Person Tracking and Event Recognition over A Sensor Network

Queen's University Belfast

CSIT
CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Persona de seguimiento y reconocimiento de eventos en una red de sensores

Queen's University Belfast

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Overview

- Motivation & Introduction
- Tracking over a sensor network
- Gender Profiling
- Multi agent surveillance architecture
- Conclusion & Summary

Secure Corridors

- "CCTV was originally seen as a preventative measure. Billions of pounds has been spent on kit, but no thought has gone into how the police are going to use the images and how they will be used in court. It's been an utter fiasco: only 3% of crimes were solved by CCTV. There's no fear of CCTV. Why don't people fear it? [They think] the cameras are not working."

  Detective Chief Inspector Mick Neville, officer in charge of the Metropolitan police unit, talking at the Security Document World Conference in London.

Secure Corridors

- "I want public transport to become world renowned for its safety and to banish the sad minority of hoodlums and trouble-makers that have blighted our buses."
  Boris Johnson, Mayor of London

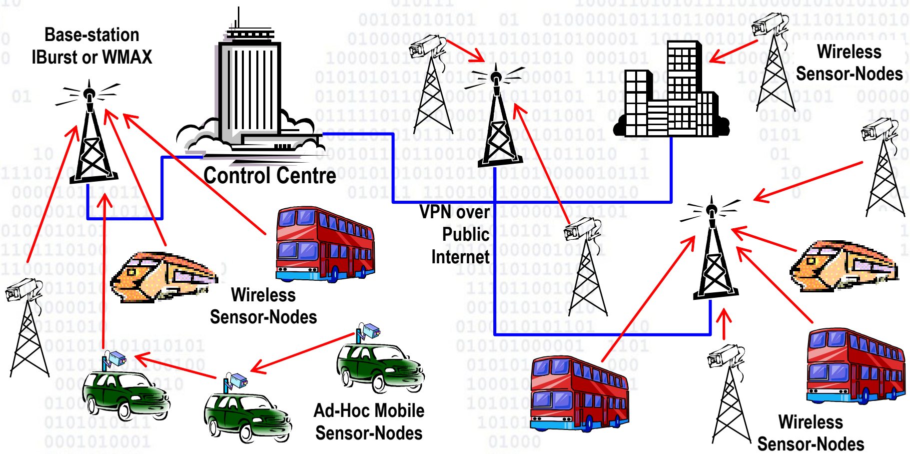
Introduction

- Massive investment in CCTV in the UK
- Impact on anti-social and criminal behaviour has been minimal
  - e.g. assaults on public transport
- CCTV operates in a passive mode
- “Active” CCTV has to alert security analysts to prevent undesirable behaviour
- Greatly increase the likelihood of being caught - a major factor in crime prevention
- Persistent analysis of CCTV video footage in real-time

