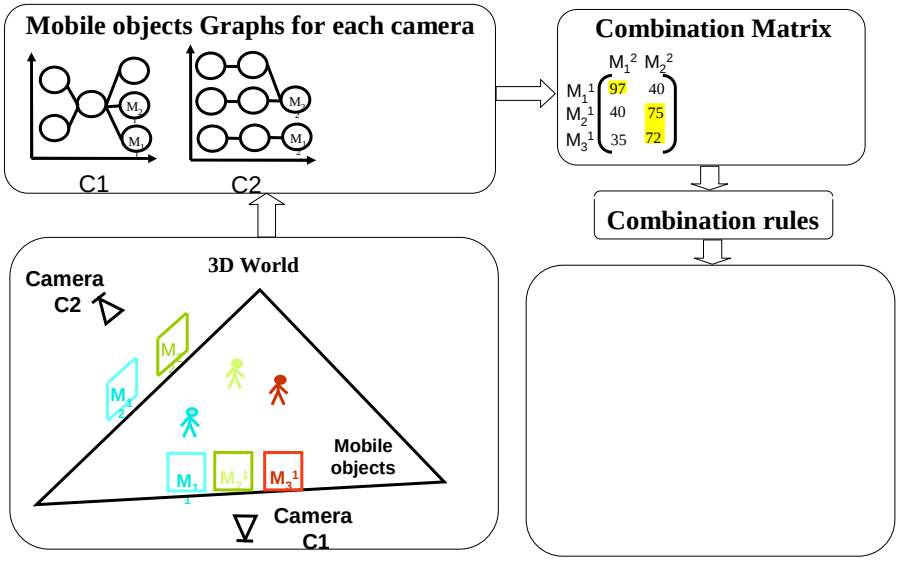
Multiple Cameras Combination



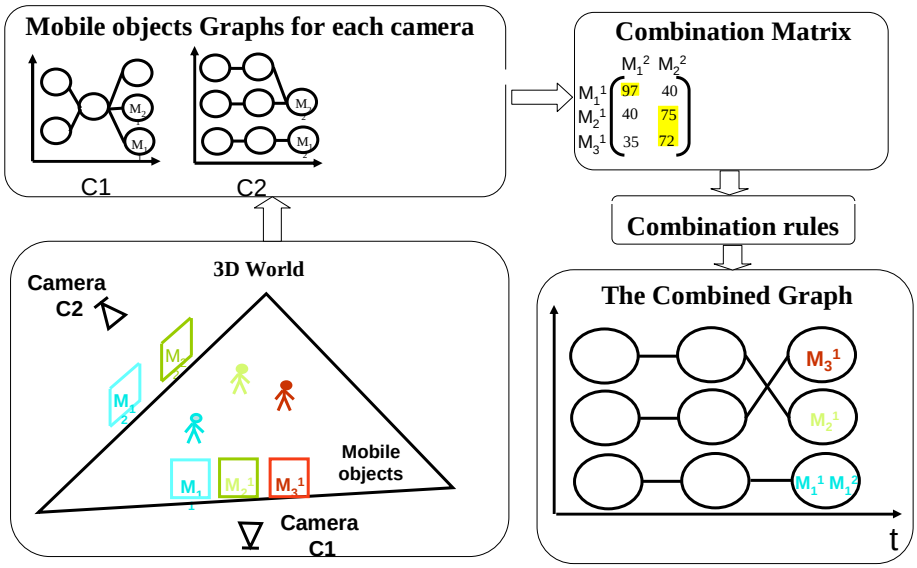
Multiple Cameras Combination



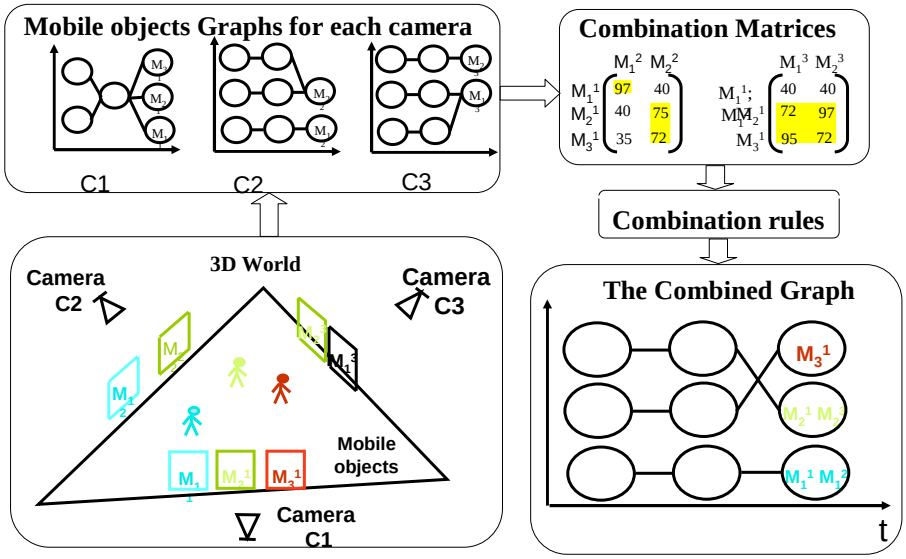
Multiple Cameras Combination



Multiple Cameras Combination

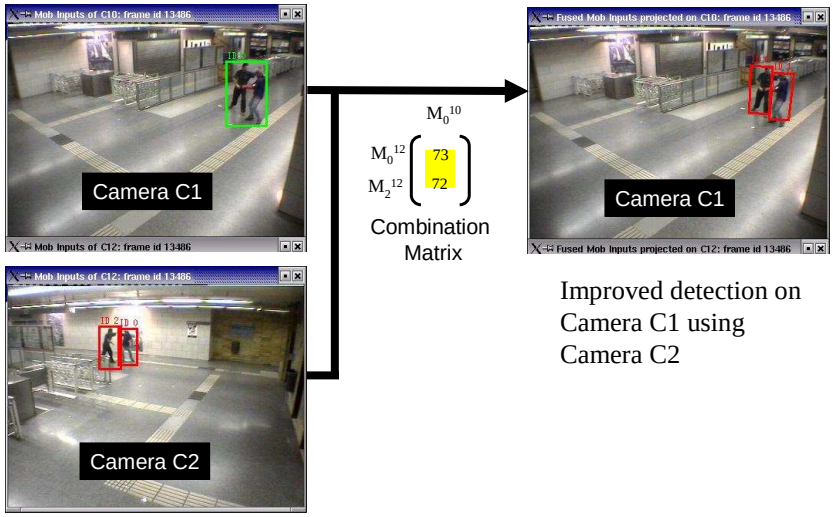


Multiple Cameras Combination



Multiple Cameras Combination

Example:



Improved detection on Camera C1 using Camera C2

Multiple Cameras Combination

Conclusion:

- Tested on 10 metro sequences with two cameras
- Globally allows to select the best camera

Limitations:

- Over estimation of the number of persons in some cases of ambiguities
  - Sensible to detection errors and camera positions
  - Work well in specific contexts small room (office..), few people

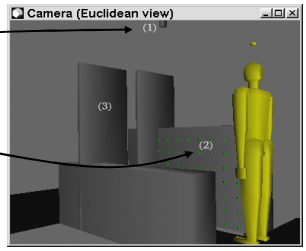
Multi sensors information fusion: Lateral Shape Recognition (B. Bui)

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 data.

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

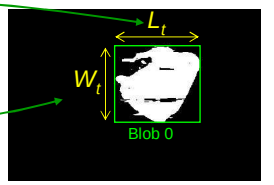
Lateral sensors (leds, 5 cameras, optical cells) on the side.

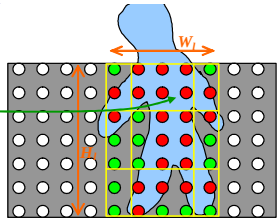


Lateral Shape Recognition: Mobile Object Model

Shape Model composed of 13 features:

- ✓ 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.



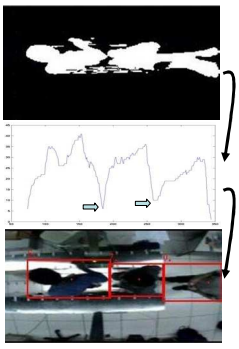


Model of a mobile object = (L_t, W_t, W_l, H_l, S_1,…, S_9) combine with a Bayesian formalism.

