Object Tracking at any Range of Scale for Video Surveillance

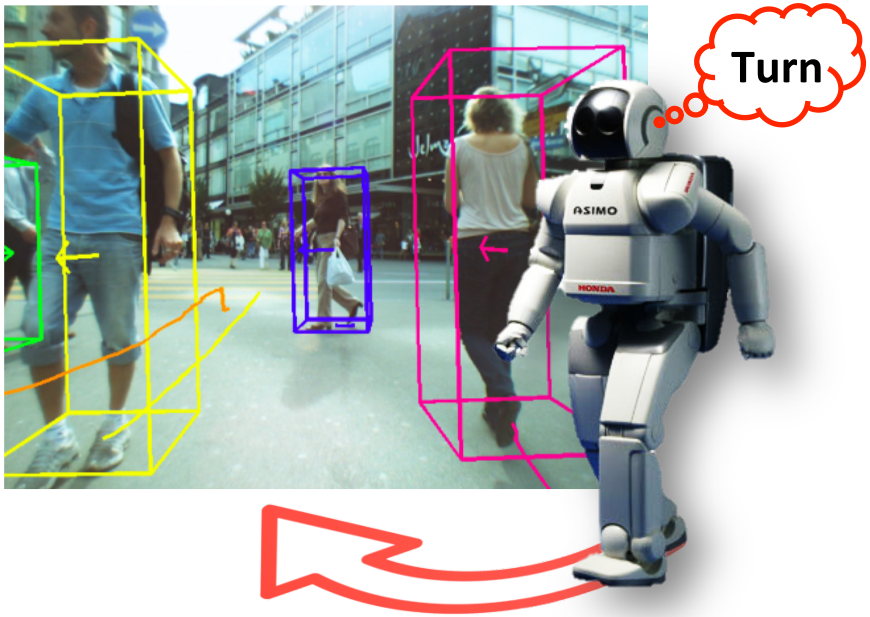
Alberto Del Bimbo
MICC Media Integration and Communication Center
Università di Firenze, Italy

delbimbo.alberto@unifi.it

Visual Object Tracking

- Visual Object Tracking is critical task in many applications like security and surveillance, urban sensing etc. Effective to build up complex systems performing image understanding

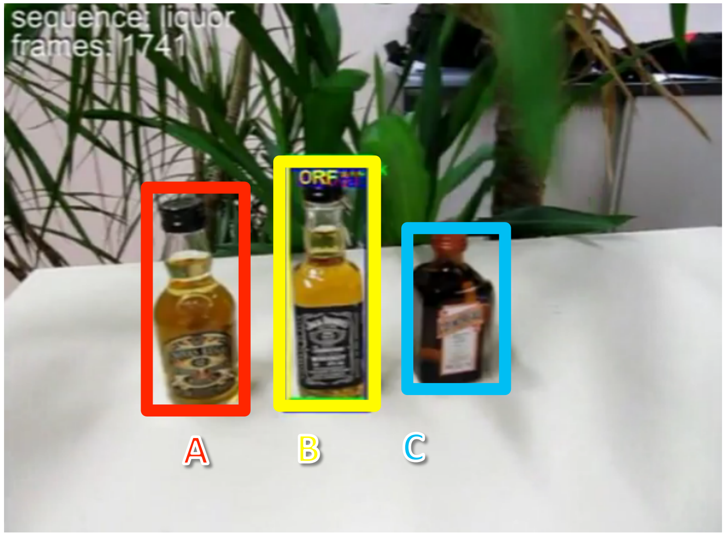
- Recognizing objects or their behaviors in a social environment



Courtesy of [Choi&Savarese2010]

- Provide decisive information to support autonomous systems (i.e. robots) to explore and interact with the 3D world

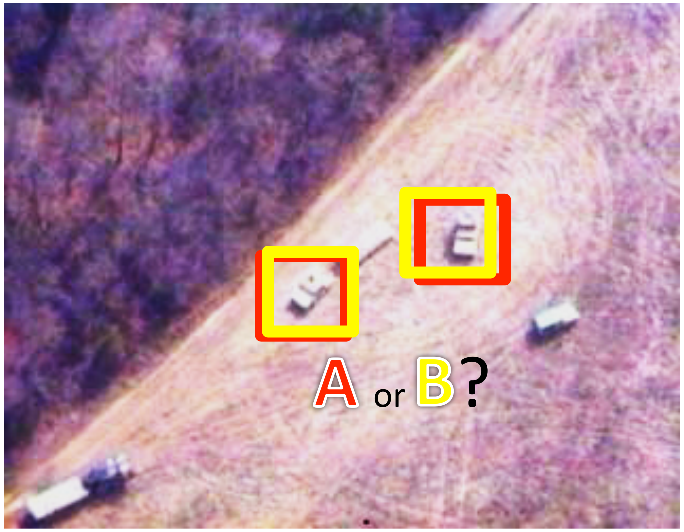
Two Distinct Lines of Approach…



- Performing target localization from adequate representation of the target shape and/or appearance (object detection)

- Labeled Localizations

- Good when camera motion has *reasonable* effects over targets appearance



- Performing target localization from target dynamics or registration of the moving sensor (visual odometry)

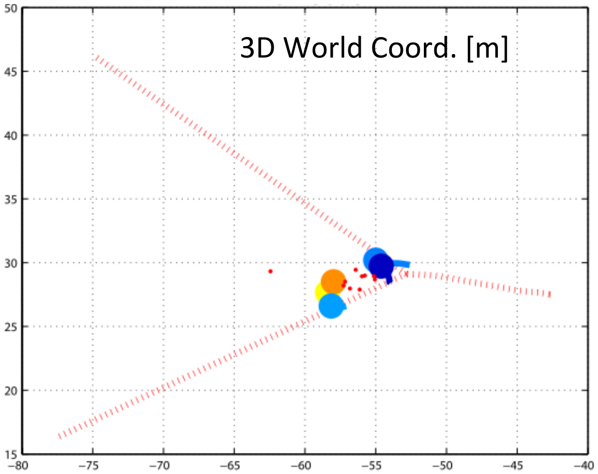
- Poorly labeled localization

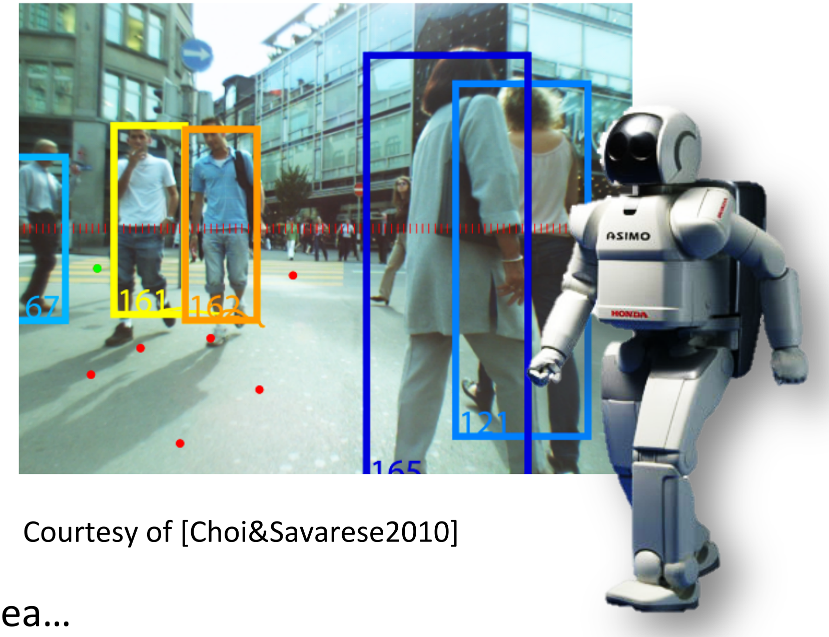
- Motion compensation allows separation between camera and target motion

Approaches combination

- The way these two broad categories are combined and weighted plays a decisive role in the robustness and efficiency of the tracker. Robustness and efficiency are application dependent

- Recent works [*Ess et al.* 2009], [*Wojeck et al.* 2009] [*Choi&Savarese* 2010] combined visual odometry and object detection to track multiple objects in 3D from a mobile robot. 3D trajectories and object image resolution simplify image understanding



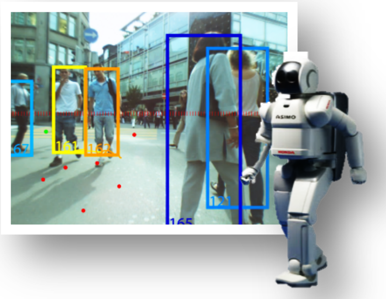


Courtesy of [Choi&Savarese2010]

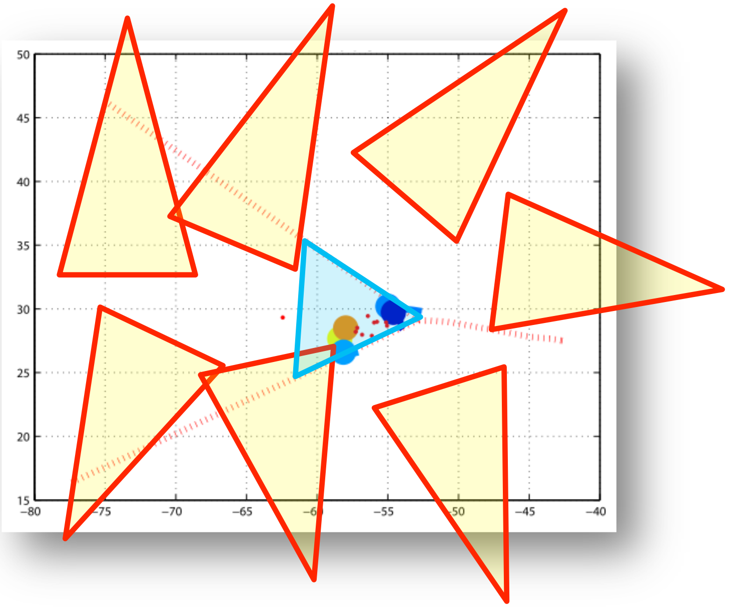
- However, more in general, considering a wide area…

Camera Sensor Coverage

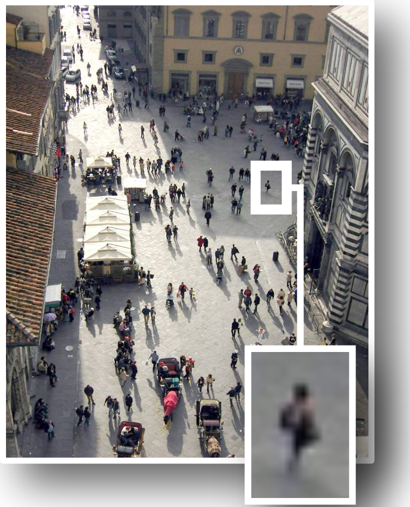
- Tradeoff between resolution and FOV



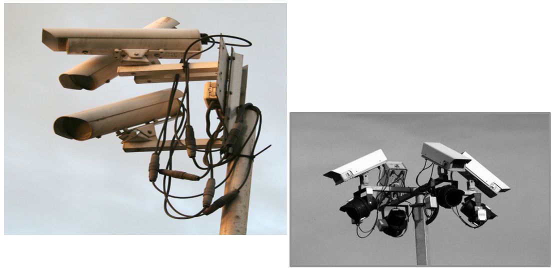
Short range



Rotating and Zooming Camera (PTZ)



Wide area

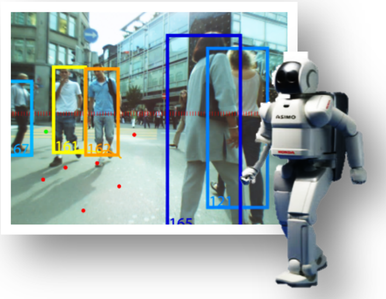


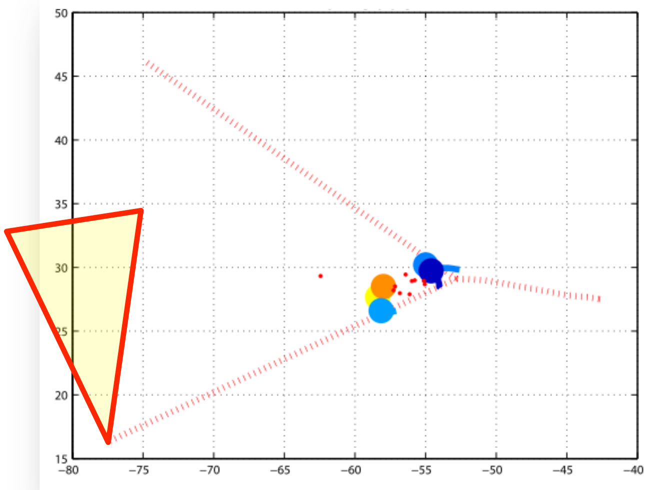
Free Moving Camera

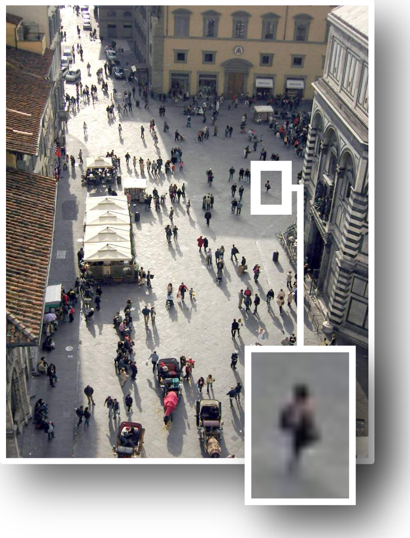
Multiple Stationary Cameras

Camera Sensor Coverage

- Fast switching between resolution and FoV







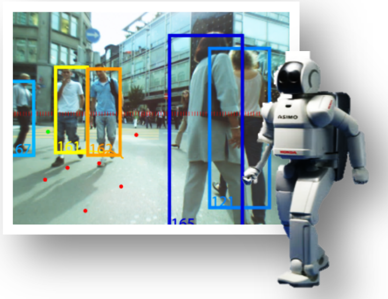
Rotating and Zooming Camera (PTZ)

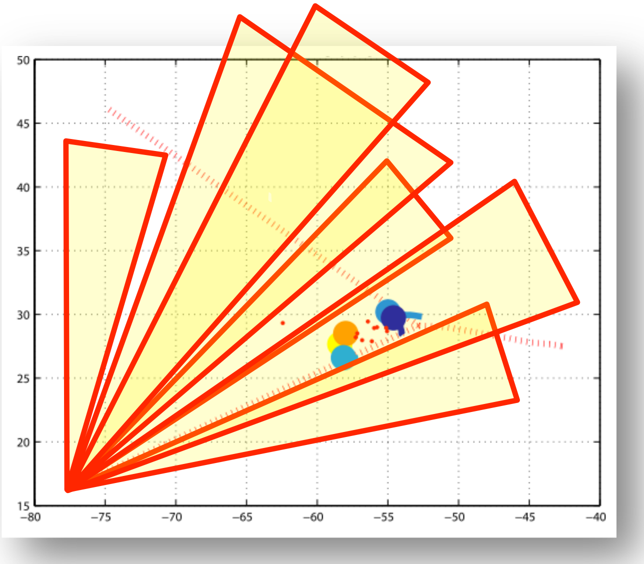
Free Moving camera

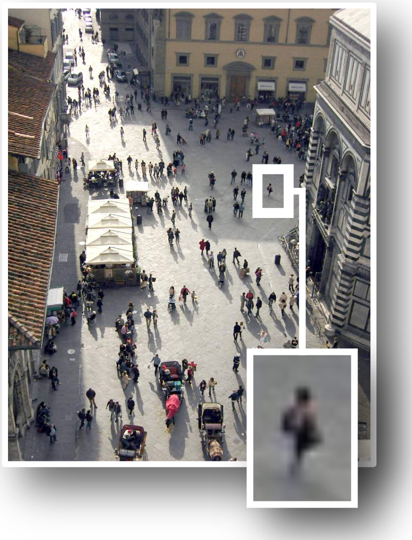
Multiple stationary cameras

Camera Sensor Coverage

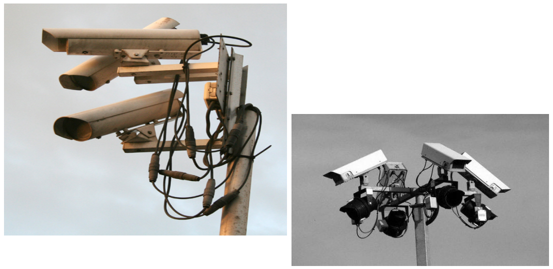
- Several static cameras can be replaced by a single PTZ camera







Rotating and Zooming Camera (PTZ)



Free Moving camera

Multiple stationary cameras

In this short course …
Objects Tracking at any Range of Scale for Video Surveillance

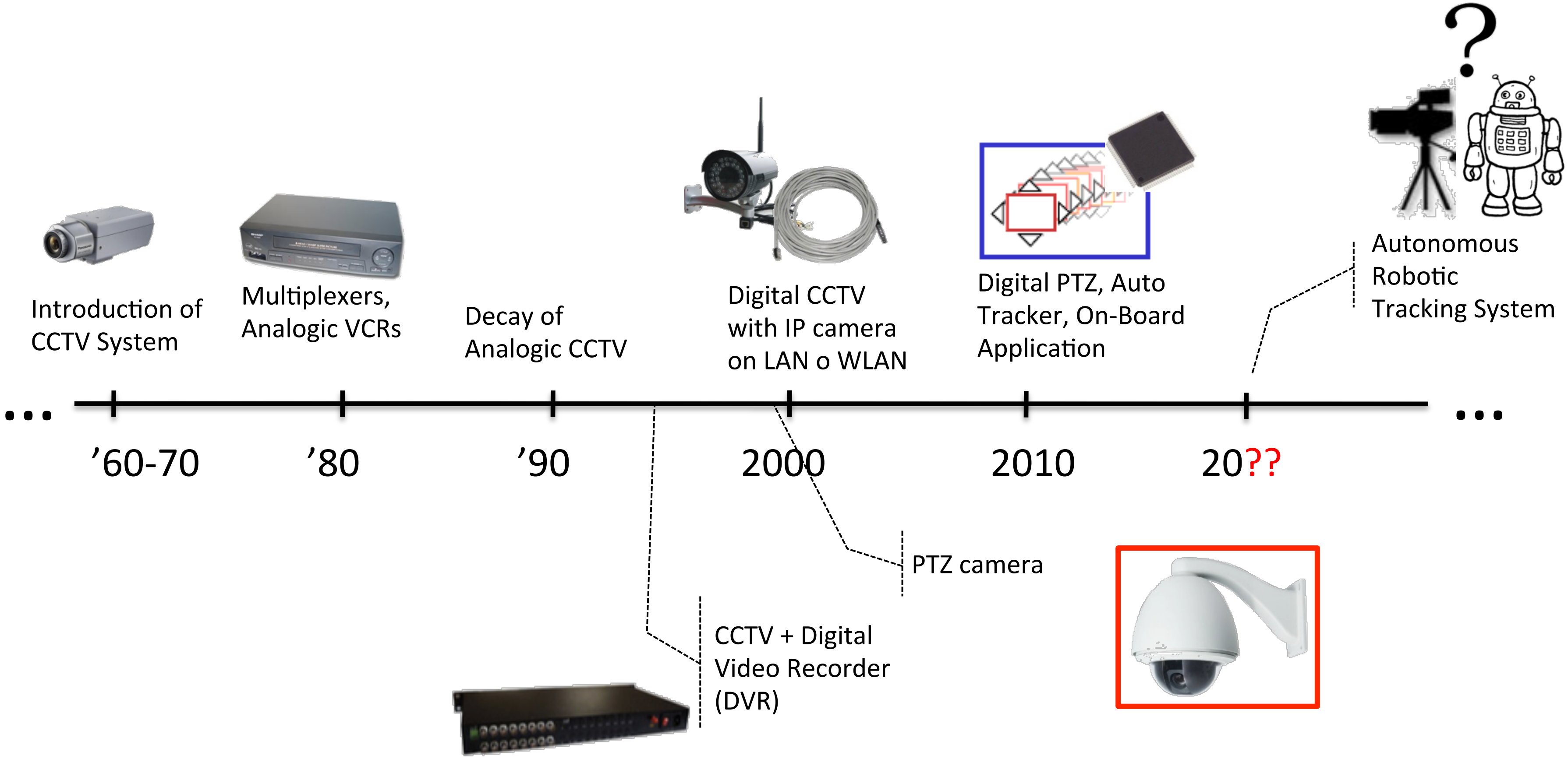
- Two subjects for discussion:
  - PART I Wide area monitoring with PTZ camera sensors, i.e. registration/alignement of the moving sensor and 3D object tracking (Alberto del Bimbo)
  - PART II Appearance based state of the art tracking system, with object recognition capability (Federico Pernici)

- Their combination can built next generation surveillance system in which not only large area can be monitored, but also recognition of specific subjects could be achieved

AdB

Wide Area Surveillance

Surveillance: Evolution of Cameras



Introduction of CCTV System

Multiplexers, Analogic VCRs

Decay of Analogic CCTV

Digital CCTV with IP camera on LAN o WLAN

Digital PTZ, Auto Tracker, On-Board Application

Autonomous Robotic Tracking System

… '60-70 '80 '90 2000 2010 20?? …

PTZ camera

CCTV + Digital Video Recorder (DVR)

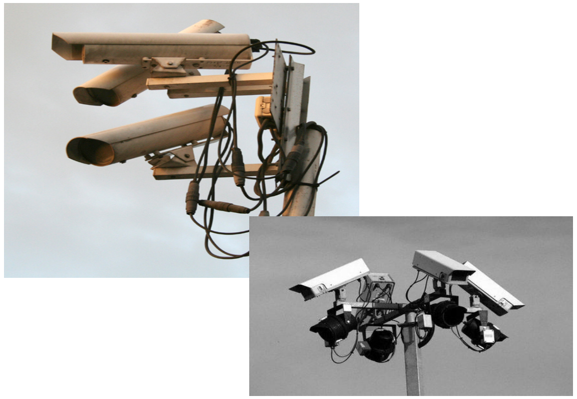
10x

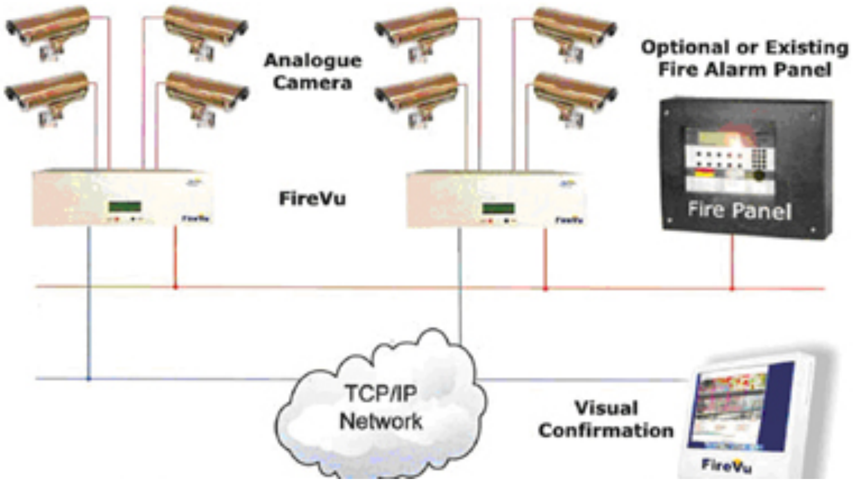
25x

	Pan Speed deg/sec	Tilt Speed deg/sec	Zoom Speed #mag/sec
Sony EVI-D30	80	50	0.6
Sony SNC-RZ30	170	76.6	8.3
Directed Perception	300	300	11.3