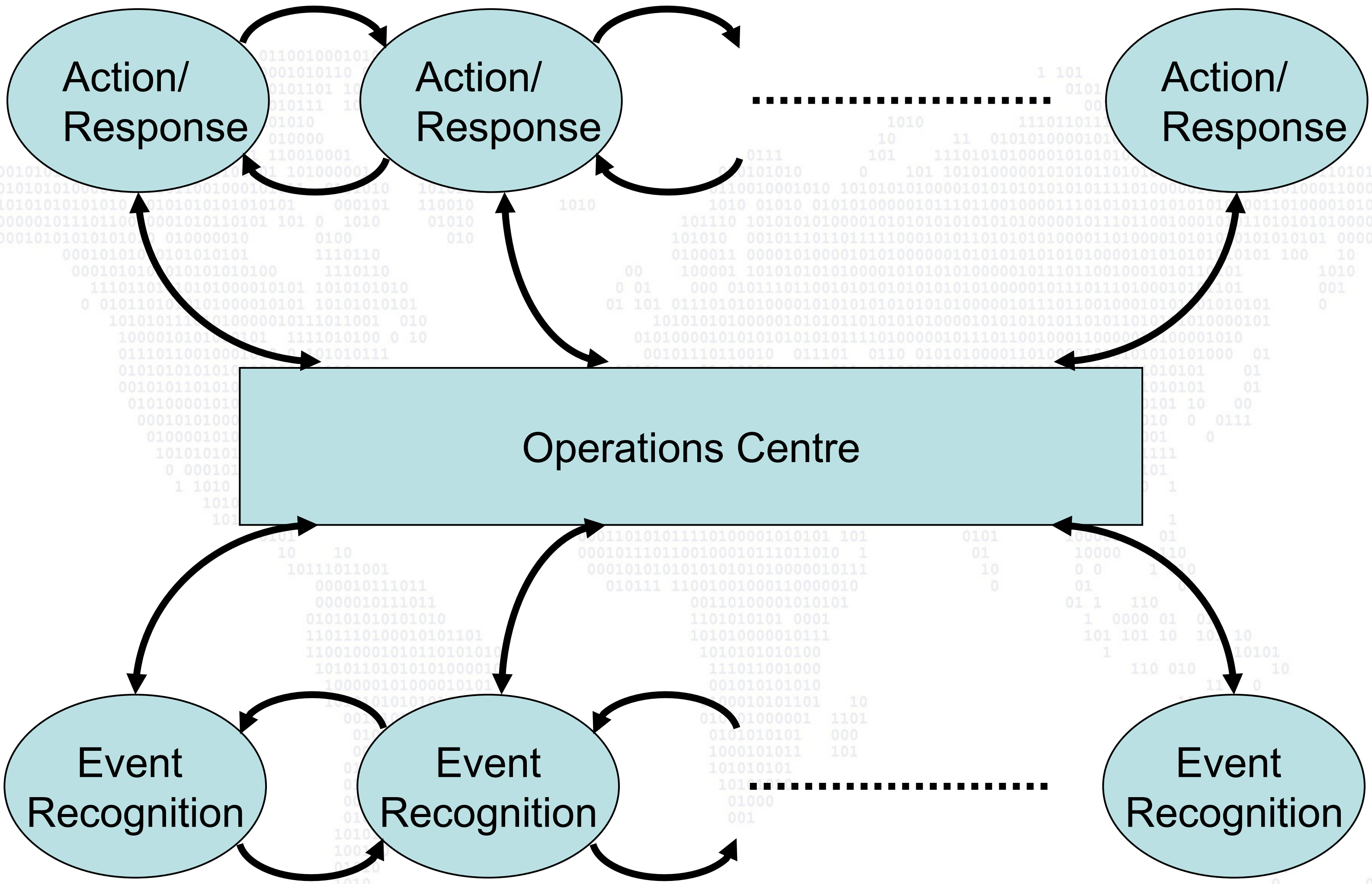
Secure Corridors – CCTV for buses

CENTRE FOR SECURE INFORMATION TECHNOLOGIES

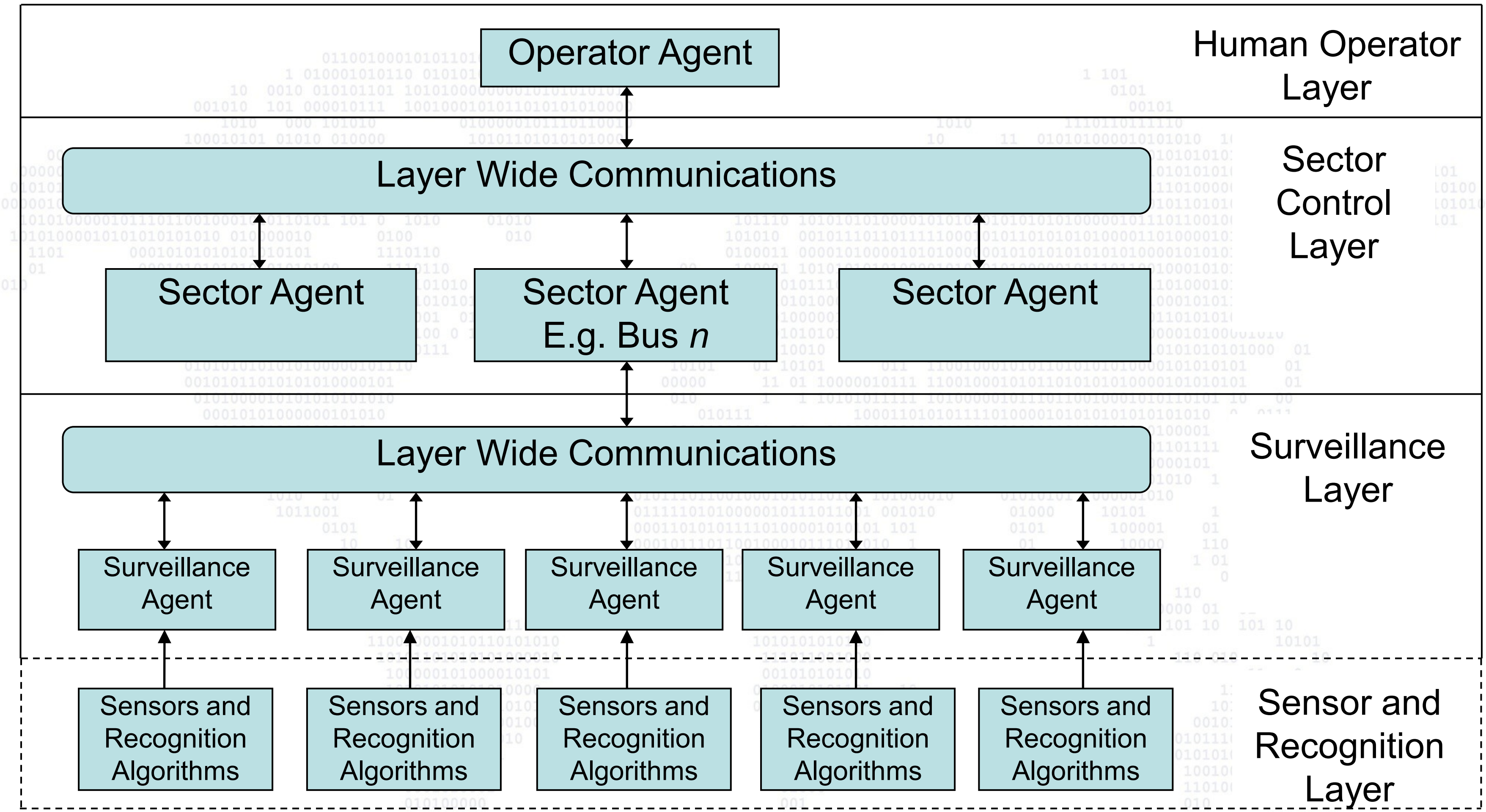
- Crime on transport platforms is a major anti-social form of crime in UK
- Aim is to reduce bus driver and passenger assaults



Event Management & Response



Architecture for Multi-Agent Surveillance



Event Recogniton

- Key requirement for active CCTV is to automatically determine the threat posed by each individual
- Focus of the computer vision community has been on behavior/action recognition