$$P(F \mid c) = \frac{P(c \mid F)P(F)}{P(c)}$$

Lateral Shape Recognition: Mobile Object Separation

Why ? To separate the moving regions that could correspond to several individuals people walking close to each other, person carrying a suitcase).
How ? Computation of pixels vertical projections and utilization of lateral sensors.

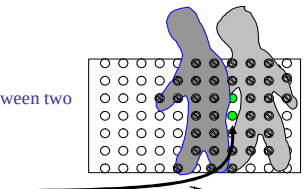


Computation of vertical projections of the moving region pixels

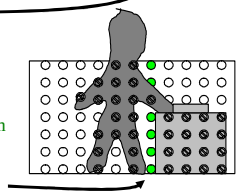
A separator is a "valley" between two "peaks"

Separation using vertical projections of pixels.

A non-occluded sensor between two bands of occluded sensors to separate two adults



A column of sensors having a large majority of non-occluded sensors enables to separate two consecutive suitcases and a suitcase or a child from the adult



Separation using lateral sensors

Lateral Shape Recognition: Experimental Results

•Recognition of “adult with child”

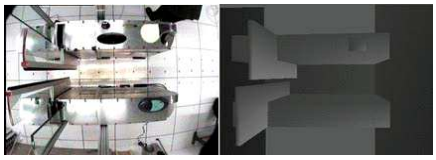
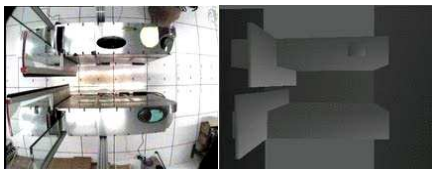


Image from the top camera

3D synthetic view of the scene

•Recognition of “two overlapping adults”



Lateral Shape Recognition: Experimental Results

• Recognition of “adult with suitcase”

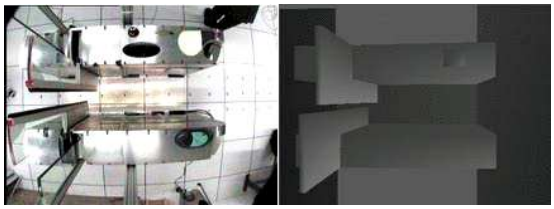


Image from the top camera

3D synthetic view of the scene

Video Understanding