AdB

Wide Area Surveillance with Fixed Cameras





Benefits:

- Camera is fixed, so the calibration is performed only once
- This scenario is typically adopted for Closed Circuit Television (CCTV)
- Commercial video analytics software exists for this kind of scenario
- Standard low-level tracking technique can be applied, generally exploiting motion detection (background subtraction)
- …

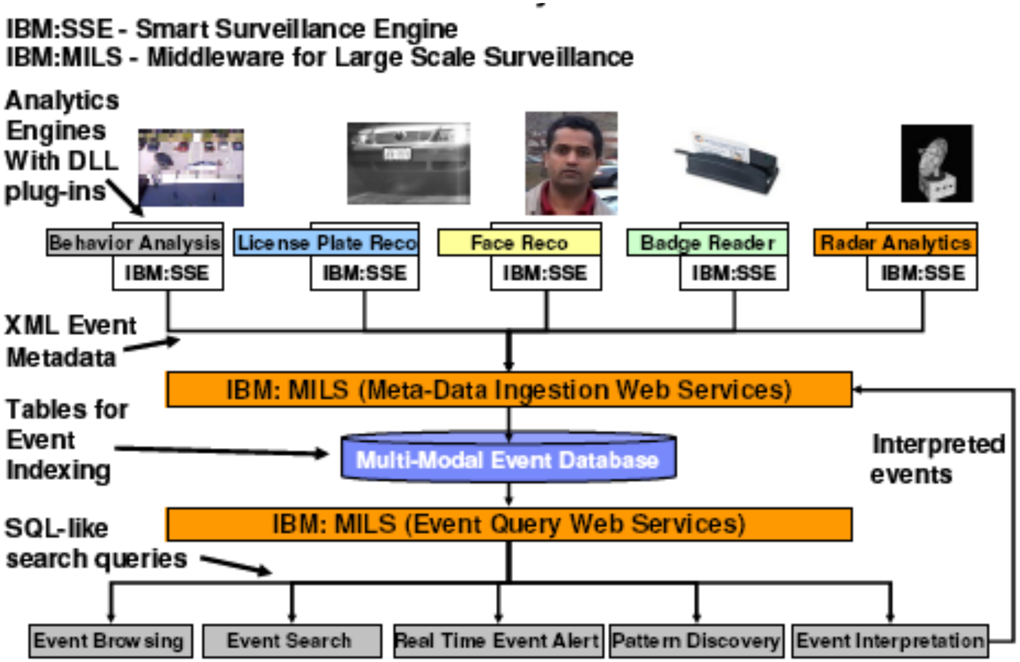
Issues:

- Installation considering a camera network, high cost for the setup and maintenance of the setup
- Once everything is fixed up, it cannot be easily modified
- Difficulties for the synchronization of videos of the distinct cameras
- Difficulties when investigators need high resolution images, f.e., face imagery of suspicious people
- High costs for management and monitoring of multiple video-streams
- ….

[Smart Surveillance System IBM Corp. USA, many others…]

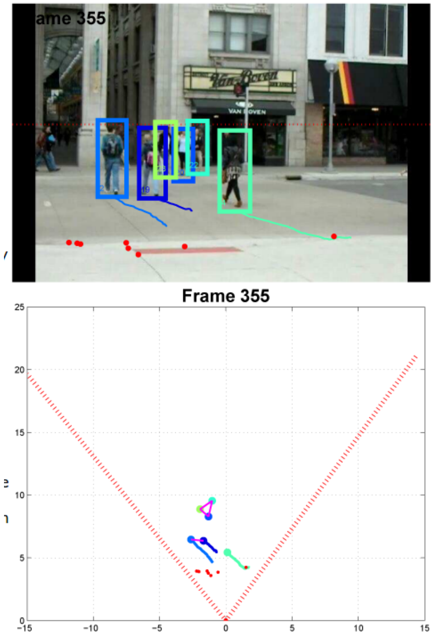
Classical surveillance systems architectures: IBM S3 Smart Surveillance System IBM Corp. USA

- The architecture of the IBM S3 system enables the use of multiple independently developed event analysis technologies. The following technologies are integrated:
  - *License Plate Recognition:* catalogs the license plate of each of the arriving and departing vehicles
  - *Behavior Analysis:* detects and tracks moving objects and classifies them into a number of predefined categories
  - *Face Detection/ Recognition:* captures and recognizes faces
  - *Badge Reading:* reads subject’s identity from badge


IBM:SSE - Smart Surveillance Engine
IBM:MILS - Middleware for Large Scale Surveillance
Analytics Engines With DLL plug-ins
Behavior Analysis
IBM:SSE
License Plate Reco
IBM:SSE
Face Reco
IBM:SSE
Badge Reader
IBM:SSE
Radar Analytics
IBM:SSE
XML Event Metadata
IBM: MILS (Meta-Data Ingestion Web Services)
Tables for Event Indexing
Multi-Modal Event Database
Interpreted events
SQL-like search queries
IBM: MILS (Event Query Web Services)
Event Browsing
Event Search
Real Time Event Alert
Pattern Discovery
Event Interpretation

AdB

Wide Area Surveillance from a Mobile Platform



Benefits:

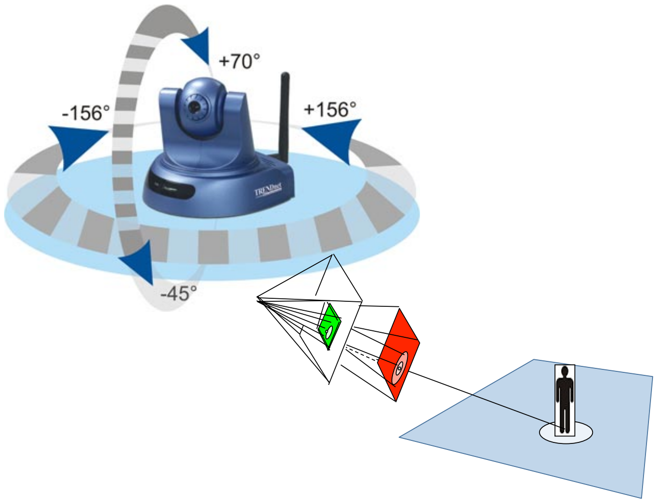
- Sensor can navigate and translate
- Very wide path coverage
- One large path can be covered from just one moving platform
- No possibility to zoom in
- Less cost for the setup, installation
- At least one stream per moving platform

Recent Work by: [*Ess* 2009, *Choi and Savarese* 2010 …]

Issues:

- Some non trivial computer vision issues to address:
  - Motion compensation
  - Use of stereo information….
- Standard techniques like background subtraction cannot be used
- Must process the stream in the real-time in order to feedback the moving platform
- Limited field of view, depth and speed do not allow to promptly cover a large area at the necessary image resolution
- Short fragmented 3D trajectories complicate scene understanding

Wide Area Surveillance with a PTZ camera



Benefits:

- Sensor can rotate and zoom
- Wide field of view
- Wide area can be covered from just one PTZ camera
- Possibility to quickly zoom-in and out
- High resolution images
- Less cost for the setup, installation
- At least one stream per wide area

Issues:

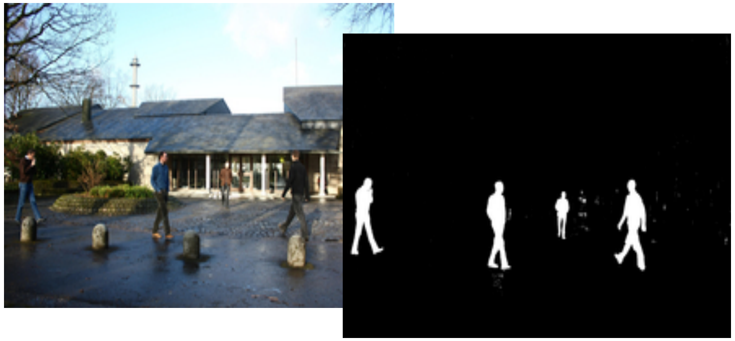
- Some non trivial computer vision issues to address:
  - Motion Compensation
  - Camera is just rotating: basically you cannot “look behind” an object…
- The camera is continuously moving so no fixed background is available
- Large scale variations of targets due to zooming or the distance between the target and the camera
- Due to pan, tilt and zoom camera pose is time-varying
- Scalability with respect to the number of targets
- ….
- All these issues must be managed in real-time

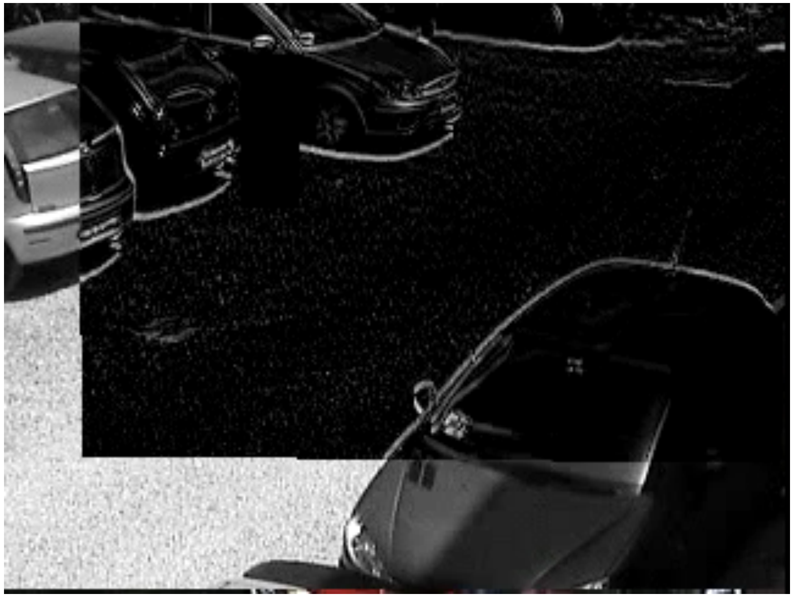
Recent Work by: [*Okuma* 2004, *Cai* 2006, *DelBimbo* 2009, *Breinstesten* 2010 ..]

AdB

PTZ Issues 1/5

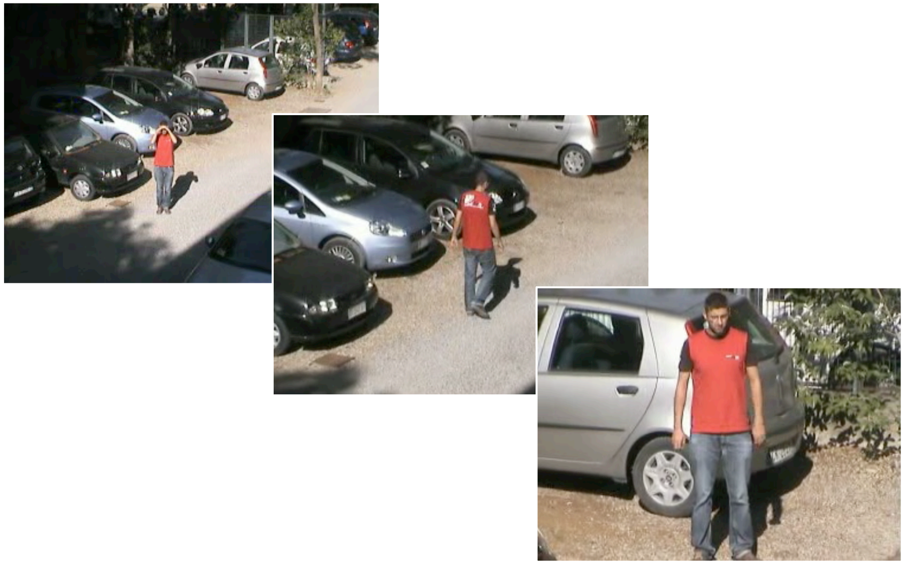
- *Moving camera sensor*: the camera is continuously moving, to capture high quality detail of targets in the scene, no fixed background will be available
- Well known target extraction methods based on change detection like background subtraction cannot work properly





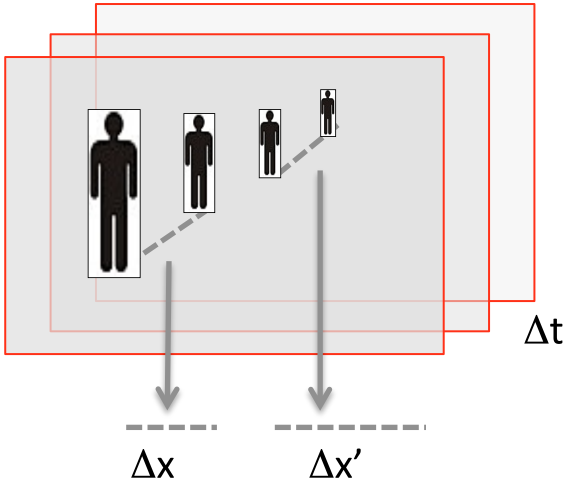
PTZ Issues 2/5

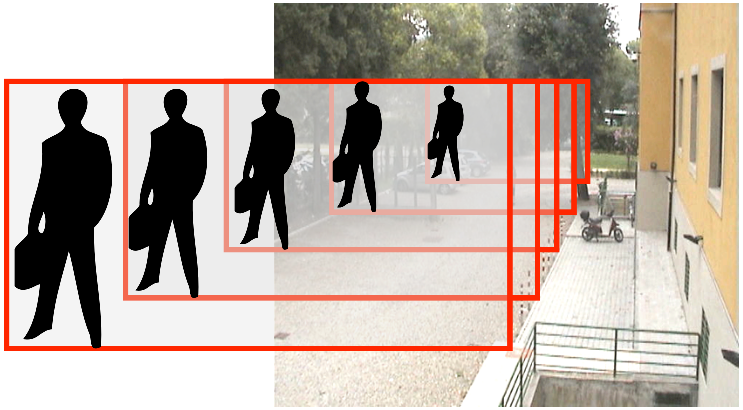
- *Large scale changes of targets*: targets moving in a wide area may undergo large scale changes and consequent appearance changes due to pan, tilt and zoom operations
- Estimation of internal and external camera parameters is challenging since when zooming in new visual structure is introduced in the image that can be difficult to match to a fixed resolution template



PTZ Issues 3/5

- *The distance between the moving target and the camera*: with PTZ cameras, tracking in the 2D image plane does not allow accurate prediction due to the fact that the same pixel displacements between far targets and near targets correspond to different distances in the real world. Near targets move faster than far targets
- Accurate estimation should be made exploiting 3D information

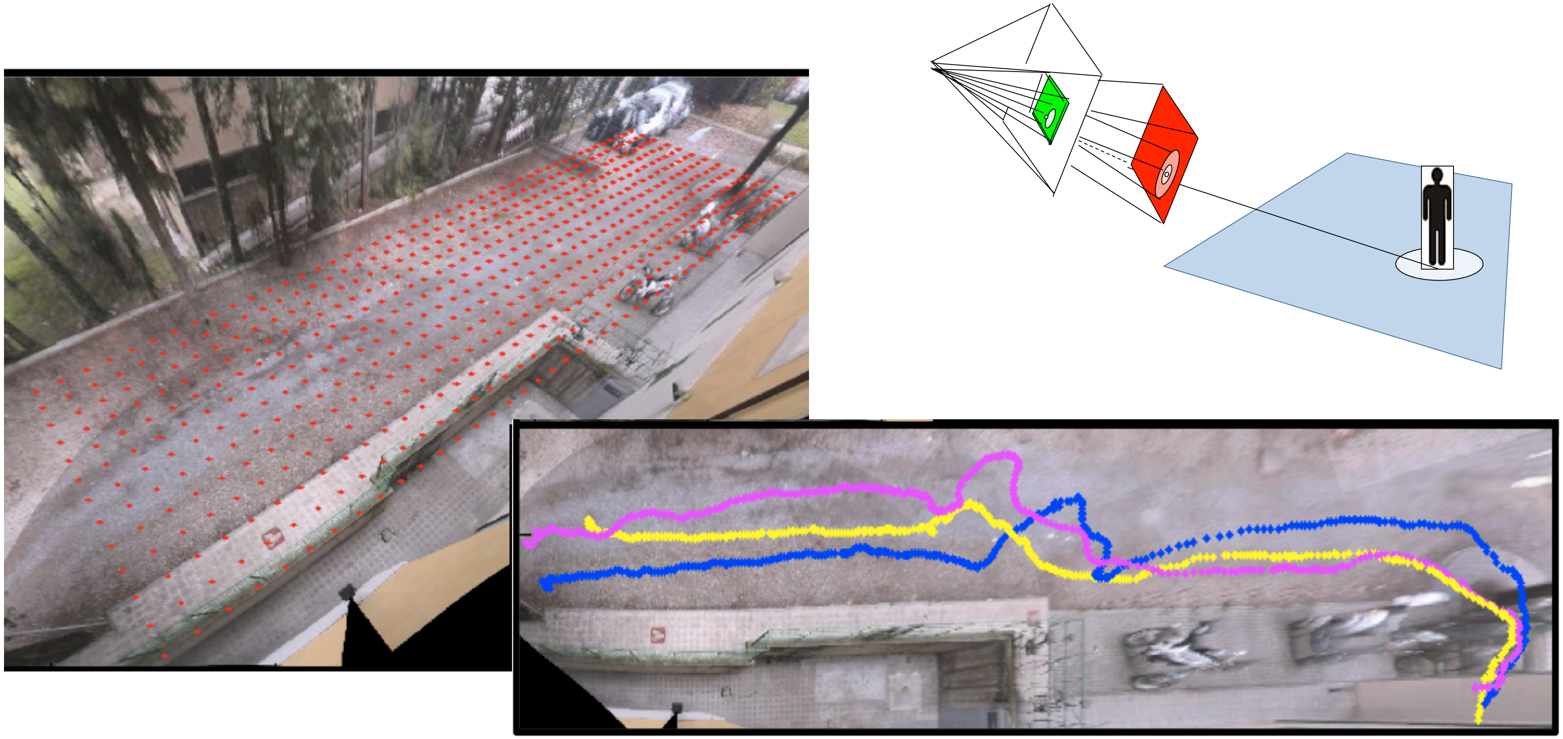




AdB

PTZ Issues 4/5

- *Time-varying camera pose*: estimation of the camera pose is necessary to estimate the precise position in 3D of the tracked target, its speed, trajectory…
- Due to pan, tilt and zoom operations, the pose between the camera view and the 3D observed scene is time-varying, so on-line camera calibration is necessary



PTZ Issues 5/5

- *Scalability with respect to the number of targets*: generally in a wide area, both indoor and outdoor, there will be a large number of people
- Multiple target tracking in real-time requires data association with sensor management

In the following…

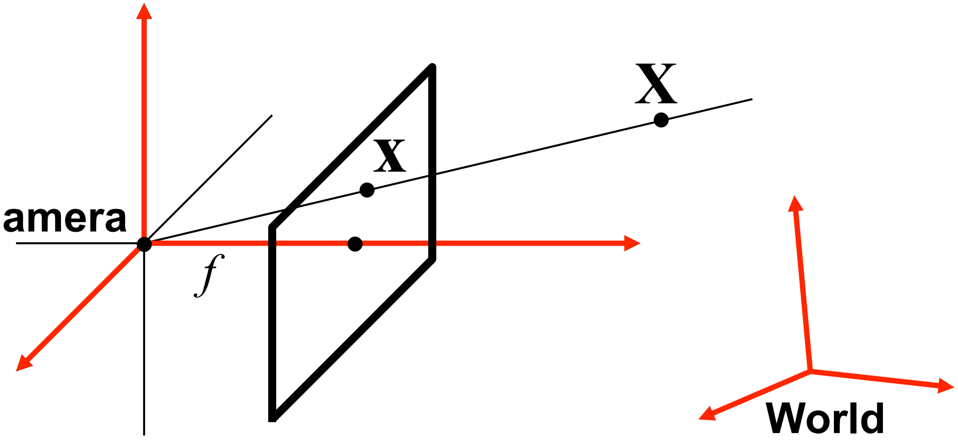
- We will discuss:
  - Online camera pose estimation
  - Target detection at any scale exploiting 3D information
  - 3D target position tracking

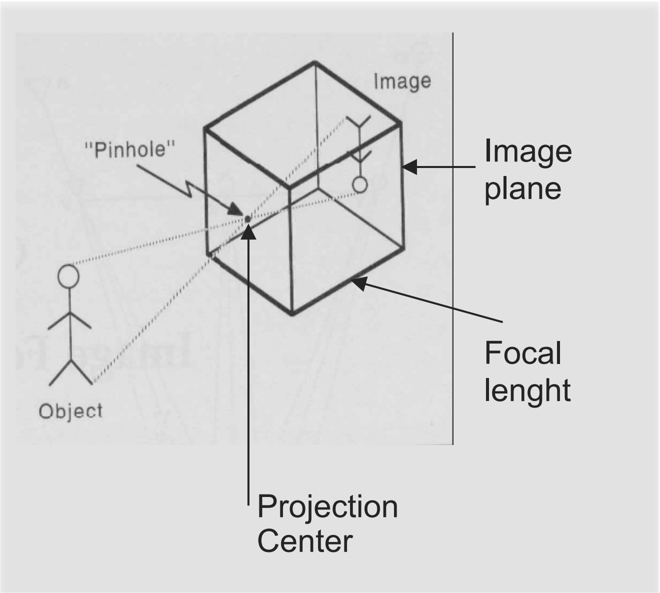
- Camera pose estimation permits to obtain an absolute reference for PTZ tracking and to compute the relationship between target positions in the 2D image and positions in the 3D world plane
- From the 3D camera pose it is possible to infer the expected imaged height of the target at any image location