- Experienced security analysts profile individuals in the scene to determine their threat

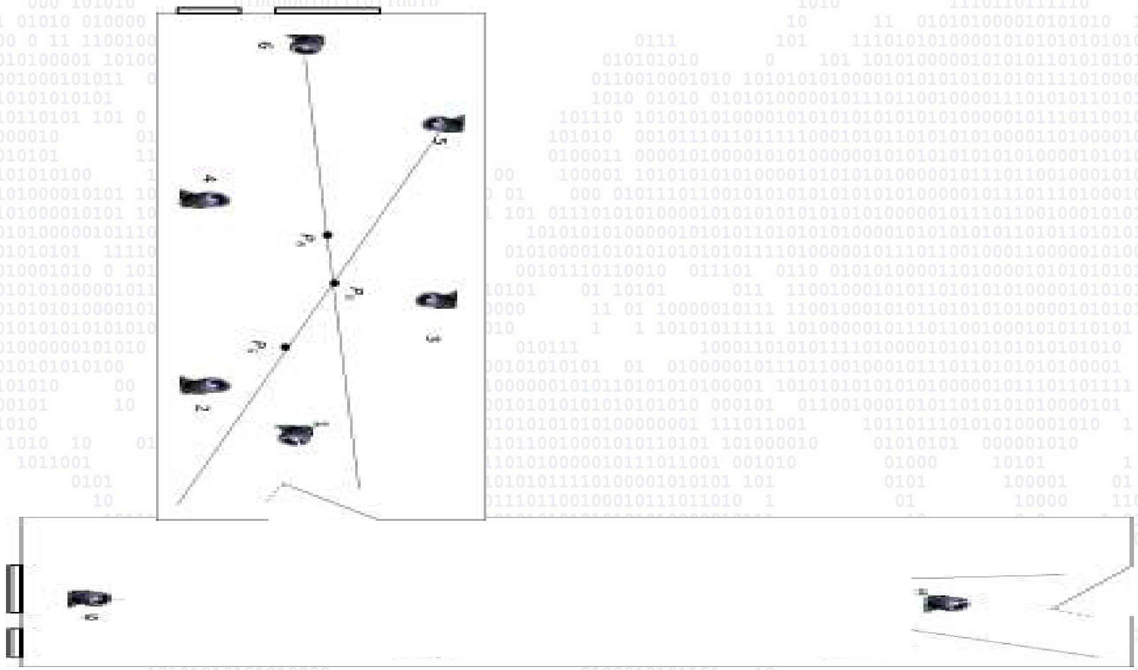
Event Recognition

- They identify individuals who look as though they may cause trouble
- Vast majority of offenders are young adolescent males
- Key to automatic threat assessment is:
  - to automatically measure the relative locations and motions of subjects in the scene
  - to automatically profile people in the scene based on their gender and age

Overview

- Introduction & Motivation
- Tracking over a sensor network
- Gender Profiling
- Multi agent surveillance architecture
- Conclusion & Summary

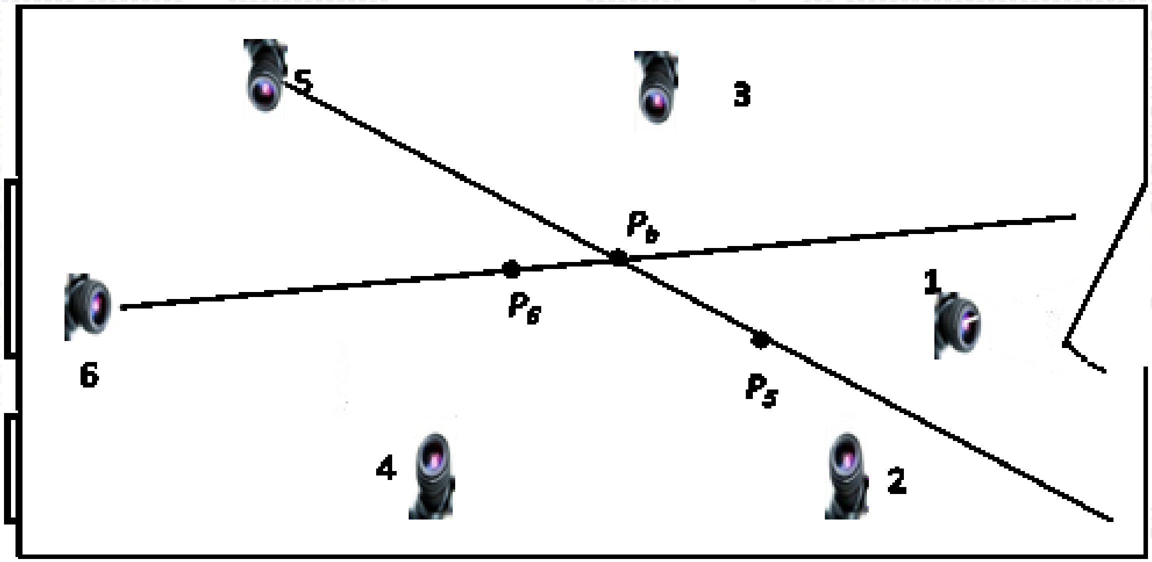
CSIT Sensor Network



CSIT Sensor Network

Open Spaces / Narrow Corridors

Laboratory

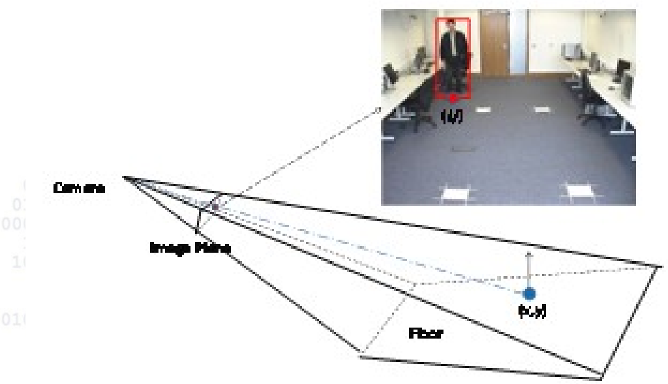


Motivation

- Moving object detection on buses likely to be poor
- Passenger obscuration by seats
- Cannot assume bottom of bounding box is coincident with floor