Camera Pose Estimation

AdB

Pinhole Camera

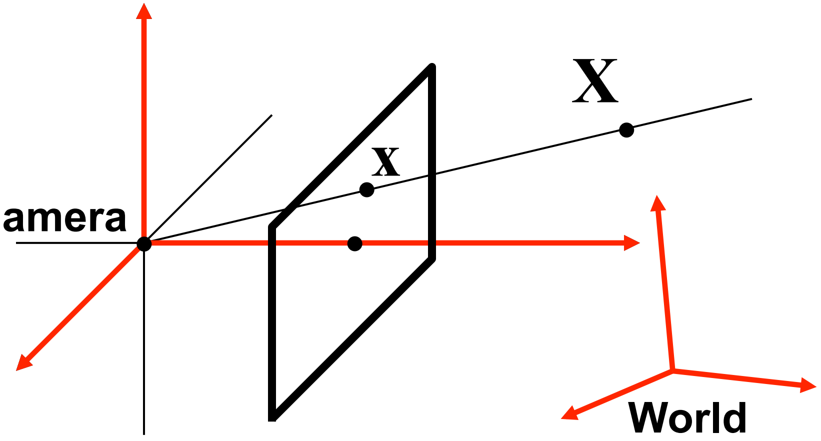




$$\begin{pmatrix}x\\y\end{pmatrix} = \frac{f}{Z}\begin{pmatrix}X\\Y\end{pmatrix}$$

- Virtual (front) image plane provides a theoretical pinhole camera which may be simpler to analyse than the real one

Camera Projection Matrix



In homogeneous coordinates

$$\mathbf{x} = \begin{bmatrix} x & y & 1 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} X & Y & Z & 1 \end{bmatrix}$$

$$\mathbf{x} = \mathbf{P}\mathbf{X} \qquad \mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix}$$

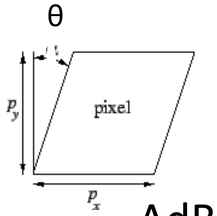
- In the general case the camera projection matrix contains both internal camera parameters (focal lenght, camera center, skew, aspect ratio, radial distortion) and external camera parameters (rotations and translation)

- **P** can be factorized as: $\mathbf{P} = \mathbf{K}\begin{bmatrix} \mathbf{R} \mid \mathbf{t} \end{bmatrix}$

R is a 3x3 rotation, and **t** is a three-dimensional translation of a generic transformation between the image and world coordinate systems

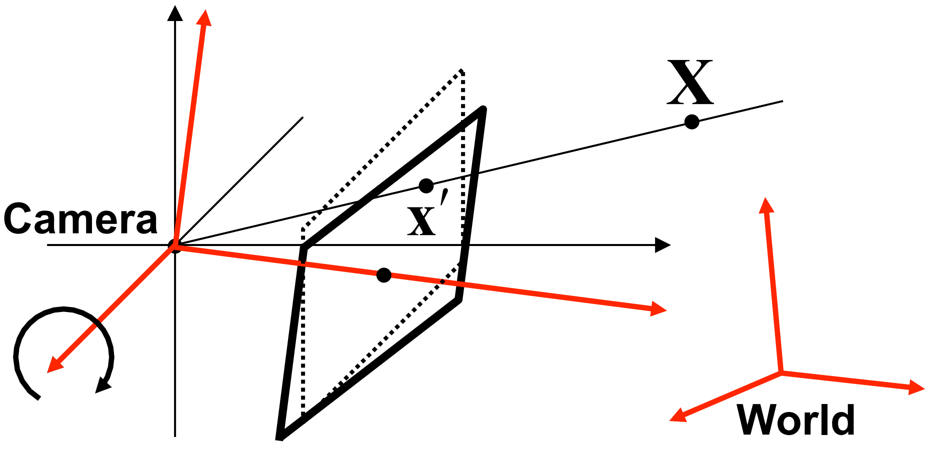
K is the camera calibration matrix that encodes the transformation from image coordinates to pixel coordinates in the image plane

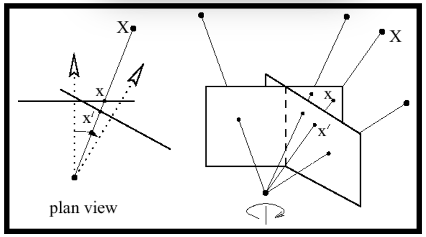
$$K = \begin{bmatrix} f/s_x & f/s_x \cot\theta & o_x \\ 0 & f/s_y & o_y \\ 0 & 0 & 1 \end{bmatrix}$$



AdB

Projection Matrix with Rotating Cameras



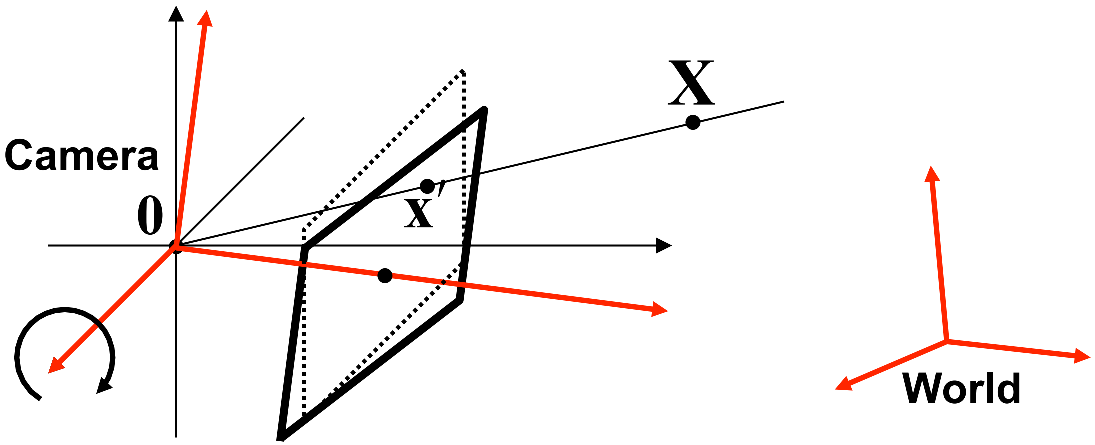


- For a camera rotating around a fixed axis each i-th grabbed image image has its own $\mathbf{P}_i$ where the Translation vector is **0**

$$\mathbf{P}_i = \mathbf{K}_i \left[\mathbf{R}_i \mid \mathbf{0} \right]$$

- For a camera with fixed optics, parameters are identical for all the images taken with the camera. For cameras which have zooming and focusing capabilities the focal length can obviously change, but also the principal point can vary
- For this case calibration is estimating the $\mathbf{K}_i$ and the $\mathbf{R}_i$ ....

Reduced Projection Matrix

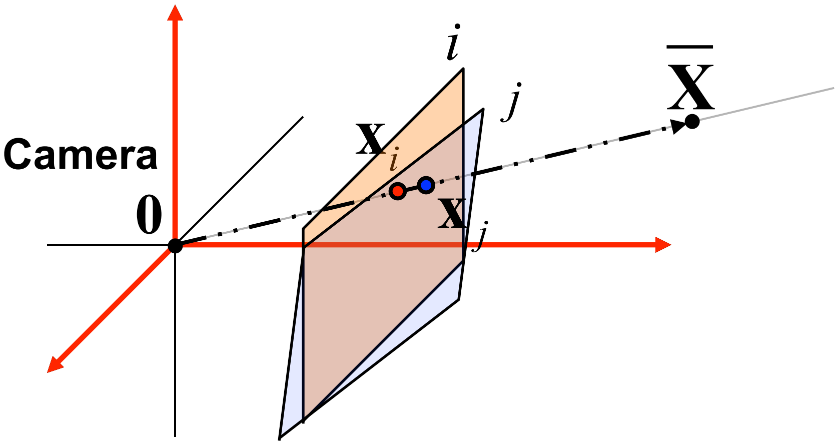


- For each generic image point $\mathbf{x}$ and real world 3D point $\mathbf{X}$

$$\mathbf{x} = \begin{bmatrix} x & y & 1 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} X & Y & Z & 1 \end{bmatrix}$$

$$\mathbf{x} = \mathbf{K}_i \left[\mathbf{R}_i \mid \mathbf{0} \right] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \mathbf{K}_i \mathbf{R}_i \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{K}_i \mathbf{R}_i \overline{\mathbf{X}} \qquad \overline{\mathbf{P}_\mathbf{i}} = \mathbf{K}_i \mathbf{R}_i$$

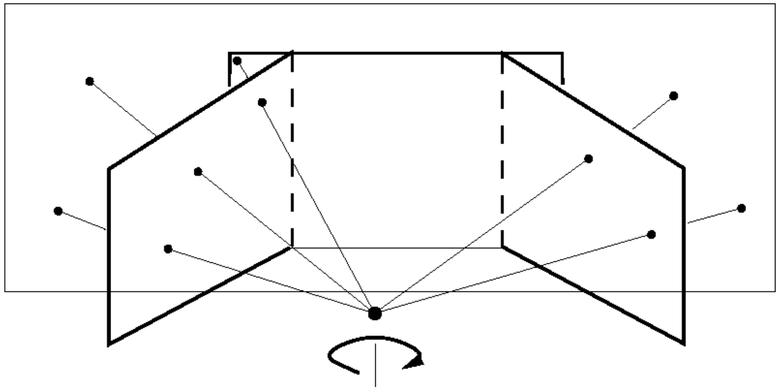
Inter image homography



For each pair of overlapped images, corresponding image points $\mathbf{x}_i$ and $\mathbf{x}_j$ of a generic 3D point $\mathbf{X}$:

$$\mathbf{x}_i = \mathbf{K}_i \mathbf{R}_i \overline{\mathbf{X}}$$

$$\mathbf{x}_j = \mathbf{K}_j \mathbf{R}_j \overline{\mathbf{X}}$$



- Eliminatating $\overline{\mathbf{X}}$ from the equations:

$$\mathbf{x}_j = \mathbf{H}_{ij} \mathbf{x}_i$$

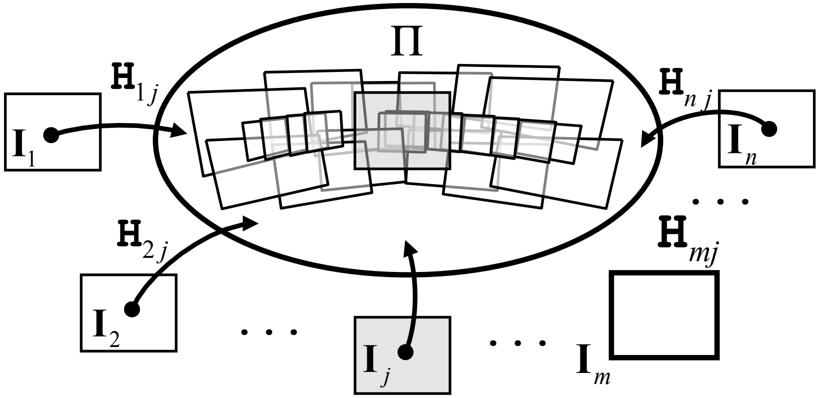
where $\mathbf{H}$ is a 3x3 inter-image homography that describes the transformation from one image plane to another

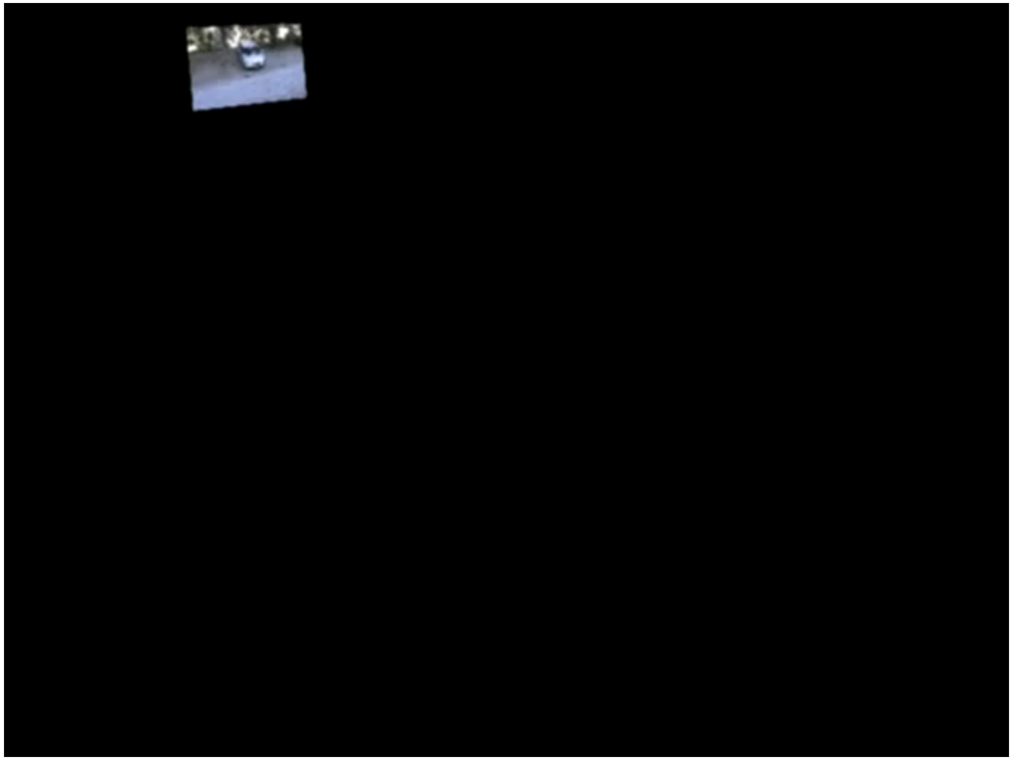
- Homography is defined by 8 DoF (5 internal 3 external): 4 pairs of points needed.

- The analytic expression in the metric and Euclidean case is: $\mathbf{H}_{ij} = \mathbf{K}_j \mathbf{R}_j \mathbf{R}_i^{-1} \mathbf{K}_i^{-1} = \mathbf{K}_j \mathbf{R}_{ij} \mathbf{K}_i^{-1}$

- Having a reference image of a scene, and taking a sequence of frames from a rotating camera each the homography between each image and the reference view contains the camera parameters of the current camera pose
- Estimates of inter-image homographies can be used to stitch a mosaic of the scene

Mosaic plane





Absolute Conic

- One of the most important concepts for self-calibration is the Absolute Conic (AC) and its projection in the images (IAC). Since it is invariant under Euclidean transformations. Its relative position to a moving camera is constant. For constant intrinsic camera parameters its image will therefore also be constant

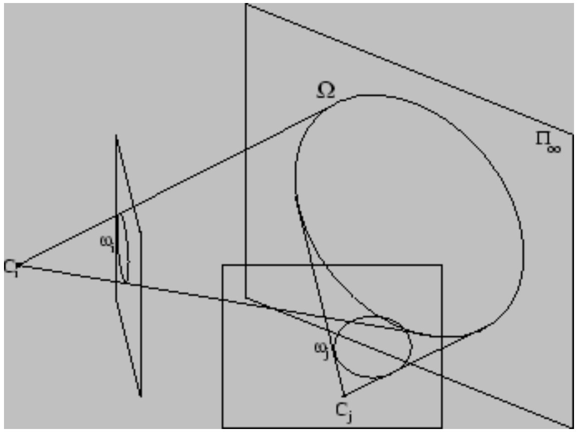
- For a Euclidean representation of the world the camera projection matrices can be factorized as: $\mathbf{P}_i = \mathbf{K}_i\mathbf{R}_i^\top[\mathbf{I}|\text{-}\mathbf{t}_i]$

  with $\mathbf{K}$ an upper triangular matrix containing the intrinsic camera parameters, $\mathbf{R}$ representing the orientation and **t** the position and the Dual Absolute Quadric can be written as Ω^* diag (1,1,1,0)

  It results that ω^* representing the dual of the Image Absolute Conic is:

$$\omega_i^* \sim \mathbf{P}_i\Omega^*\mathbf{P}_i^\top \quad . \qquad \omega_i^* \sim \mathbf{K}_i\mathbf{K}_i^\top$$

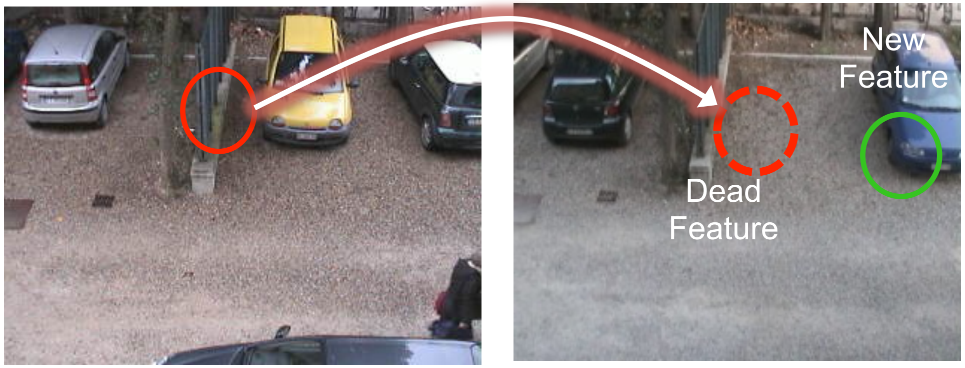
Absolute Conic (AC) and its projection in the images (IAC)



Relationships with Visual SLAM

- SLAM is a process by which a mobile robot can build a map of an environment and at the same time use this map to deduce its location

- Standard formulation is fragile to incorrect association of observations to landmarks. The scene can change drastically. The same happens with PTZ cameras in operation



- Standard Visual SLAM
  - 6DOF (internal calibrated camera) good linear properties
- PTZ Visual SLAM
  - 8 DOF (5 internal 3 external) higly non linear
  - 3DOF (focal length, pan and tilt) can be assumed in practice

PTZ Camera Pose Estimation
Related Work and State of the Art

AdB

Related Work: PTZ Camera Pose Estimation

Drift-free real-time sequential mosaicing [*Civera* et al. IJCV2009]

- Performs real-time sequential mosaicing of a scene observed by a rotating sensor. Based on EKF-SLAM for localization of the pose of the camera
  - Only rotating camera, without taking into account for zoom variation
  - Kalman state contains the 3 DoF of the camera pose and the coordinates of image landmarks
  - Not suited for zooming: zooming is non-linear and near-ambiguities may arise especially at large focal length

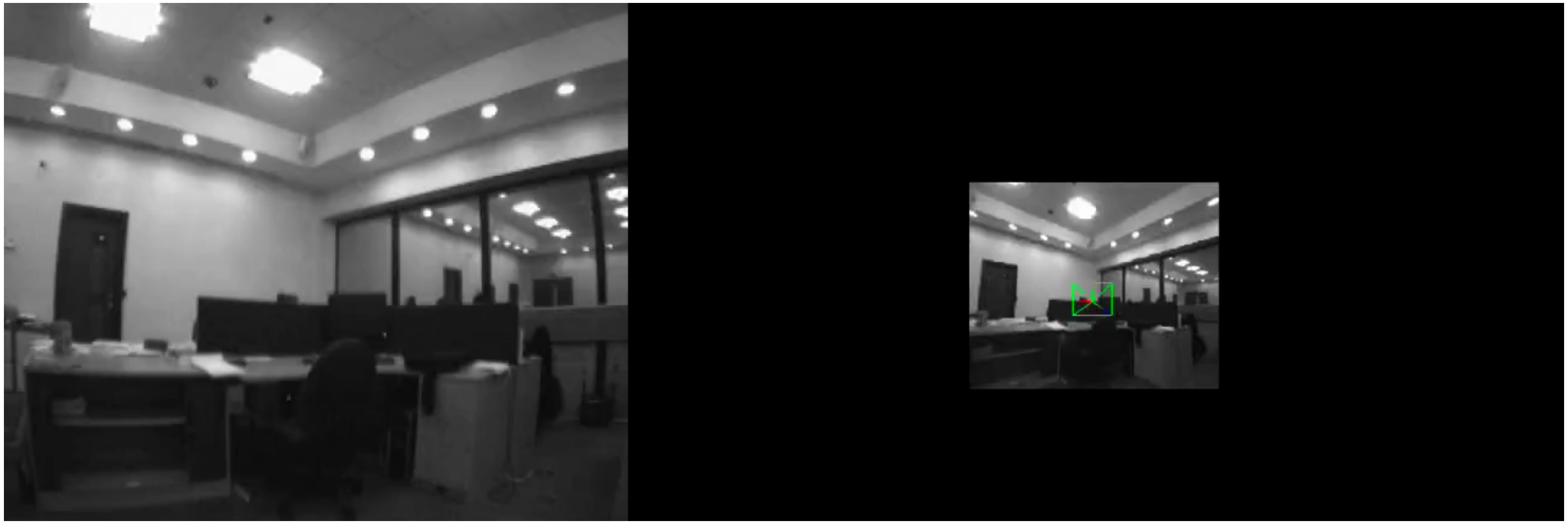
- Good accuracy also with moving people and moderate variations of illumination
- Scales badly with the number of features in the map state, and has very poor accuracy when the number of landmarks grows beyond a few hundreds
- Does not manage abrupt motion as the motion model only considers smooth rotational motion

Related Work: PTZ Camera Pose Estimation

Use of keyframes instead of sequential filtering [*Lovegrove* et al. ECCV 2010]

- Camera parameters between consecutive images of a PTZ camera sequence obtained by pixel-based global image alignment
  - Variable focal length
  - Does not use local features

- Global image alignment does not provide robustness in the presence of moving objects (there are no moving targets in the sequence)

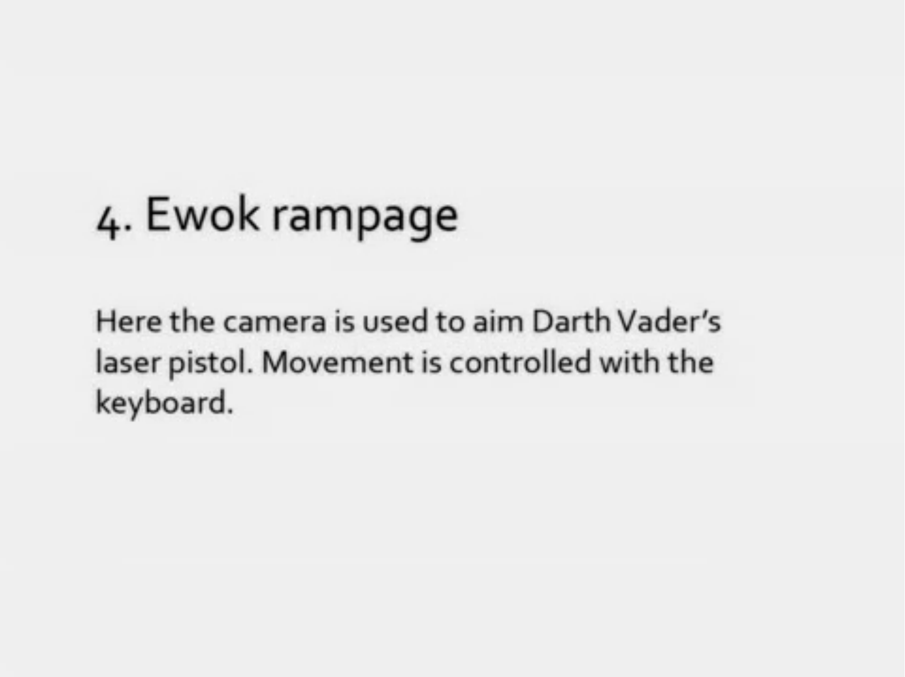


Related Work: Moving Camera Pose Estimation

Parallel Tracking and Mapping [*Kleyn* and *Murray* Int Symp on Mix Aug Reality 2009],
[*Nistér* et al. PCV 2006]

- Performs real-time camera pose estimation exploiting keyframes and online Bundle Adjustment: optimizes camera pose based on all the observed landmarks from the last 20 frames

- Geometry is for handheld cameras, not for rotating and zooming cameras
- Only for small workspaces, not suited for Wide Areas



4. Ewok rampage

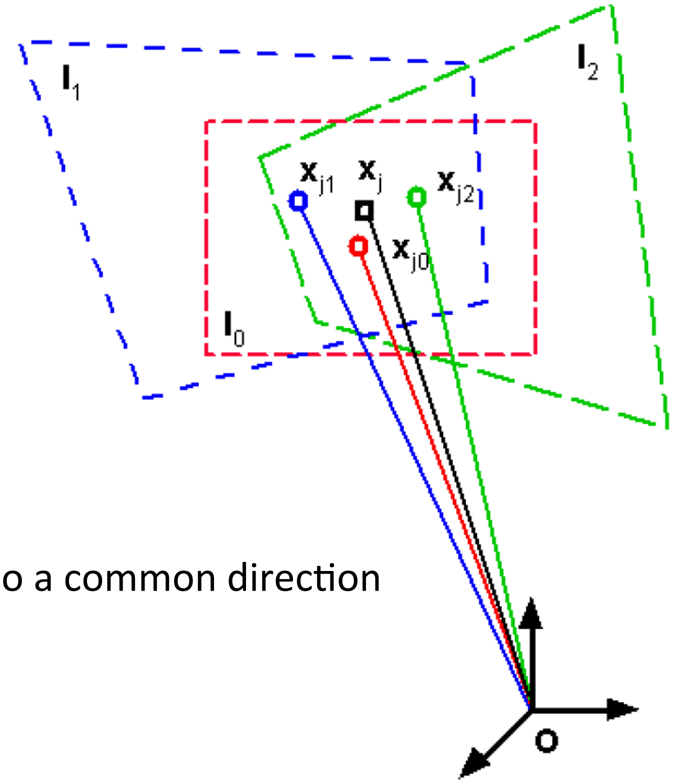
Here the camera is used to aim Darth Vader's laser pistol. Movement is controlled with the keyboard.