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Single Camera



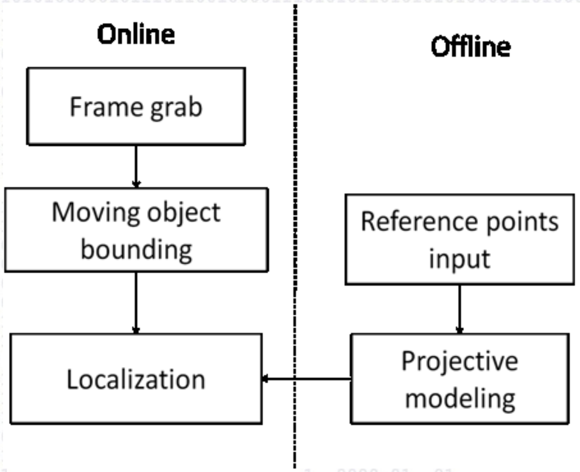
$$\begin{bmatrix} si \\ sj \\ s \end{bmatrix} = \mathbf{Q} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} \mathbf{R} & \mathbf{T} \end{bmatrix}$$

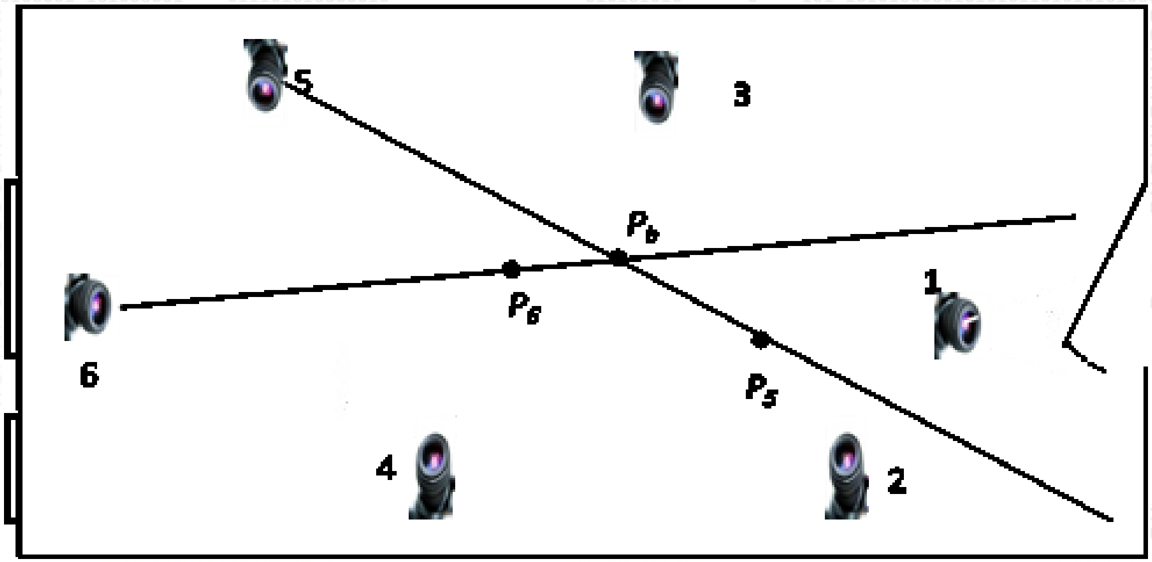
- Assumption: centre of detected object’s bottom boundary is coincident with feet
- *z* value is zero, only need to find *(x,y)* co-ordinates
- Measure positions of four key points
- Solve for the eight unknowns in **Q**

Single Camera

- Foreground extracted by background subtraction
- Bounding box placed around foreground
- Corner coordinates passed to localization module
- Solves for x and y



Two camera

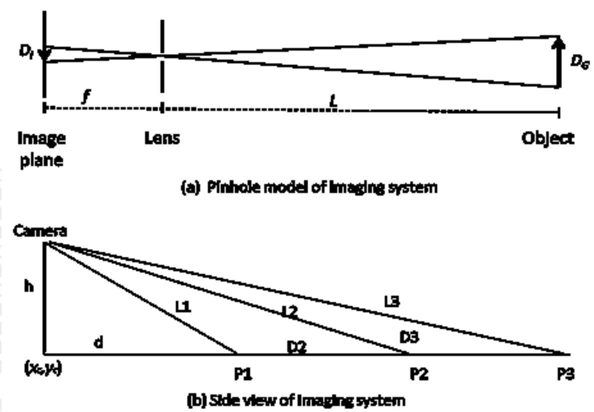


Two-camera Localisation

- For each camera determine its position in the room coordinates
- Determine the object position in the room coordinates for each camera
- Draw a line from the camera to object
- Object position is the intersection of lines

Two camera

- Select three points P_1, P_2 and P_3
- Determine L_1, L_2, and L_3
- Know D_2 and D_3
- Solve for *d*, then *P*



(a) Pinhole model of imaging system

(b) Side view of imaging system

$$\frac{D_i}{f} = \frac{D_G}{L}$$

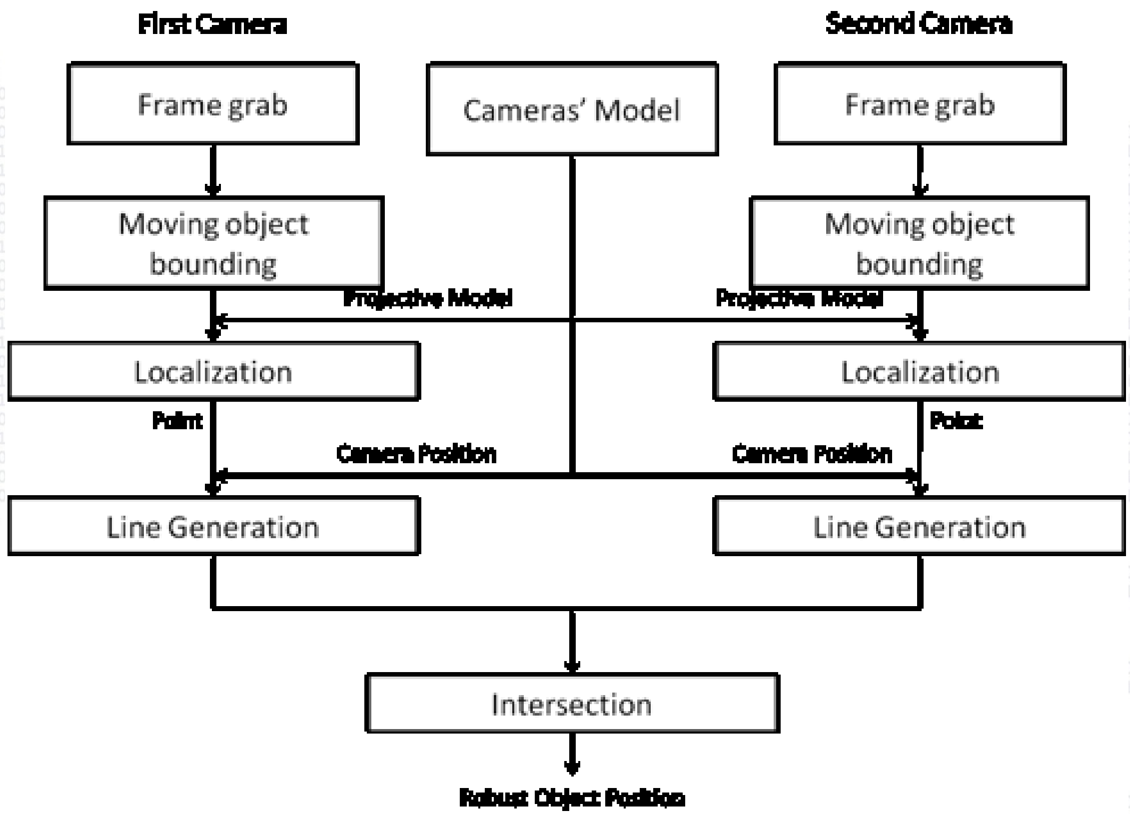
$$h^2 + d^2 = L_1^2$$

$$h^2 + (d + D_2)^2 = L_2^2$$

$$h^2 + (d + D_2 + D_3)^2 = L_3^2$$

$$\vec{P} = \vec{P_1} + d \frac{\vec{P_1} - \vec{P_2}}{|\vec{P_1} - \vec{P_2}|}$$

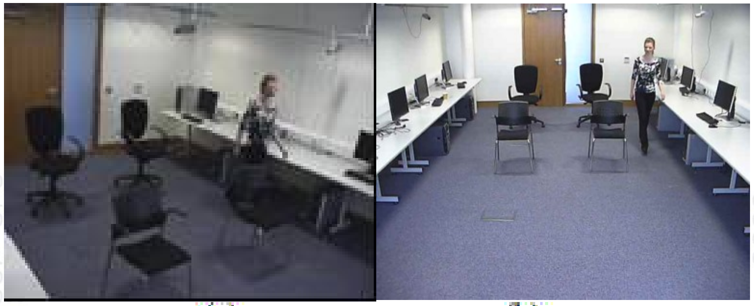
Two camera


First Camera
Second Camera
Frame grab
Cameras' Model
Frame grab
Moving object bounding
Moving object bounding
Projective Model
Projective Model
Localization
Localization
Point
Point
Camera Position
Camera Position
Line Generation
Line Generation
Intersection
Robust Object Position