Software available with license at
http://www.robots.ox.ac.uk/~gk/PTAM/

Using Bundle Adjustement

- With overlapped frames, at each frame the camera pose can be estimated from the sum of the squared distances of measured feature locations to the true image points for all points across all views

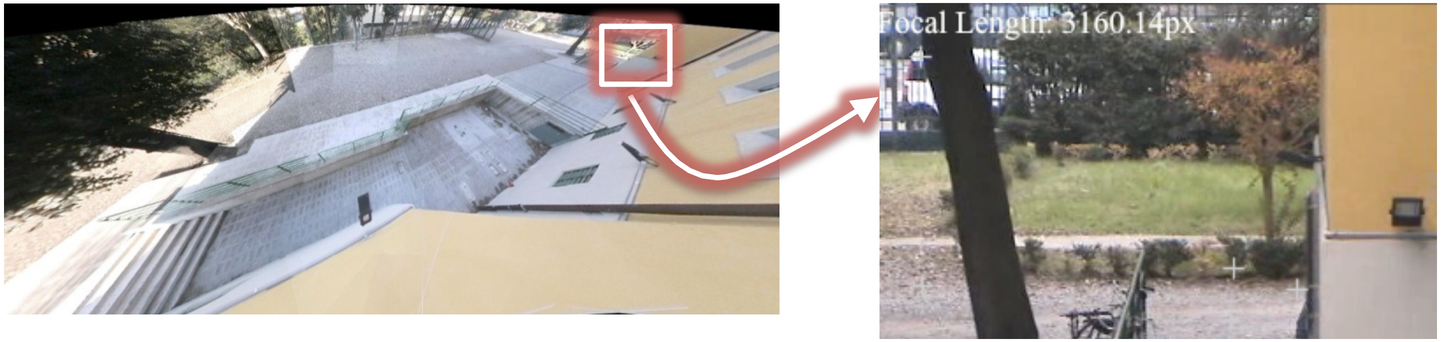
$$MLE = \arg\min_{\mathbf{K}_i ,\mathbf{R}_i ,\bar{\mathbf{X}}_j} \sum_{i=1}^{n}\sum_{j=1}^{m}\left\| \hat{\mathbf{x}}_{ij} - \mathbf{K}_i \mathbf{R}_i \bar{\mathbf{X}}_j \right\|^2$$



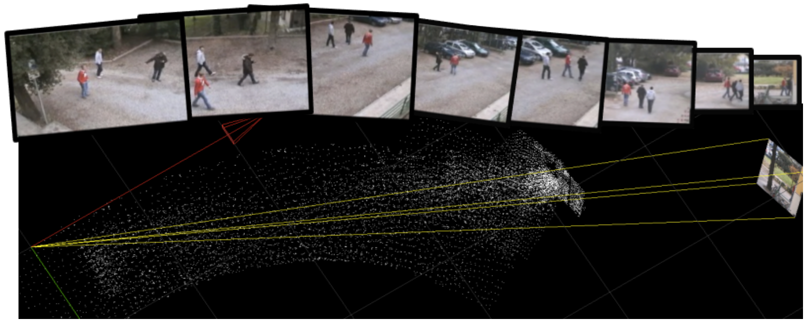
- The bundle of rays (shown colored) is adjusted to converge to a common direction

Calibration Overfitting with PTZ cameras

- Using bundle adjustment optimization with PTZ cameras based on a temporal window of the grabbed sequence results into a bias on the estimated parameters



320x240 resolution



Online bundle adjustment

✗ Estimated focal length = 3160 px

✓ Right focal length = 2160 px

Related Work: PTZ Camera Calibration

- *Hartley R.I.* “Self-Calibration from Multiple Views with a Rotating Camera” ECCV1994
- *L. de Agapito, E. Hayman and Ian D. Reid* Self-calibration of rotating and zooming cameras” IJCV2001
- *B. Tordoff and D. Murray* “The impact of radial distortion on the self-calibration of rotating cameras” CVIU2004
- *E. Hayman D. Murray* “The Effects of Translational Misalignment when Self-Calibrating Rotating and Zooming Cameras” PAMI2003
- *L.i Wang S. B. Kang H-Y. Shum G. Xu* “Error Analysis of Pure Rotation-based Self-Calibration” PAMI2004
- *Y. Seo and K. S. Hong* “About the Self-calibration of a Rotating and Zooming Camera: Theory and Practice” ICCV1999
- *H.Y.Shum and R.Szeliski* “Systems and experiment paper: Construction of panoramic mosaics with global and local alignment” IJCV2000
- *S. Sinha, M. Pollefeys* “Towards Calibrating a Pan-Tilt-Zoom Cameras Network” ECCV-OMNIVIS 2004

Detection and Tracking

Detection is Essential to Tracking

- Extracting target regions with high reliability is essential to tracking. With PTZ cameras detection at the right target scale is the true problem to perform tracking in real-time

- In order to follow moving targets for a potentially long period of time it is necessary to have some model of target appearence that is matched at each frame. Continuous tracking allows reasoning about where a target is, where it's going, and where it's been. Also, provides a compact representation of an object in the form of a trajectory

- Approaches to visual tracking can be broadly categorized into:
  - Global methods (using motion masks and/or appearance)
  - Local, feature-based methods

- Observations are usually passed through a filtering process:
  - Kalman filter (for simple motion, and “point like” objects, such as a car’s centroid)
  - Particle filter (for complex, highly non linear motion, and/or objects with complex shapes)

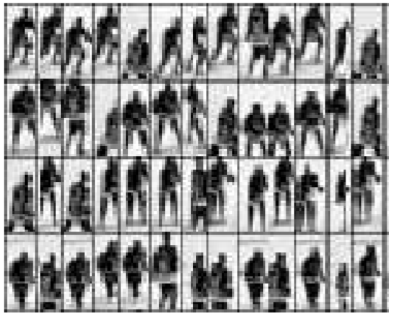
- Multiple target tracking needs for a data association mechanism

Detection and tracking with PTZ
Related Work and State of the Art

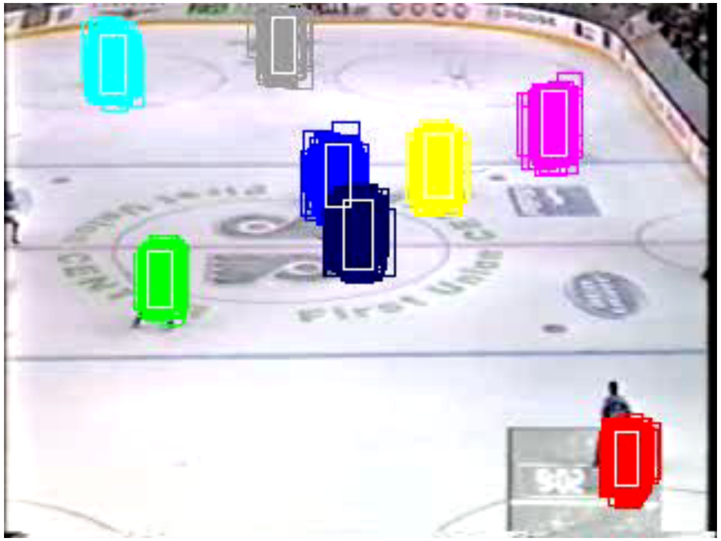
AdB

PTZ Multi-Target Tracking without calibration

A Boosted Particle Filter: Multitarget Detection and Tracking [*Okuma et al* ECCV ’04]



- Tracking is obtained with the combination of a Boosting detector and Particle filtering. Specifically tailored to track hockey players. Trajectories are extracted from the 2D plane without taking into account for 3D information
- The Boosting detector is trained to estimate the target scale. Moving targets are extracted from the sequence off line and used to train the detector



- For each detected target a particle filter is initialized for tracking. Particle updating is performed by combining the detector respose with the filter prediction. Imaged scale is embedded in the state
- Does not consider interaction between targets (nearest neighbor association)

AdB

PTZ Multi-Target Tracking without calibration

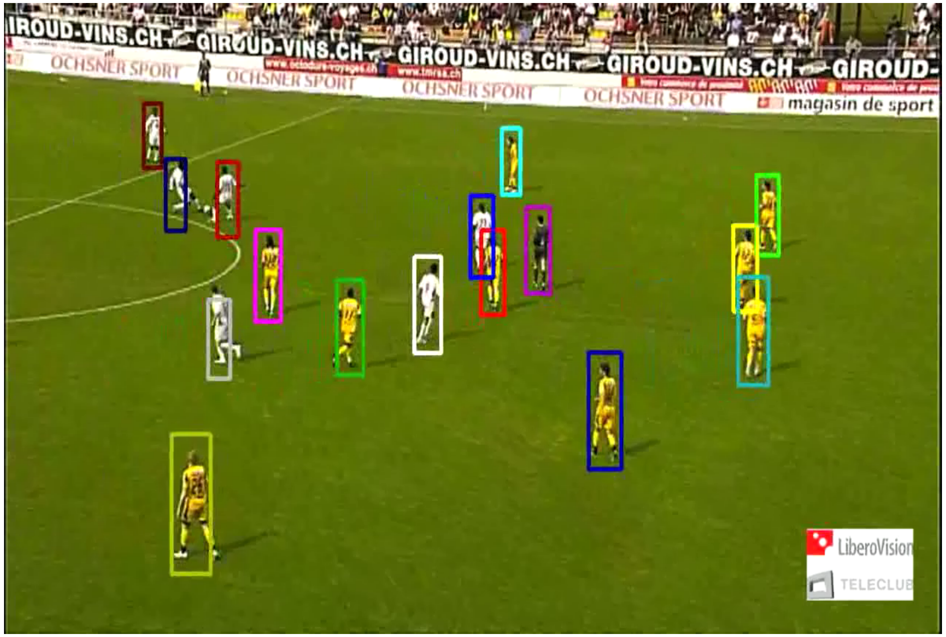
Online Multi-Person Tracking-by-Detection from a Single, Uncalibrated Camera [*Breitenstein et al* PAMI '10]

- Tracking is obtained with the combination of a HoG Detector and particle filtering Tracker. Trajectories are extracted from the 2D plane without taking into account 3D information

- Target size set to the average of the last 4 associated detections. One particle filter is instantiated per detected target. Target state contains position and speed

- Performs initialization and termination of target tracks

- Since tracking is performed on the 2D plane, the approach is unable to deal with abrupt motion of the camera. So not suited to PTZ cameras



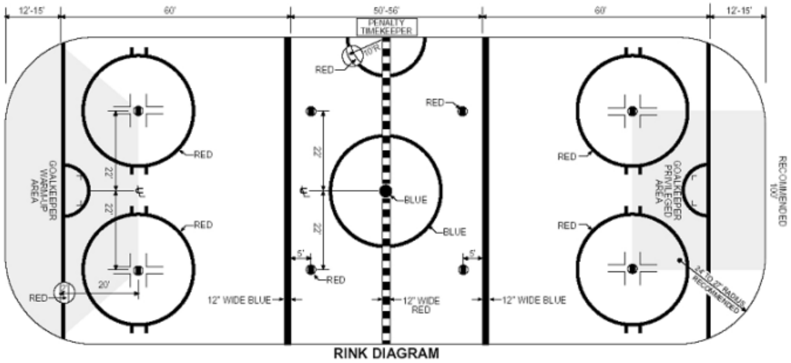
PTZ Multi-Target Tracking with calibration

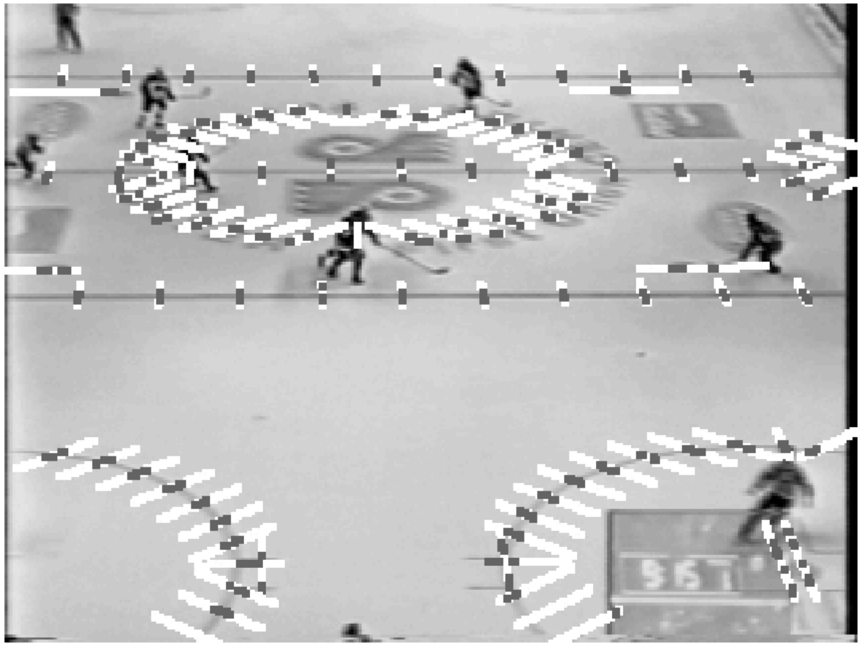
Robust Visual Tracking for Multiple Targets [*Cai et al* ECCV ’06]

- Extends the approach by [Okuma 2004] to perform tracking in 3D World Coordinates.Targeted to the special case of the hockey game.

- The world to image homography is estimated from the hockey rink model. A rectification technique is employed to find the correspondence between the video frame coordinates and the standard hockey rink coordinate. Target scale is computed by examining windows larger/ smaller than the current size

- Model based alignment: zooming can impair the method due to the fact that features to map are not anymore visible

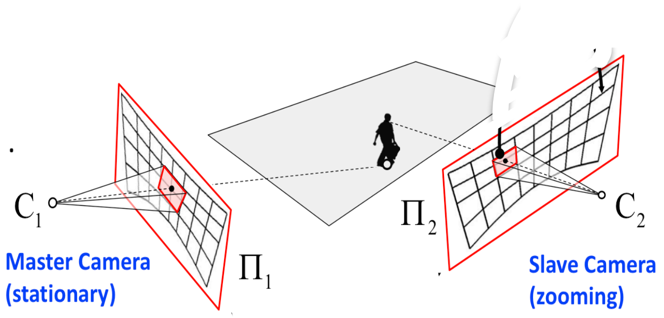


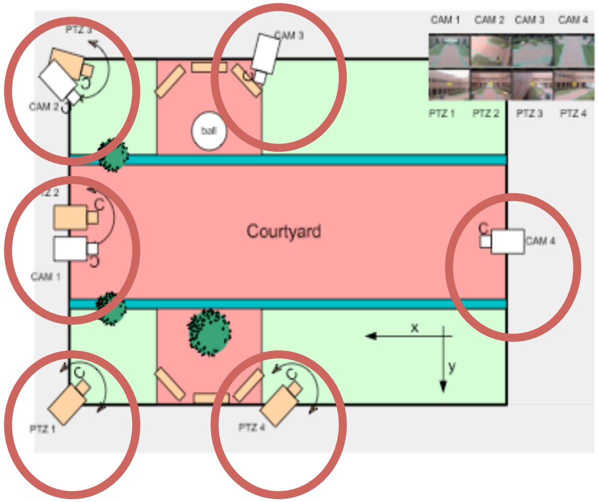


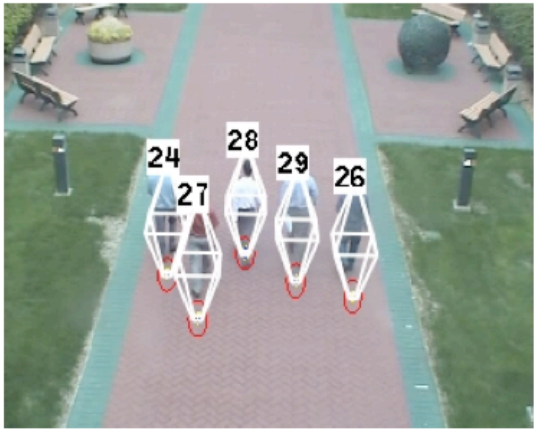
PTZ – fixed camera master slave configuration

Collaborative Real-Time Control of Active Cameras in Large Scale Surveillance Systems [*Krahnstoever et al* M2SFA2 ’08]

- Highly complex system with 4 PTZ cameras and 4 fixed sensors. Fixed cameras are calibrated and used to provide references to PTZ cameras in a master-slave relation
- Requires calibration, synchronization and coordination





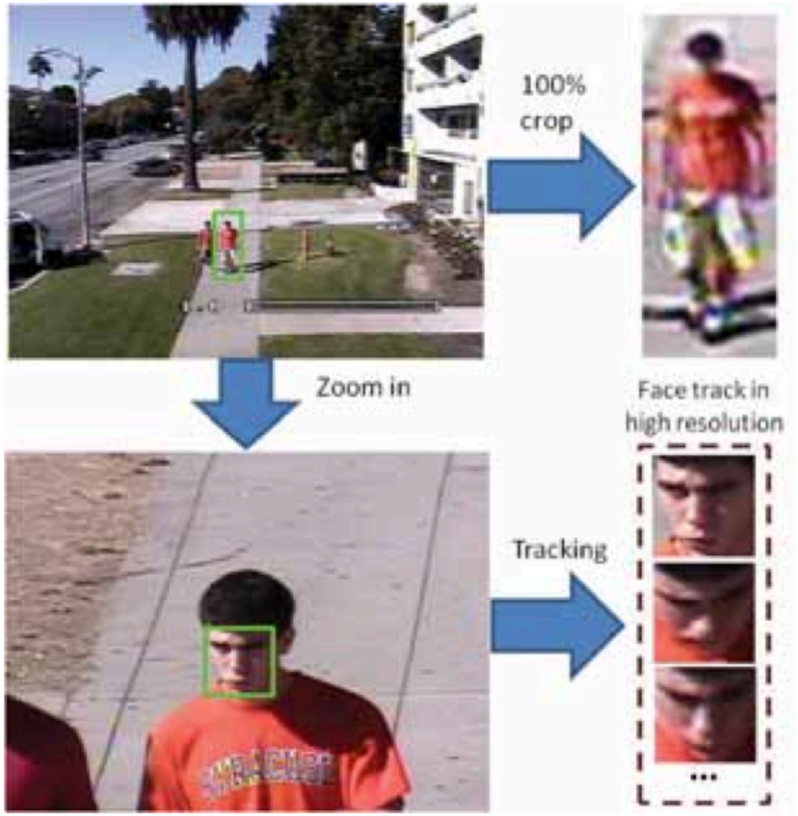


PTZ camera network

High Resolution Face Sequences from a PTZ Network Camera [*Dinh et al* 2011]

- Acquires high resolution sequences of a person's face using a pan-tilt-zoom camera network

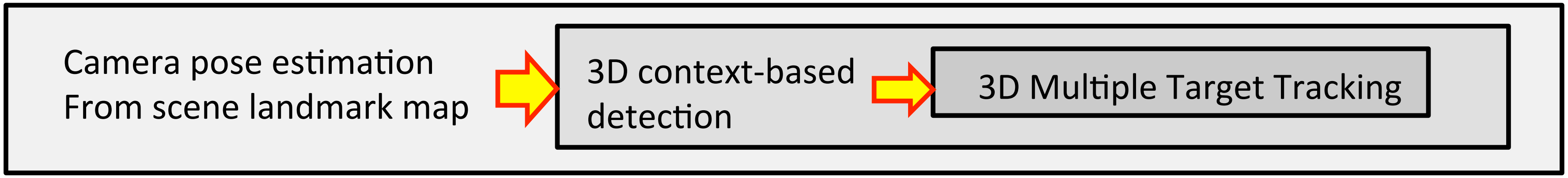
- Different states coresponding to regions of interest at a discrete set of zooming levels:
  - *Pedestrian detection mode*: operates at the the widest angle.
  - *Upper-body mode*: as a pedestrian is detected camera parameters are automatically tuned to focus onto the upper body of the detected person, where the face should appear
  - *Face detection mode*: at the narrowest angle, the face is detected using a face detector