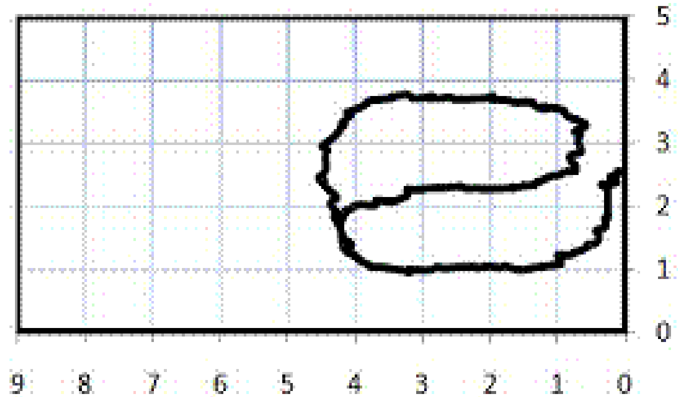
CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Experiments

- Selected cameras 5 and 6 because they had largest coverage of any camera-pair
- 11 individuals were asked to roughly follow a route in the lab
- Chairs placed in the route
- 12400 video frames were collected
- Sequences manually analysed to give ground truth of real world position in each frame



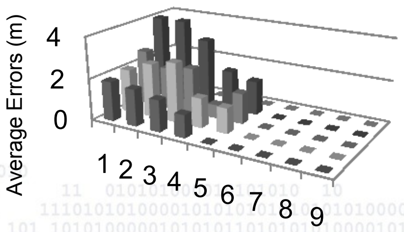
(a) (b)



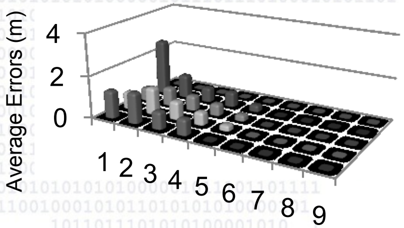
(c)

Position based error analysis

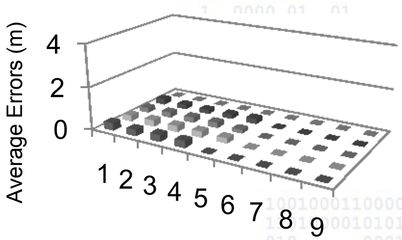
- Lab floor divided into 1m×1m blocks.
- Localization errors for each block are averaged over subjects



(a)

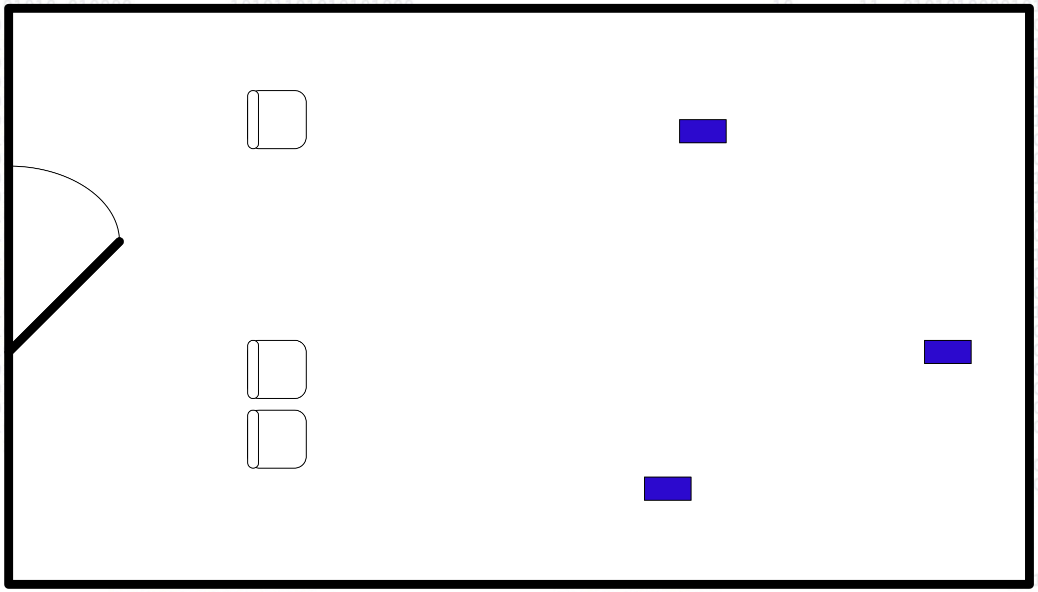


(b)