Related Work: PTZ Camera Tracking

- E. H. T. Thorhallsson and D. W. Murray., “Zoom-invariant tracking using points and lines in affine views - an application of the affine multifocal tensors,” in *Proceedings of the IEEE International Conference on Computer Vision,* 1999. 2
- J. A. Fayman, O. Sudarsky, E. Rivlin, and M. Rudzsky, “Zoom tracking and its applications,” *Machine Vision and Applications,* vol. 13, no. 1, pp. 25–37, 2001. 2
- X. Zhou, R. Collins, T. Kanade, and P. Metes, “A Master-Slave system to acquire biometric imagery of humans at distance,” in *Proceedings of ACM International Workshop on Video Surveillance,* 2003. 1
- C. Micheloni, B. Rinner, and G. Foresti, “Video analysis in pan-tilt- zoom camera networks,” *IEEE Signal Processing Magazine,* vol. 27, no. 5, pp. 78 –90, sept. 2010. 1
- C.-C. Chen, Y. Yao, A. Drira, A. Koschan, and M. Abidi, “Cooperative mapping of multiple PTZ cameras in automated surveillance systems,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2009. 1
- N. Krahnstoever, M.-C. Chang, and W. Ge, “Gaze and body pose estimation from a distance,” in *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance,* 2011. 1
- K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe, “A boosted particle filter: Multitarget detection and tracking.” in *Proceedings of the European Conference on Computer Vision,* 2004. 1, 2, 3, 8, 10, 12
- N. d. F. Yizheng Cai and J. Little., “Robust visual tracking for multiple targets.” in *Proceedings of the European Conference on Computer Vision,* 2006. 1, 2, 3
- M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, “Online multi-person tracking-by-detection from a single, uncalibrated camera,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* no. 99, p. 1, 2010. 2, 8, 10, 12

A solution at MICC UNIFI



Camera pose estimation
From scene landmark map

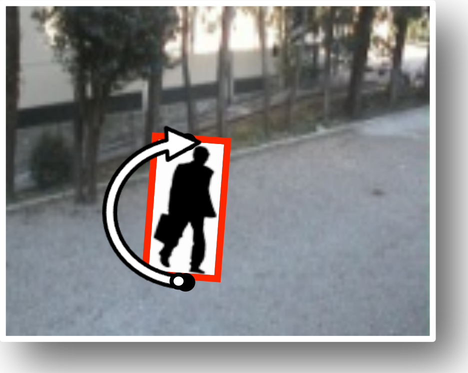
3D context-based detection

3D Multiple Target Tracking

- Landmark map update
- Landmark appearance update
- Landmark initialization and termination

- Expected targets height
- Landmark background estimation
- 3D temporal coherence

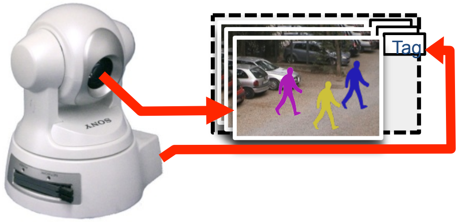
- Filtering
- Data association
- Targets appearance

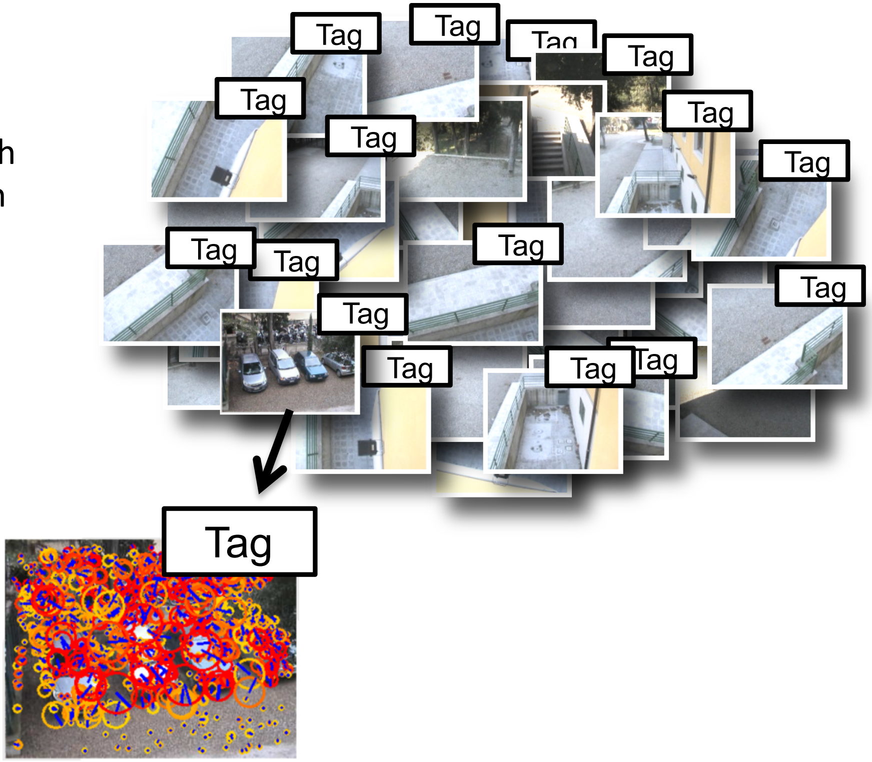


Scene Landmark Map Construction

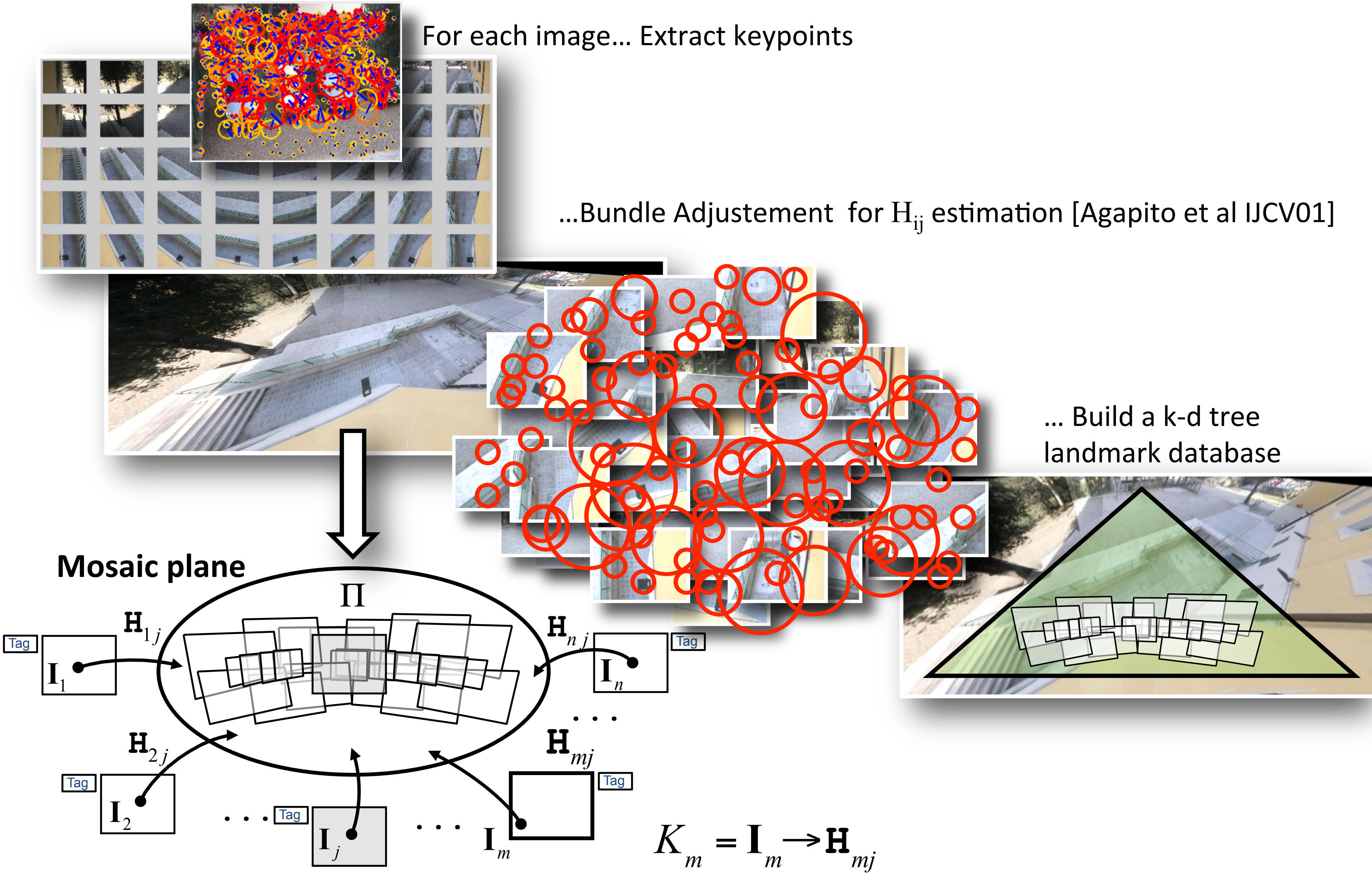
- In a batch phase we collect a base set of M images taken at different levels of pan-tilt and zoom so as to describe well the scene observed by the sensor. Each of these views is associated with:
  - The raw device camera values defined by the pan, tilt and zoom value from the PTZ sensor (Tag)
  - The extracted visual landmarks (Speeded Up Robust Features SURF to grant real-time performance)
  - The homography $\mathbf{H}_{ij}$ relating each view to a common reference plane $\mathbf{\Pi}_j$ (homographies are estimated through bundle adjustment)

- Tag = encoded output of camera engine. Each image collected in the map is associated with a Tag of the PTZ actuators





Map Construction Animated



For each image... Extract keypoints

...Bundle Adjustement for H_{ij} estimation [Agapito et al IJCV01]

... Build a k-d tree landmark database

Mosaic plane

$$K_m = \mathbf{I}_m \rightarrow \mathbf{H}_{mj}$$

3D Ground Plane Estimation

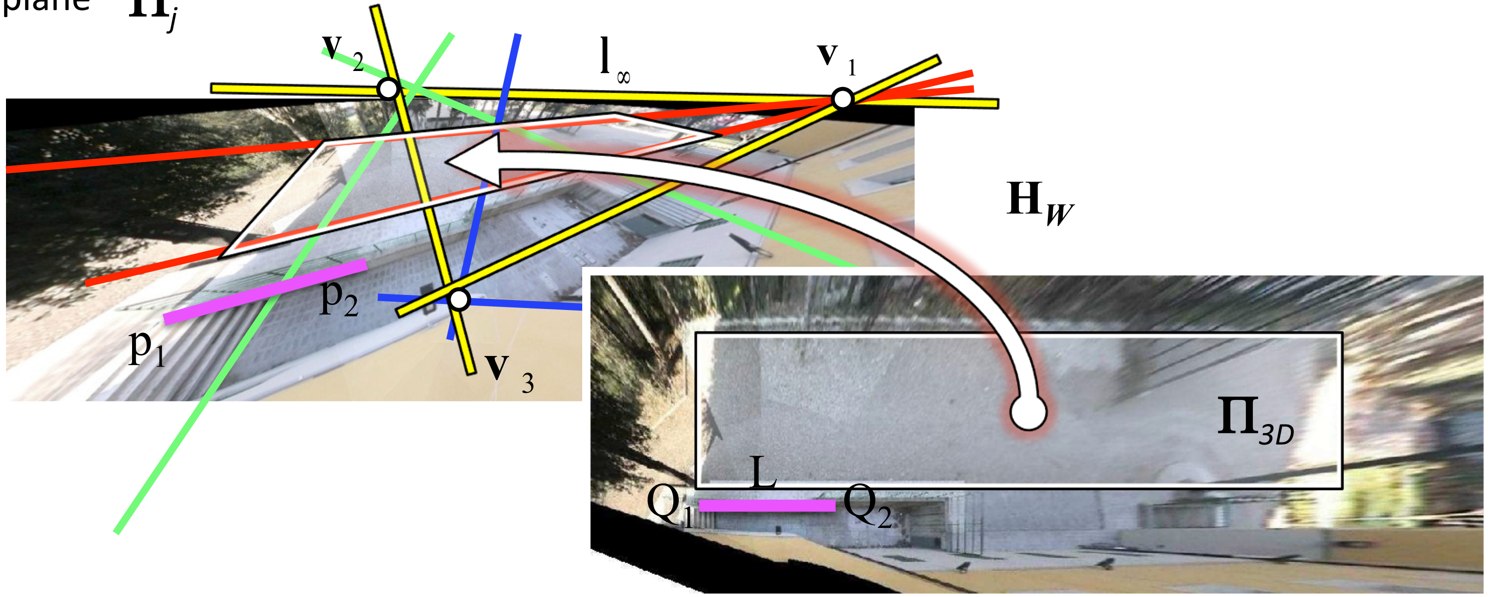
- To perform 3D tracking of targets we must estimate the homography that relates the mosaic plane to the 3D world, defined as $\mathbf{H}_W : \Pi_{3D} \rightarrow \Pi_j$ = $\mathbf{H}_p\mathbf{H}_s$ [Liebowitz, Zisserman CVPR98]

$$\mathbf{H}_p = \begin{pmatrix} \beta^{-1} & -\alpha\,\beta^{-1} & 0 \\ 0 & 1 & 0 \\ l_1 & l_2 & 1 \end{pmatrix}$$

Rectification Homography

l_1, l_2, α and β from the projections of the circular points of the imaged scene plane. Three pairs of lines (perpendicular between pairs) to identify the vanishing points and the vanishing line necessary to compute the rectification homography

- $\mathbf{H}_s$ computed from the projection of a pair of points at known distance from 3D world to mosaic plane Π_j



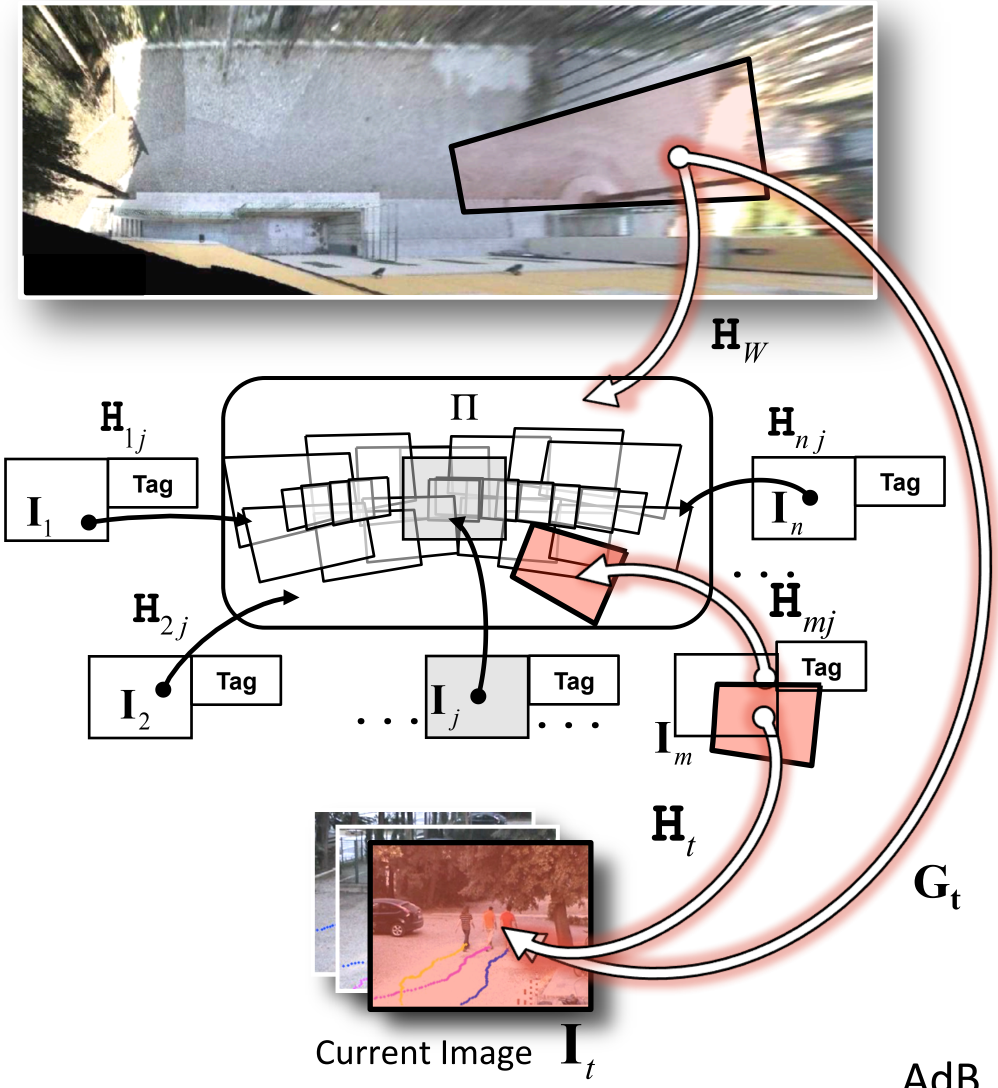
AdB

3D World to 2D Image Homography

- 3D World to 2D image homography:

$$\mathtt{G}_t = \underbrace{\mathtt{H}_t^{-1}}_{\mathbf{I}_t \leftarrow \mathbf{I}_m} \cdot \underbrace{\mathtt{H}_{mj}^{-1}}_{\mathbf{I}_m \leftarrow \Pi_j} \cdot \underbrace{\mathtt{H}_W}_{\Pi_j \leftarrow \Pi_{3D}}$$

3D world scene

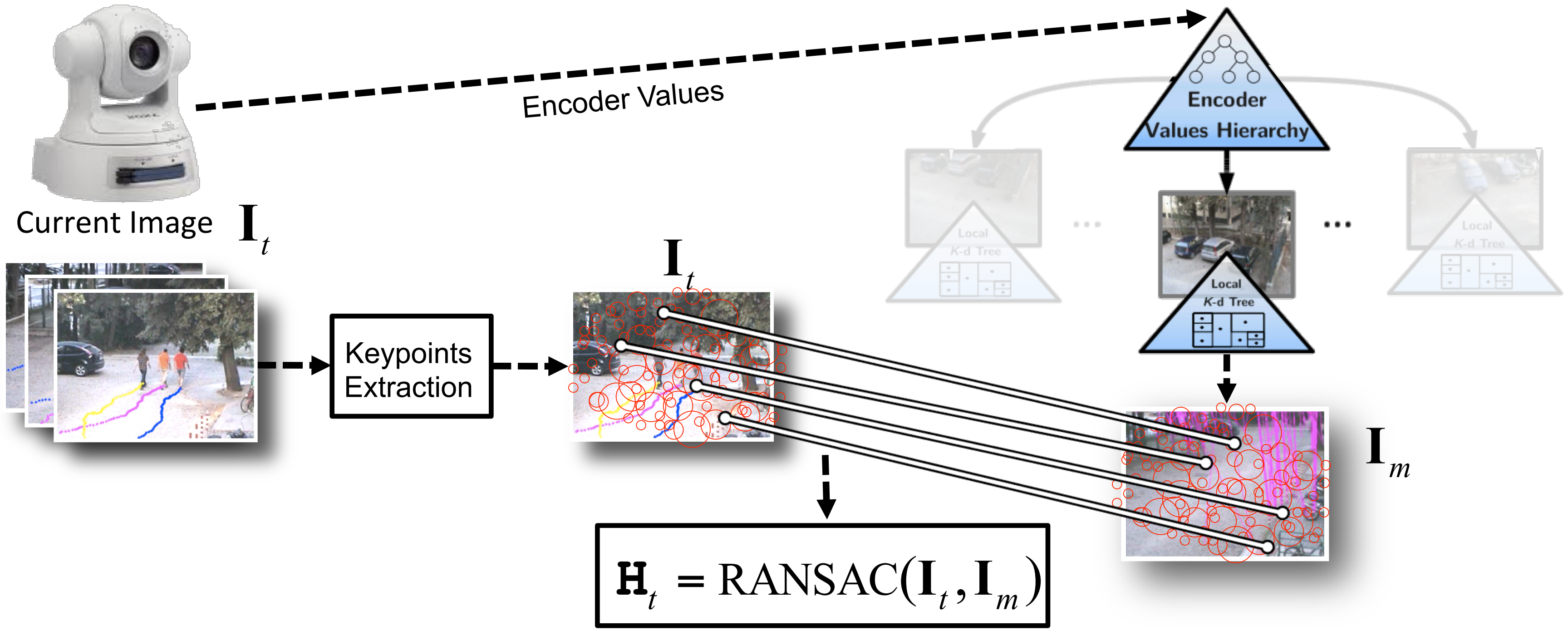


Current Image

AdB

Real-time Camera Pose Estimation

- To estimate the camera pose we must build at runtime the global transformation that relates the current frame observed by the PTZ camera with the 3D ground plane
  - Search over the device-tagged reference views obtained offline (encoder values and keypoint matching by nearest neighbor). Not accurate for camera pose estimation but retrieves a frame with most of the landmarks
  - Obtain the homography $\mathbf{H}_t : \mathbf{I}_t \rightarrow \mathbf{I}_m$ mapping the current frame onto the retrieved view by RANSAC



Encoder Values

Current Image $\mathbf{I}_t$

Keypoints Extraction

$\mathbf{I}_t$

Encoder Values Hierarchy

Local K-d Tree

Local K-d Tree

Local K-d Tree

$\mathbf{I}_m$

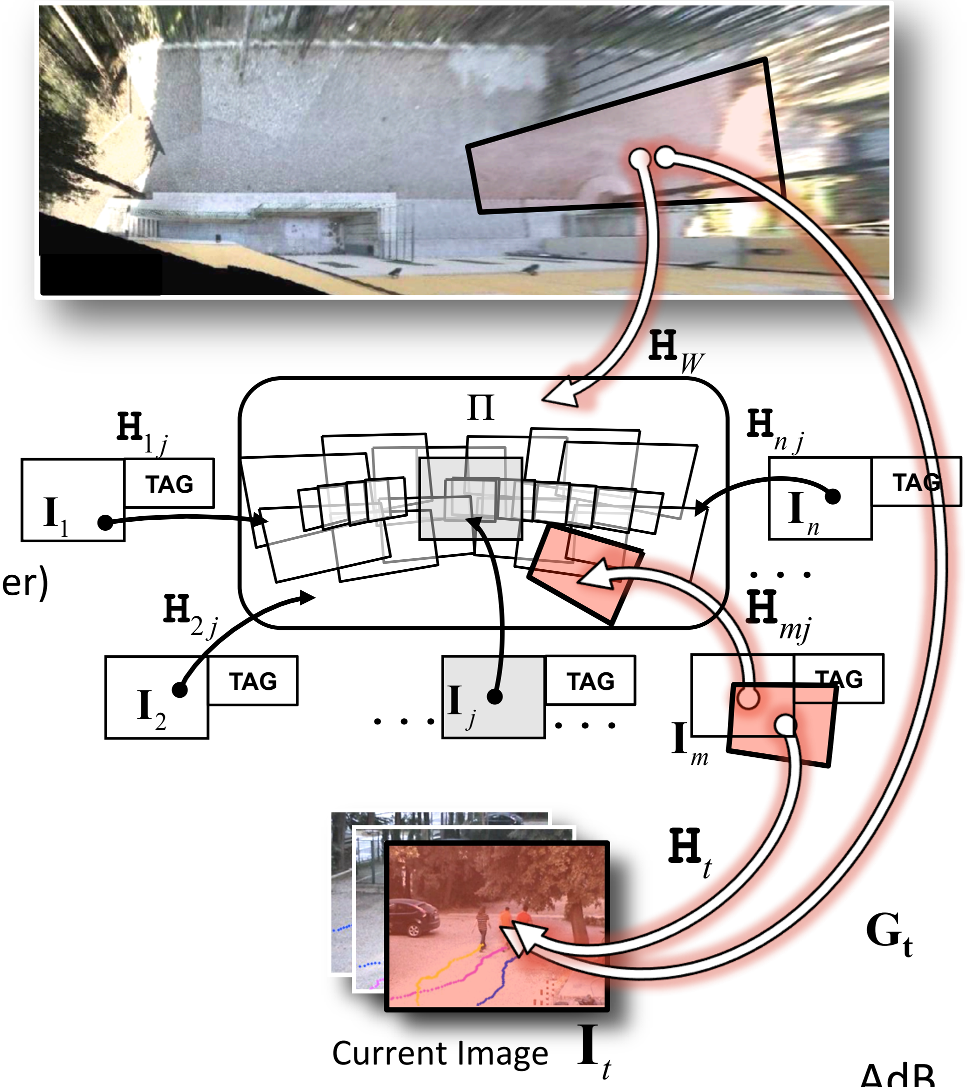
$$\mathbf{H}_t = \mathrm{RANSAC}(\mathbf{I}_t, \mathbf{I}_m)$$

Real Time Camera Pose Estimation Animated

- Obtain the time variant world to image homography:

$$\mathtt{G}_t = \underbrace{\mathtt{H}_t^{-1}}_{\mathbf{I}_t \leftarrow \mathbf{I}_m} \cdot \underbrace{\mathtt{H}_{mj}^{-1}}_{\mathbf{I}_m \leftarrow \Pi_j} \cdot \underbrace{\mathtt{H}_W}_{\Pi_j \leftarrow \Pi_{3D}}$$

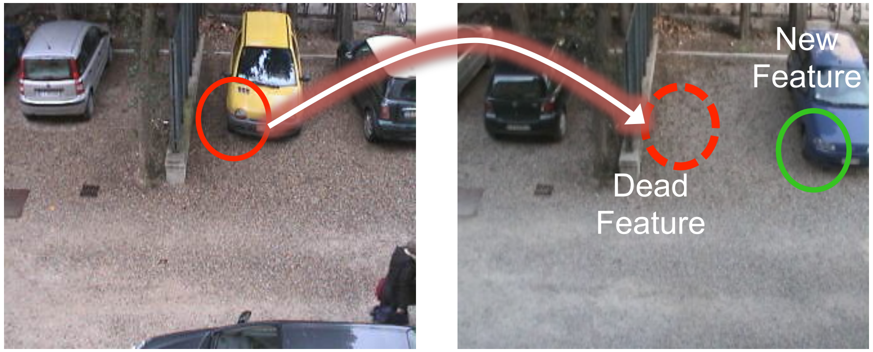
- Very robust $\mathbf{G}_t$ because:
  - $\mathbf{H}_t$ well conditioned (spaced keypoints)
  - $\mathbf{H}_{mj}$ very accurate ("bundle adjusted")
  - $\mathbf{H}_W$ best as possible (3D markers do better)

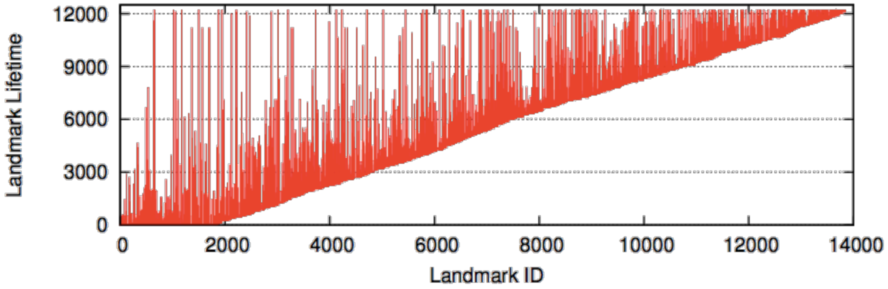


Landmark Initialization and Termination

- The scene may have changes due to lighting, entering exiting objects….Only few keypoints that are extracted offline are still alive at the time of operation and this number could not be sufficient for good calibration

- Birth-death management of landmarks is necessary. Landmarks added by the online procedure ensure a very good stability for the pose estimation

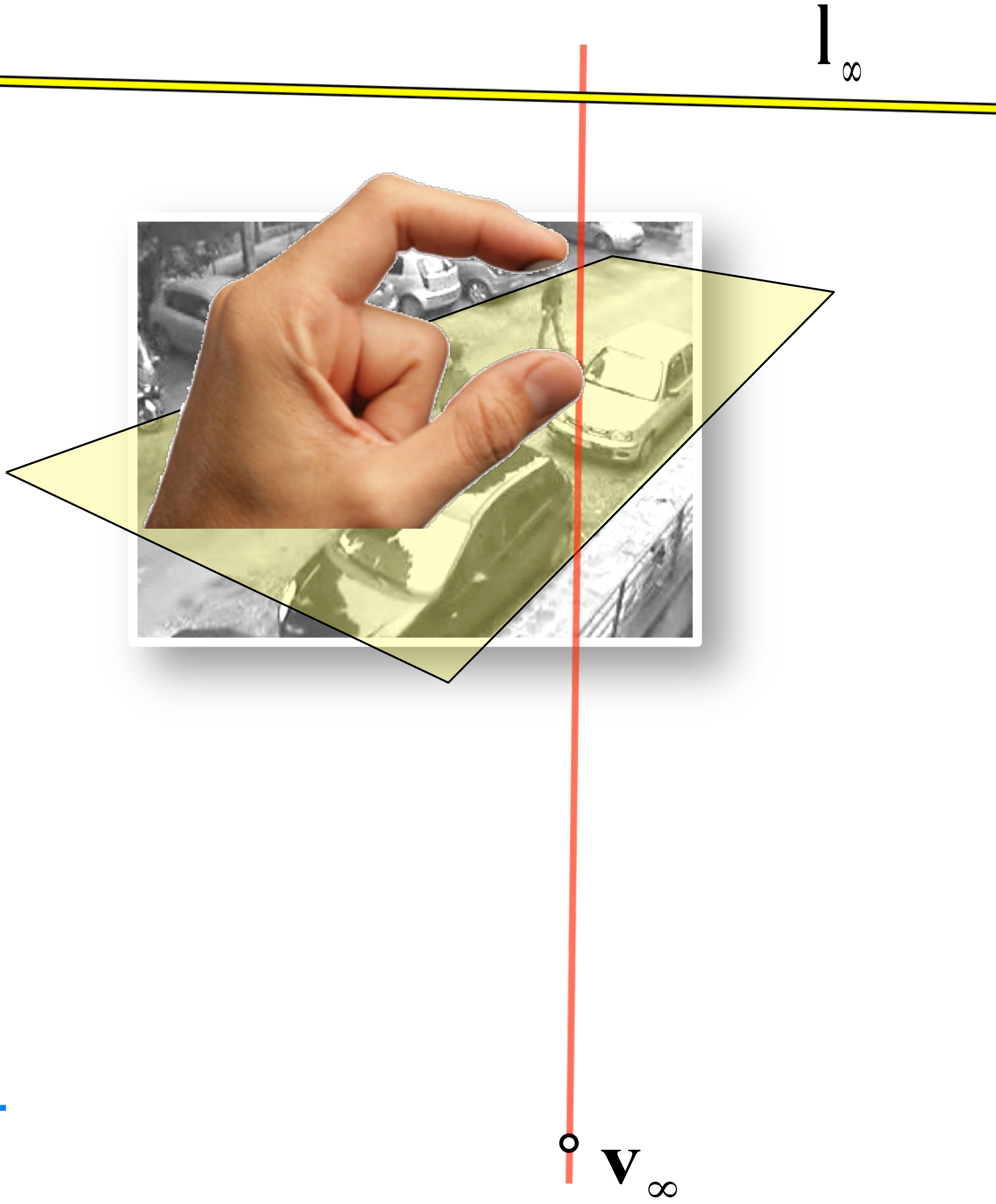




Landmarks with ID ≥ 2000 have been added at run-time

In operation……

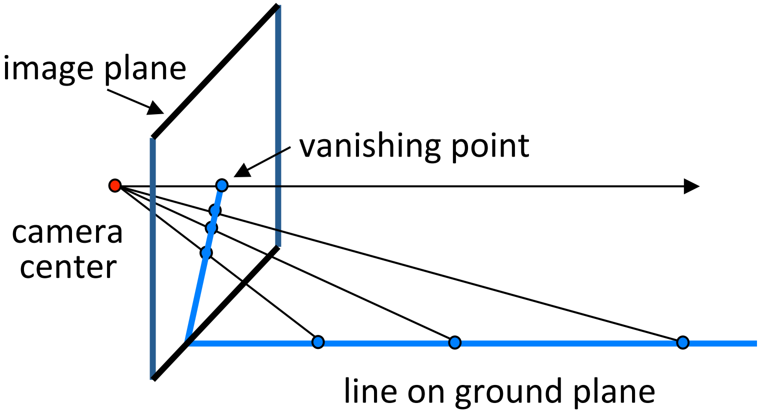
Exploiting 3D Information Improves Detection



- Clearly, image understanding is a 3D problem. Scene parts are all interconnected

- If we knew what makes a well-formed scene, we could use that knowledge to model the coherent scene in an image

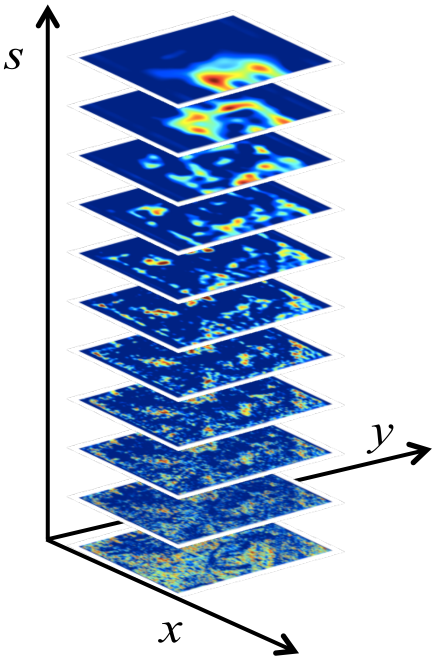
- The horizon (i.e. vanishing line) allows to reason in 3D about imaged object shape



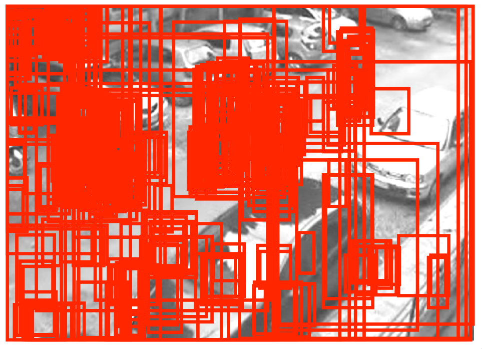
Pedestrian Detection with SVM Thresholded Responses

- State of the art pedestrian detectors use SVM trained filters to detect pedestrian in the image. Recent benchmark [*Dollar et. al* CVPR2009] shows that the best performing pedestrian detector have miss rate of over 80% at 1 FPPI

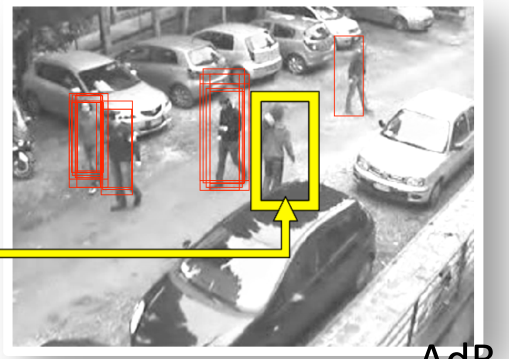
- Performance is improved by taking the outputs of the SVM at any scales and then thresholding the results by exploiting other knowledge
- This introduces a number of false positives. Errors can be removed by post-processing detections with suppression of non maxima responses (NMS)



- Low threshold:
  - Many false positive
  - Few missed detections

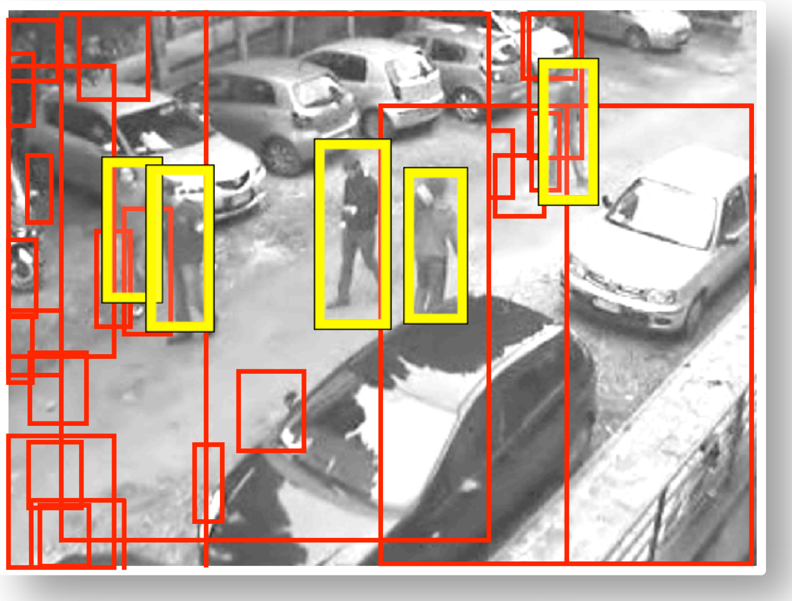


- Standard threshold:
  - A pedestrian missed

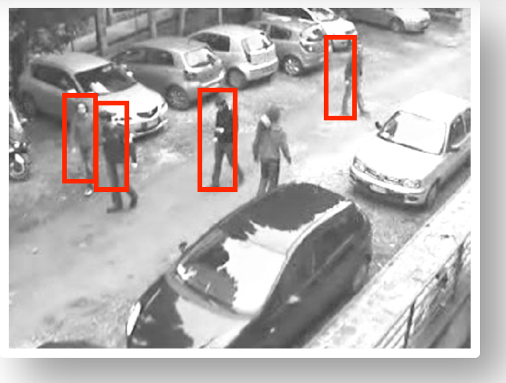


Non Maxima Suppression

- NMS applied [*Felzenswalb et al* CVPR 08]
  - when applied to low thresholded responses:
    - Several false positive still persist
    - Missed detections are less likely to happen



  - when applied to standard thresholded responses:
    - gives good precision but a bad recall



- The underlying assumption is that the detector response degrades gracefully in the 2D image plane (2D locality context NMS). But the 2D locality context treats all image positions and scales equally likely, and indeed they are not…

AdB

Exploiting Homologies

- A planar homology is a plane projective transformation with five degrees of freedom. It has a line of fixed points, called the axis, and a fixed point not on the line, called the vertex
- Planar homologies arise naturally in an image when two planes related by a perspectivity in 3D space are imaged. A homology can be computed from three matched features since it has five degrees of freedom and each feature gives two constraints

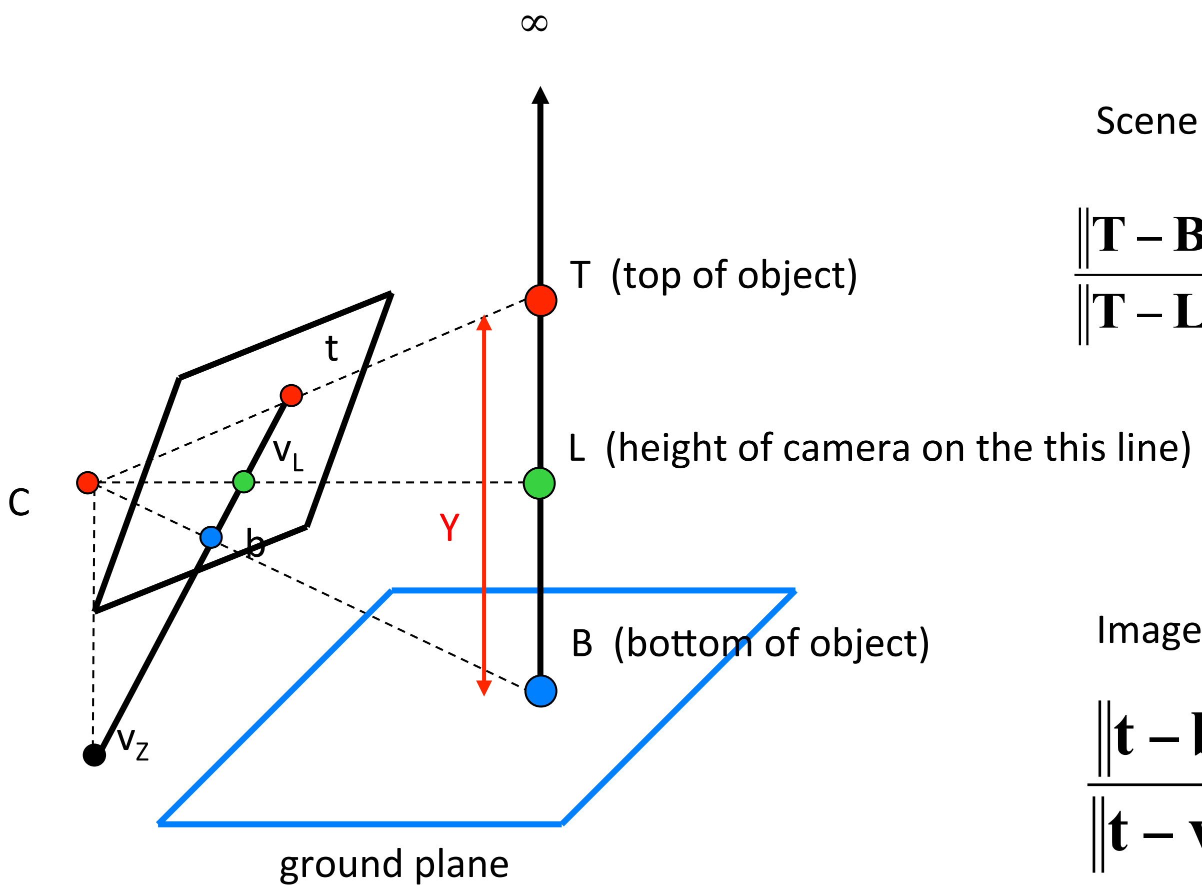
- The projective transformation representing the homology can be parameterized directly in terms of the vector representing the axis l, the vector representing the vertex v, and the characteristic cross-ratio μ as

$$\mathtt{W}_t = \mathtt{I} + (\mu - 1)\frac{\mathbf{v}_{t,\infty} \cdot \mathbf{l}_{t,\infty}^{\mathrm{T}}}{\mathbf{v}_{t,\infty}^{\mathrm{T}} \cdot \mathbf{l}_{t,\infty}}$$

- The cross ratio μ is formed by corresponding points, the vertex and the intersection of their join with the axis, and it is an invariant of the homology

AdB

Measuring Height



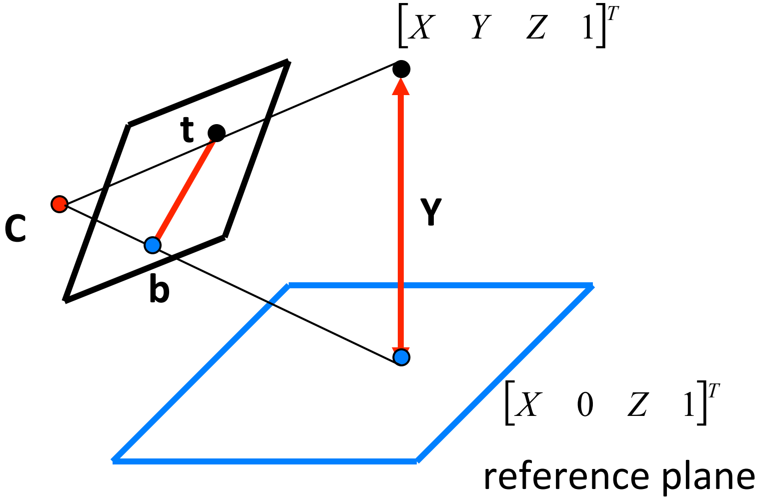
Scene Cross Ratio

$$\frac{\|\mathbf{T}-\mathbf{B}\|\,\|\infty-\mathbf{L}\|}{\|\mathbf{T}-\mathbf{L}\|\,\|\infty-\mathbf{B}\|}=\frac{1}{\|\mathbf{T}-\mathbf{L}\|}Y=\alpha Y$$

Image Cross Ratio

$$\frac{\|\mathbf{t}-\mathbf{b}\|\|\mathbf{v}_Z-\mathbf{v}_L\|}{\|\mathbf{t}-\mathbf{v}_L\|\,\|\mathbf{v}_Z-\mathbf{b}\|}=\alpha Y$$

Measuring Height


$\begin{bmatrix} X & Y & Z & 1 \end{bmatrix}^T$
t
C
b
Y
$\begin{bmatrix} X & 0 & Z & 1 \end{bmatrix}^T$
reference plane

Algebraic Derivation

$$\rho\,\mathbf{b} = \mathbf{\Pi}\begin{bmatrix} X & 0 & Z & 1 \end{bmatrix}^T = Xa\,\mathbf{v}_X + Zb\,\mathbf{v}_Z + \mathbf{l}$$

$$\mu\,\mathbf{t} = \mathbf{\Pi}\begin{bmatrix} X & Y & Z & 1 \end{bmatrix}^T = Xa\,\mathbf{v}_X + Yb\,\mathbf{v}_Y + \alpha Z\,\mathbf{v}_Z + \mathbf{l}$$

- Eliminating ρ and μ yields $$\alpha Y = \frac{-\|\mathbf{b}\times\mathbf{t}\|}{\mathbf{l}^T\mathbf{b}\|\mathbf{v}_Y\times\mathbf{t}\|}$$
- Can calculate α given a known height in scene

AdB

Imaged Height Context

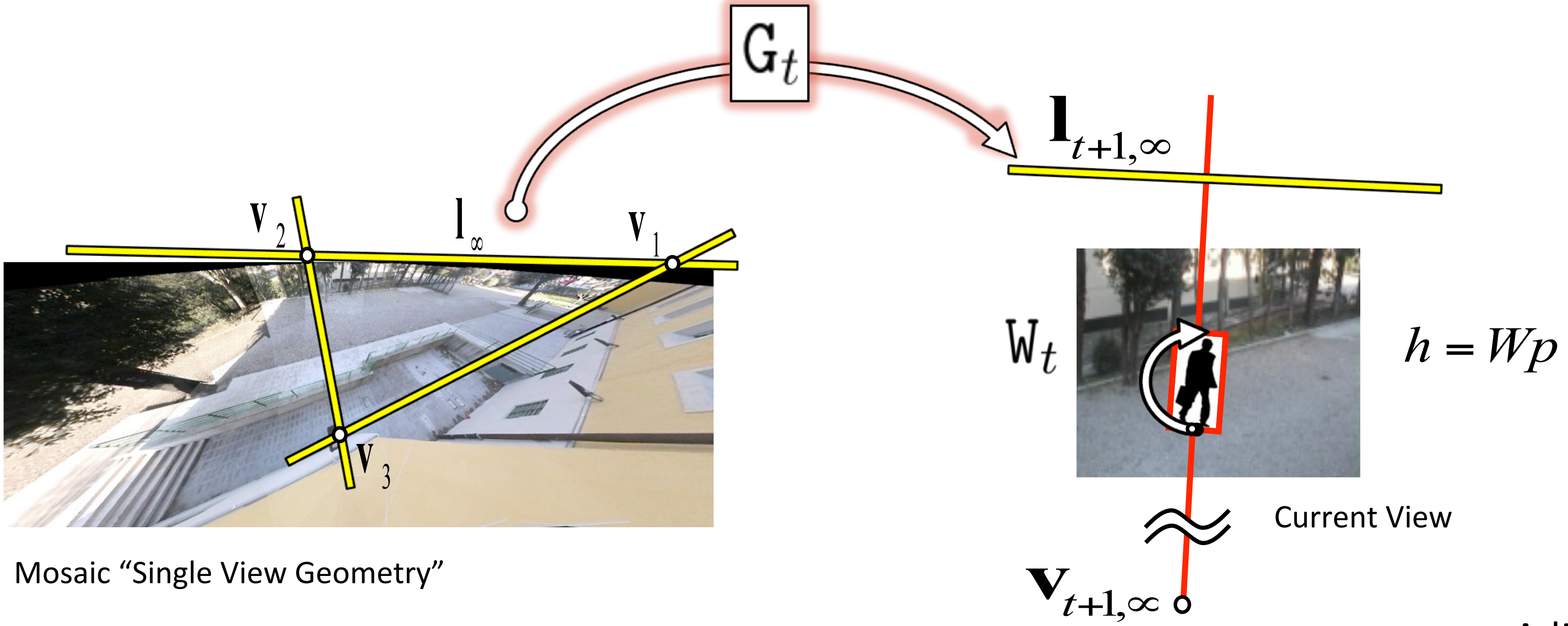
- At each time t the estimate of camera pose and the relationship G permit to infer the expected height of a target. Mosaic plane “Single View Geometry” is transferred onto the current view. Extremities related by a planar homology [*Criminisi et al.* IJCV2000]:

- Assumption: targets are closely vertical in the 3D scene plane

$$\mathtt{W}_t = \mathtt{I} + (\mu - 1)\frac{\mathbf{v}_{t,\infty} \cdot \mathbf{l}_{t,\infty}^{\mathrm{T}}}{\mathbf{v}_{t,\infty}^{\mathrm{T}} \cdot \mathbf{l}_{t,\infty}}$$

$$v_{t+1,\infty} = KK^T l_{t+1,\infty}$$

$$\mathbf{l}_{t+1,\infty} = G_t \left[0\ 0\ 1\right]^T$$

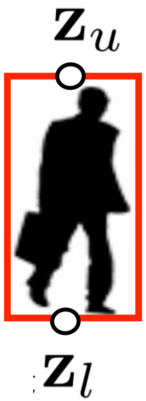


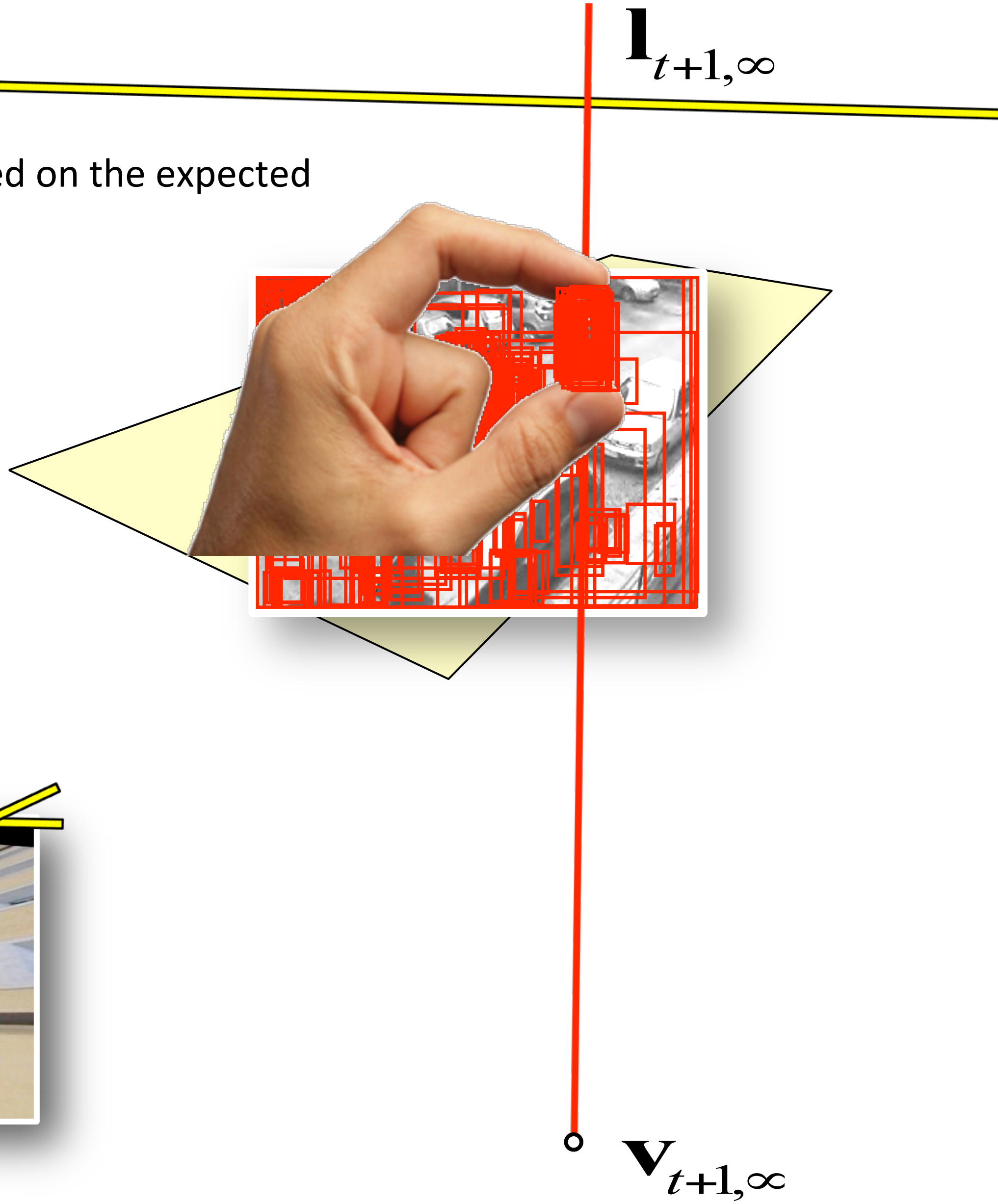
Mosaic “Single View Geometry”

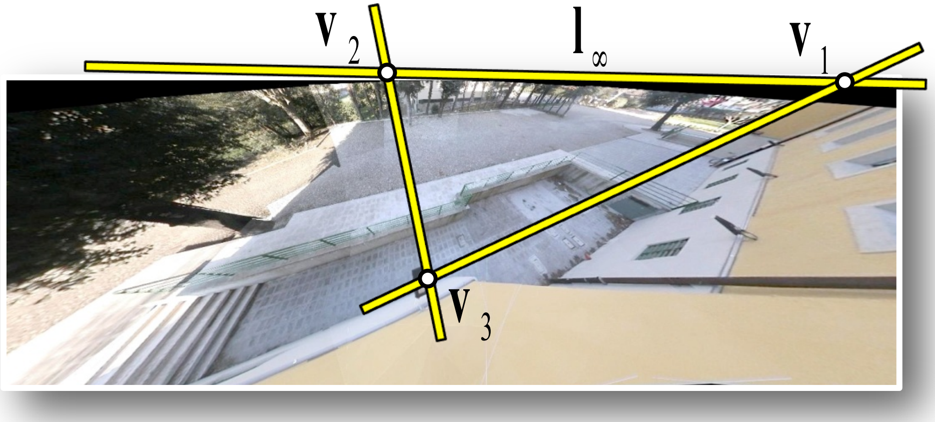
AdB

Detection Exploiting 3D Context

- Target bounding boxes are filtered out based on the expected height…

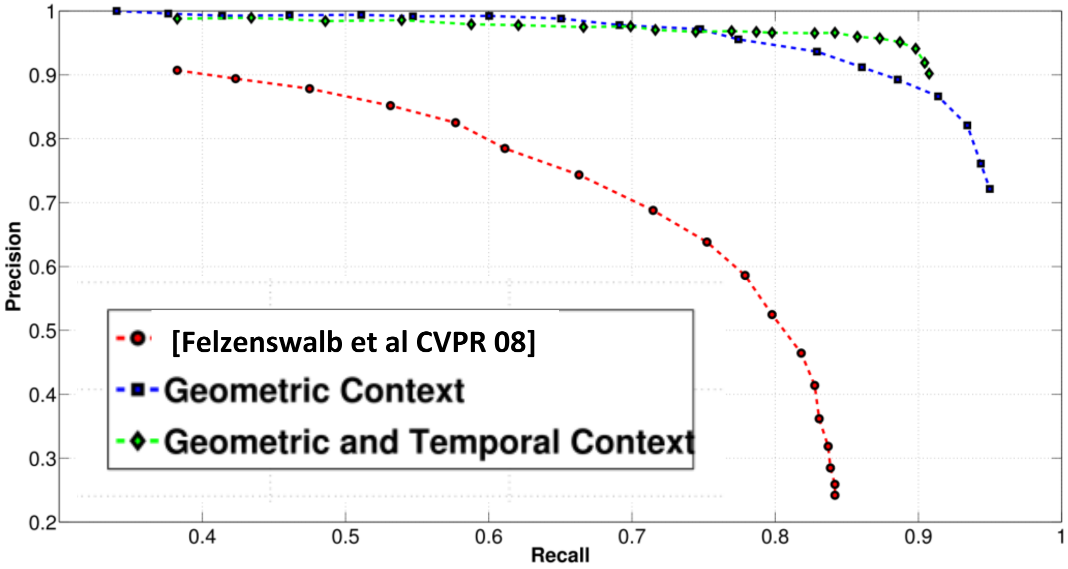




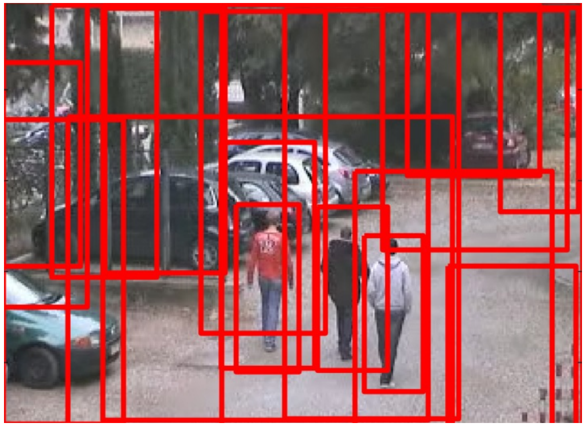


Mosaic “Single View Geometry”

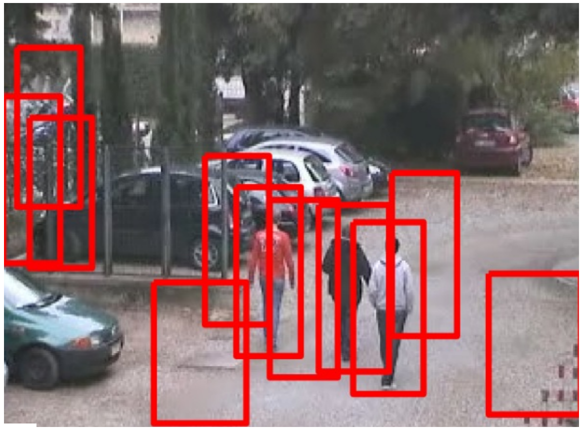
Experimental Results



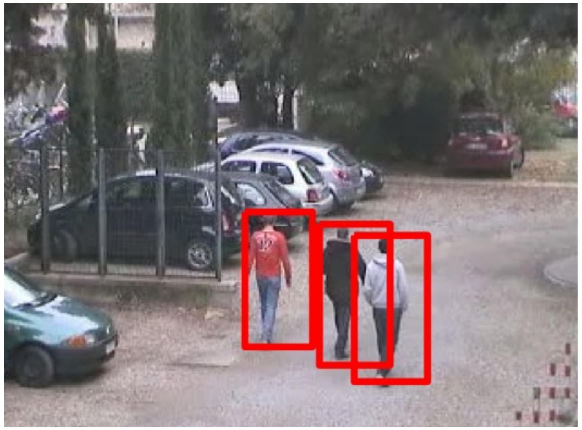
- Applying 3D geometric and temporal context filtering to Felzenswalb's detector
- Does not run in real-time anyway



[CVPR08] low Threshold



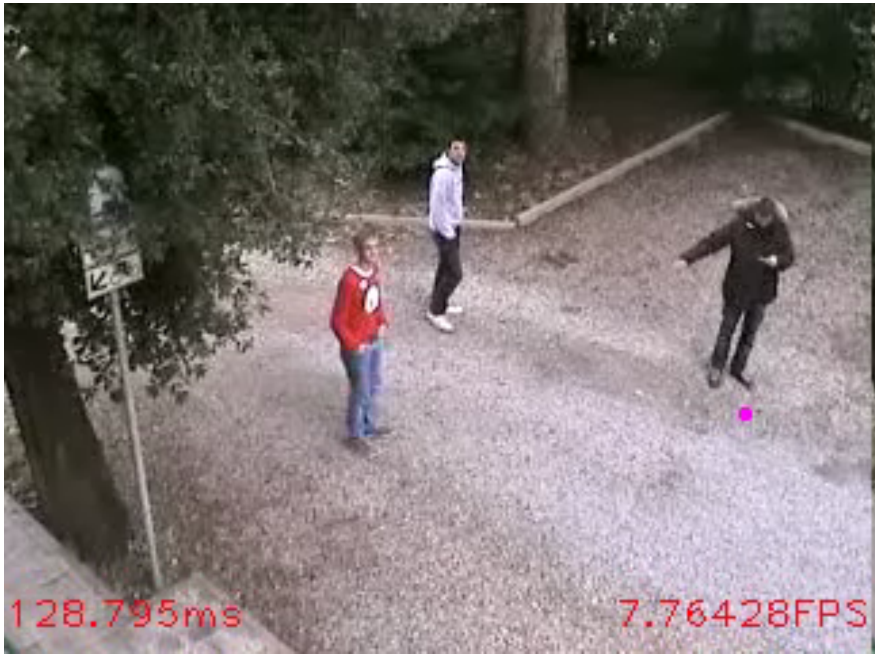
.. + Geometric context



... +temporal context

AdB

Real Time Detection in Comparison with Dalal & Triggs' Detector



Context FIltered

[Dalal-Triggs CVPR 05]

Tracking by Extended Kalman Filter

- The observation model for each target moving on the ground plane is:

$$\mathbf{z} = \mathbf{g}\,(\mathrm{x}_t\,\mathrm{y}_t) + \mathbf{v}_t$$

$$\mathtt{G}_t = \mathtt{H}_t \mathtt{H}_{mj}^{-1} \mathtt{H}_W$$

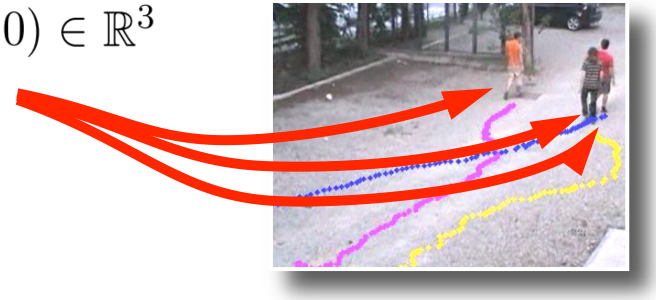
where **v** is a Gaussian noise term that models target localization error in the frame *t*, **g** is obtained linearizing the 3D world to 2D image homography G

- The motion model in world coordinates has constant velocity:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_{t-1},$$

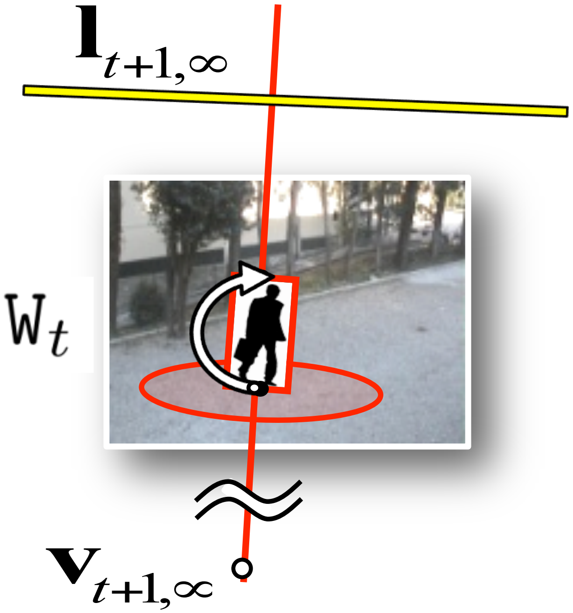
- The state of each target is defined by foot position and speed

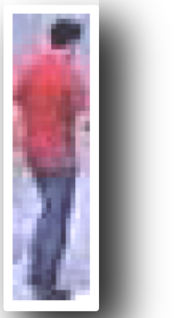
$$\mathbf{x}_t = [x_t, y_t, \dot{x}_t, \dot{x}_t] \quad (x_t, y_t, 0) \in \mathbb{R}^3$$



- Association of target measurements to target tracks is made with muti-stage approach

Tracking with Greedy and Soft Association

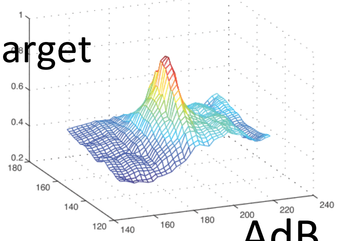




Template resized according to W_t

Sliding window search

- Greedy association of target positions to tracks is performed by Appearance Template Matching: brute force method for tracking single objects
  - Define a search area
  - Place the template defined from the previous frame at each position of the search area and compute a similarity measure between the template and the candidate
  - Select the best candidate with the maximal similarity measure
  - Use Spatiogram (not scale invariant, gives a peaked response around the target



$$h_I(b) = \langle n_b, \mu_b, \Sigma_b \rangle\,, \quad b = 1, \dots, B,$$

AdB

Tracking with Greedy and Soft Association

- Measurements that have not been associated to any track are softly associated computing probabilities of targets association

- At each time a validation region is established around the predicted target position and n points are randomly sampled. At each point a rectangular template is taken. The three most similar are considered to form hypothesis of the target position

Extracted measures "j" are jointly associated with each target "t" through the weight β_{jt} (Cheap JPDAF)

$$\beta_j^t = \frac{G_{tj}}{S_t + S_j - G_{tj} + B} \qquad S_j = \sum_{t=1}^{T} G_{tj}$$

$$G_{tj} = \mathcal{N}[\ \nu_j(k)\ ] \qquad S_t = \sum_{j=1}^{m} G_{tj}$$

These probabilities of association are used to update the state of each target by calculating the weighted innovation

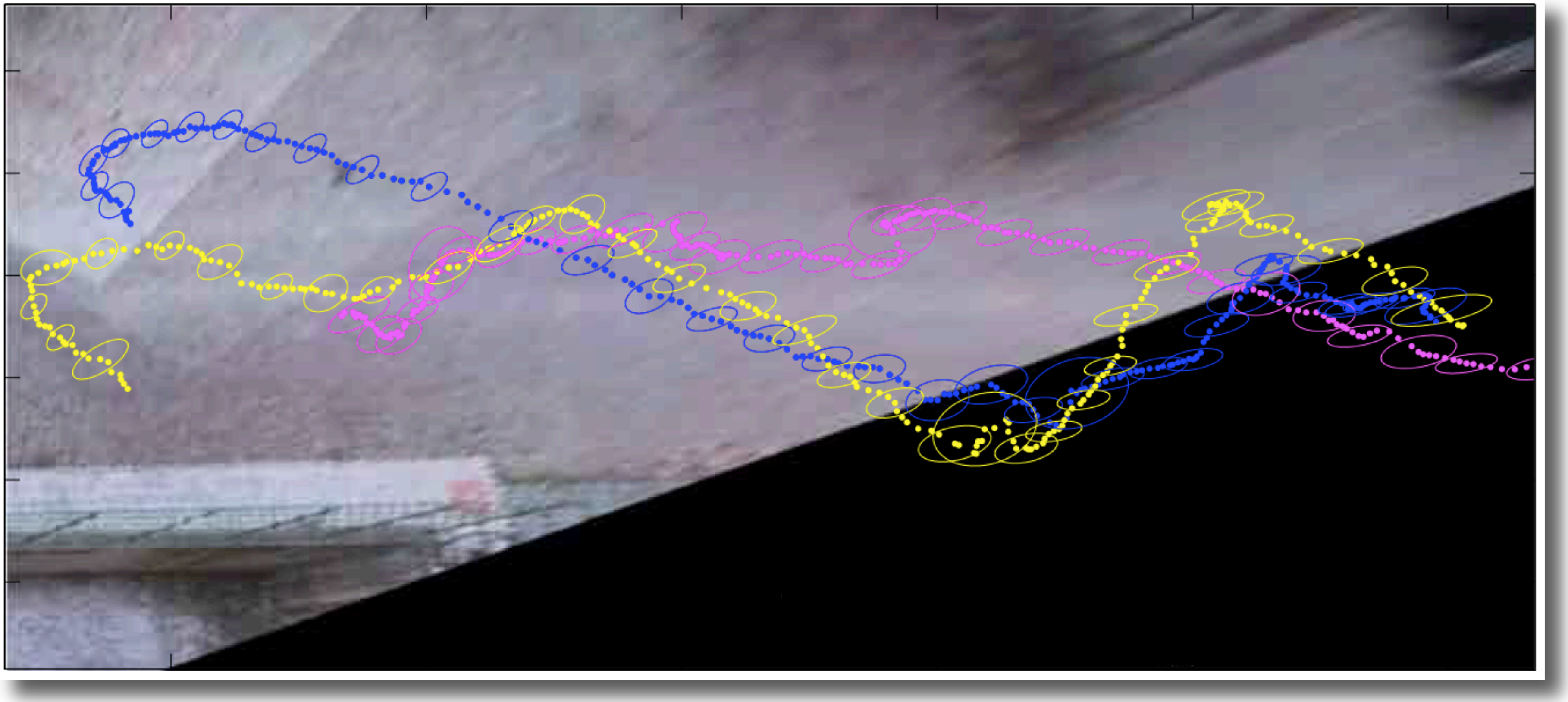
$$\tilde{\mathbf{z}}_j^t \triangleq \mathbf{z}_j - \hat{\mathbf{z}}^t$$ Innovation for measure j

$$\tilde{\mathbf{z}}^t = \sum_{j=1}^{m} \beta_j^t\ \tilde{\mathbf{z}}_j^t$$ Weighted innovation for target t

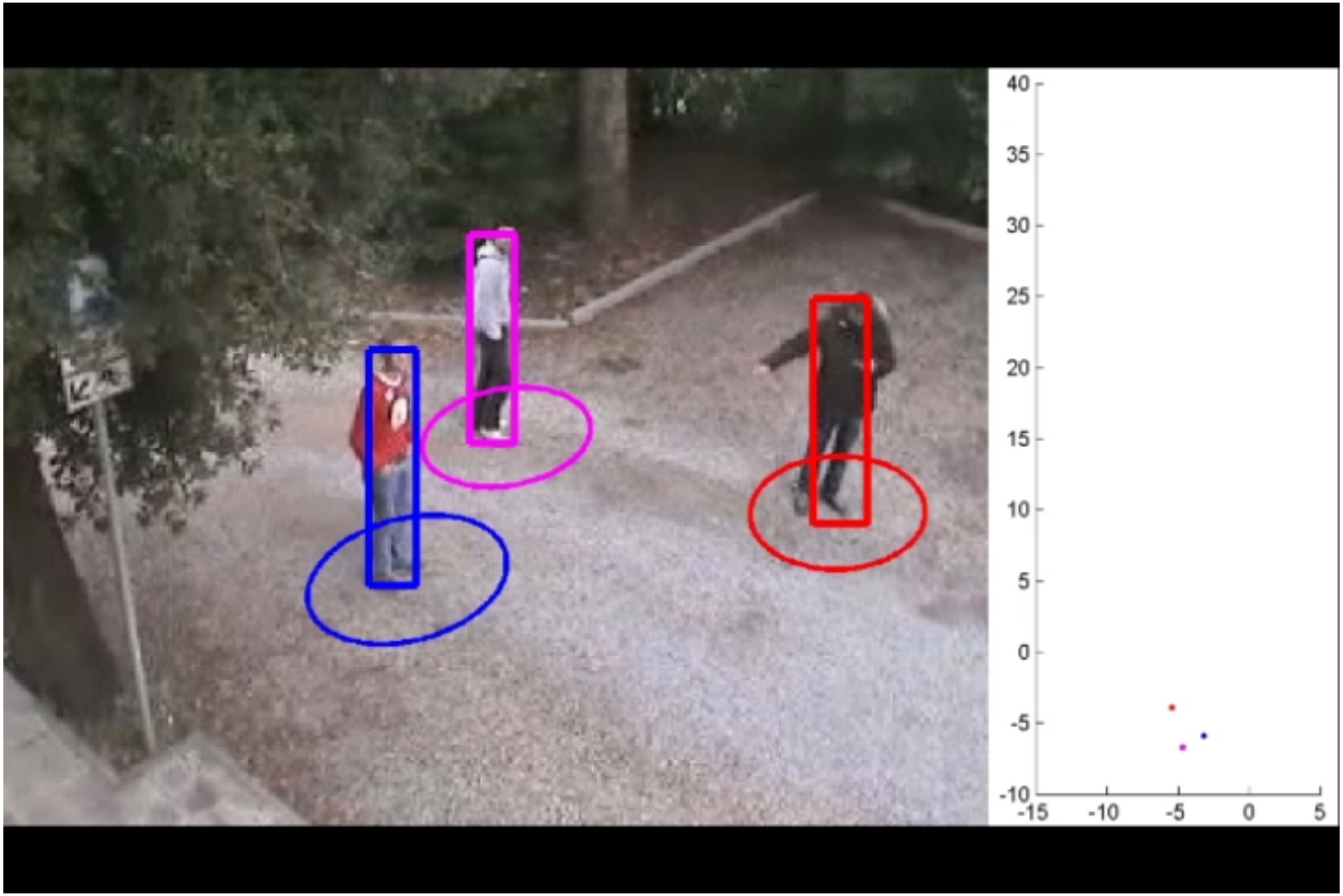
- Kalman gain determines how much we should trust observations

Experimental Results

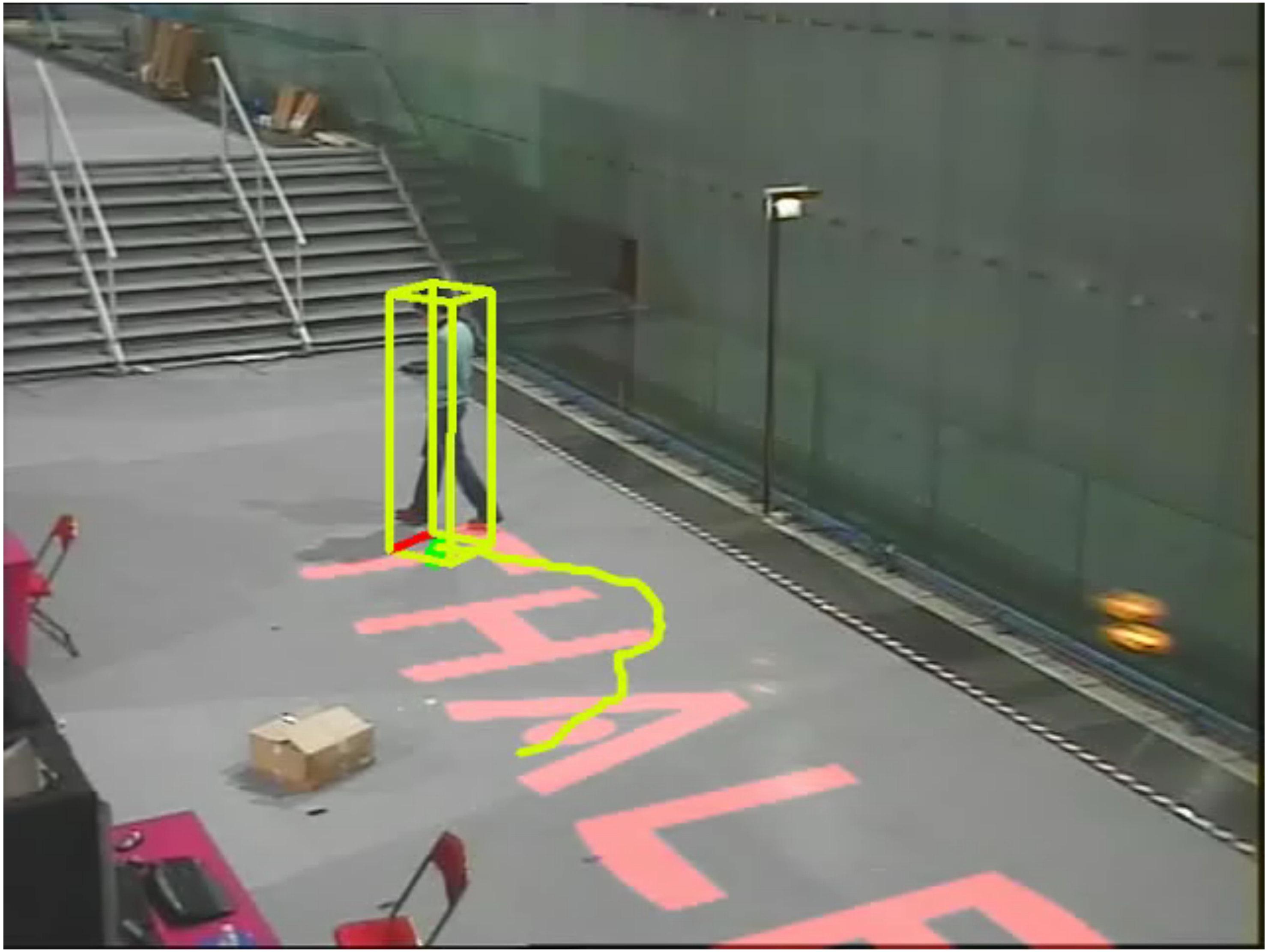
Constant standard deviation error (less than 0.3 meters) in recovering 3D trajectories of multiple moving targets in an area of 70x15 meters



Tracking at work



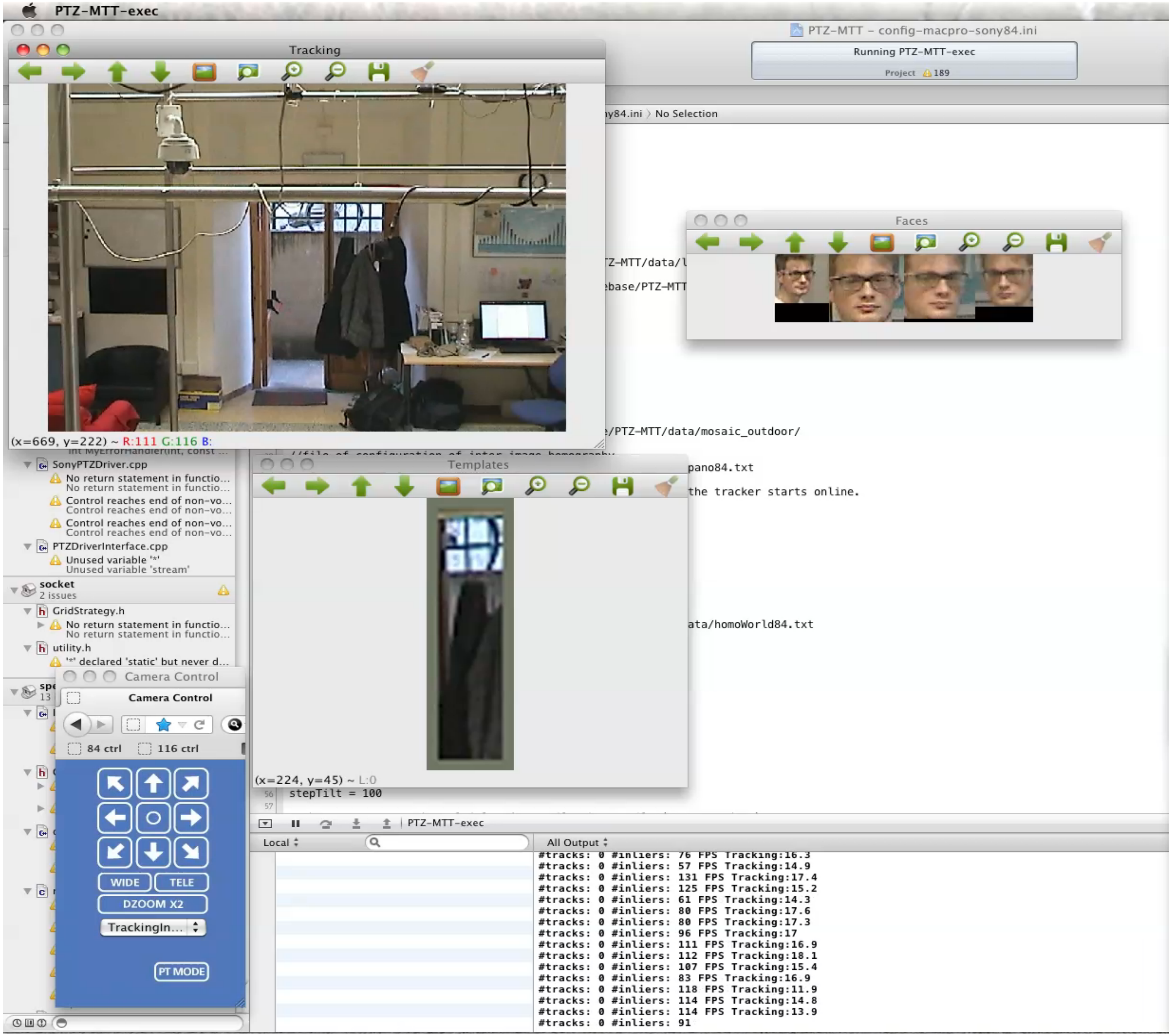
Tracking at work



Tracking at work

AdB

Tracking at work



(x=669, y=222) ~ R:111 G:116 B:

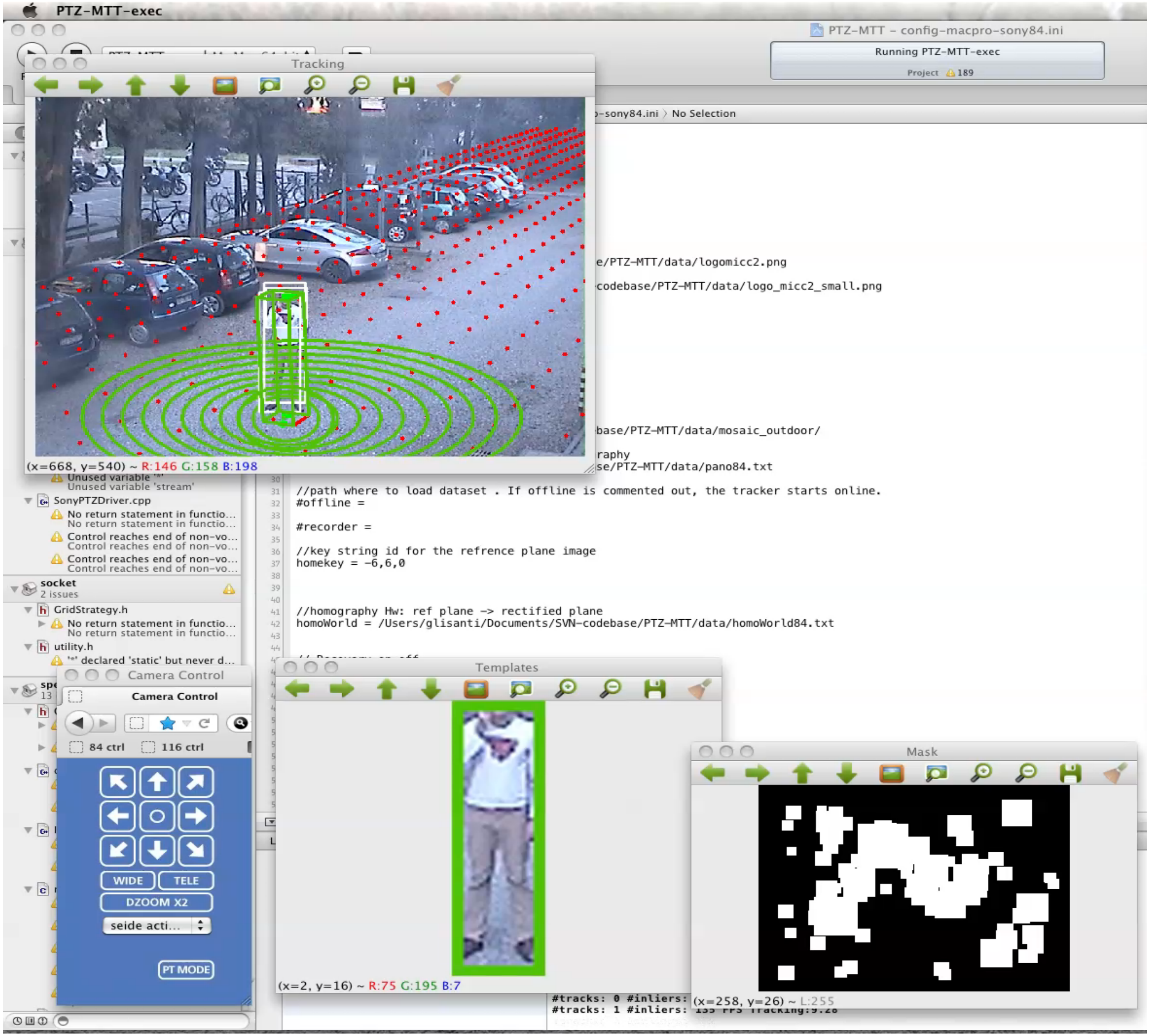
(x=224, y=45) ~ L:0

```
y84.ini 〉 No Selection
Z-MTT/data/l
base/PTZ-MTT
/PTZ-MTT/data/mosaic_outdoor/
pano84.txt
the tracker starts online.
ata/homoWorld84.txt
stepTilt = 100
```

```
int MyErrorHandler(int, const ...
SonyPTZDriver.cpp
  No return statement in functio...
  No return statement in functio...
  Control reaches end of non-vo...
  Control reaches end of non-vo...
  Control reaches end of non-vo...
  Control reaches end of non-vo...
PTZDriverInterface.cpp
  Unused variable '*'
  Unused variable 'stream'
socket
2 issues
GridStrategy.h
  No return statement in functio...
  No return statement in functio...
utility.h
  '*' declared 'static' but never d...
```

```
#tracks: 0 #inliers: 76 FPS Tracking:16.3
#tracks: 0 #inliers: 57 FPS Tracking:14.9
#tracks: 0 #inliers: 131 FPS Tracking:17.4
#tracks: 0 #inliers: 125 FPS Tracking:15.2
#tracks: 0 #inliers: 61 FPS Tracking:14.3
#tracks: 0 #inliers: 80 FPS Tracking:17.6
#tracks: 0 #inliers: 80 FPS Tracking:17.3
#tracks: 0 #inliers: 96 FPS Tracking:17
#tracks: 0 #inliers: 111 FPS Tracking:16.9
#tracks: 0 #inliers: 112 FPS Tracking:18.1
#tracks: 0 #inliers: 107 FPS Tracking:15.4
#tracks: 0 #inliers: 83 FPS Tracking:16.9
#tracks: 0 #inliers: 118 FPS Tracking:11.9
#tracks: 0 #inliers: 114 FPS Tracking:14.8
#tracks: 0 #inliers: 114 FPS Tracking:13.9
#tracks: 0 #inliers: 91
```

Tracking at work



(x=668, y=540) ~ R:146 G:158 B:198

Unused variable '*'
Unused variable 'stream'

SonyPTZDriver.cpp
- No return statement in functio...
- Control reaches end of non-vo...
- Control reaches end of non-vo...

GridStrategy.h
- No return statement in functio...

utility.h
- '*' declared 'static' but never d...

```
e/PTZ-MTT/data/logomicc2.png
codebase/PTZ-MTT/data/logo_micc2_small.png

base/PTZ-MTT/data/mosaic_outdoor/

raphy
se/PTZ-MTT/data/pano84.txt

//path where to load dataset . If offline is commented out, the tracker starts online.
#offline =

#recorder =

//key string id for the refrence plane image
homekey = -6,6,0



//homography Hw: ref plane -> rectified plane
homoWorld = /Users/glisanti/Documents/SVN-codebase/PTZ-MTT/data/homoWorld84.txt
```

(x=2, y=16) ~ R:75 G:195 B:7

(x=258, y=26) ~ L:255

```
#tracks: 0 #inliers:
#tracks: 1 #inliers: 135 FPS Tracking:9.28
```

AdB

A Few Publications

- A. Del Bimbo, F. Dini, G. Lisanti, and F. Pernici, “Exploiting Distinctive Visual Landmark Maps in Pan-Tilt-Zoom Camera Networks,” Computer Vision and Image Understanding (CVIU), vol. 114, iss. 6, pp. 611-623, 2010.
- A. Del Bimbo, G. Lisanti, I. Masi, and F. Pernici, “Continuous Recovery for Real Time Pan Tilt Zoom Localization and Mapping,” in 2011 IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS 2011), Klagenfurt, Austria, 2011.
- A. Del Bimbo, G. Lisanti, I. Masi, and F. Pernici, “Person Detection using Temporal and Geometric Context with a Pan Tilt Zoom Camera,” in Proc. of International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 2010.
- A. Del Bimbo, F. Dini, G. Lisanti, and F. Pernici, “Sensor fusion for cooperative head localization,” in Proc. of International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 2010.
- A. Del Bimbo, G. Lisanti, I. Masi, and F. Pernici, “Device-Tagged Feature-based Localization and Mapping of Wide Areas with a PTZ Camera,” in Proc. of CVPR Int.’l Workshop on Socially Intelligent Surveillance and Monitoring (SISM), San Francisco, CA, USA, 2010.
- A. Del Bimbo, G. Lisanti, and F. Pernici, “Scale Invariant 3D Multi-Person Tracking using a Base Set of Bundle Adjusted Visual Landmarks,” in Proc. of ICCV Int’l Workshop on Visual Surveillance (VS), Kyoto, Japan, 2009.