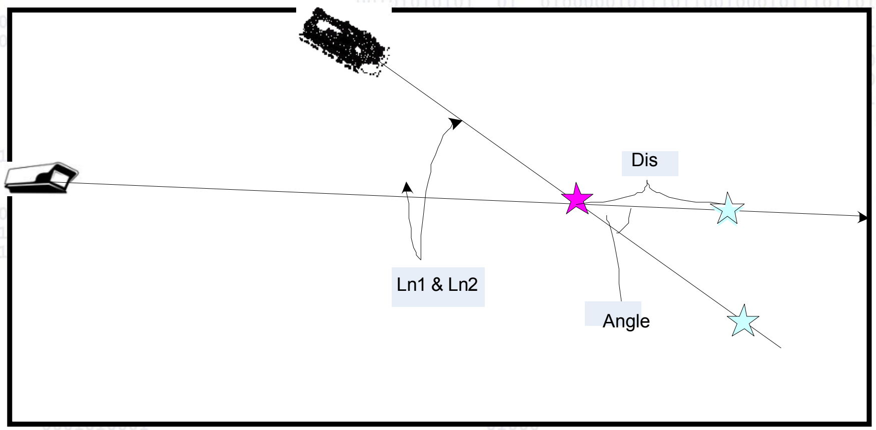
(c)

Three Camera



Intersection

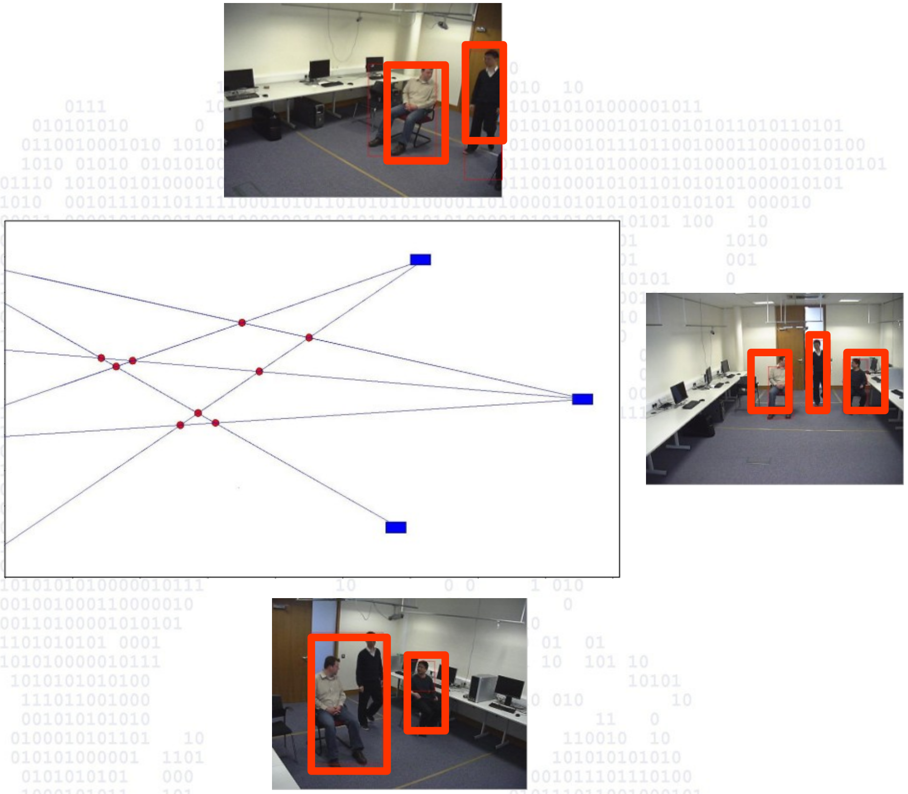
- Position: Ground floor co-ordinates, denoted by (*x,y*).
- Dis: The distance to the nearest localisation point inferred by a single camera observation.
- Ln1, Ln2: The two intersecting lines generated by the two single-camera observations.
- Angle: The angle between Ln1 and Ln2.



CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

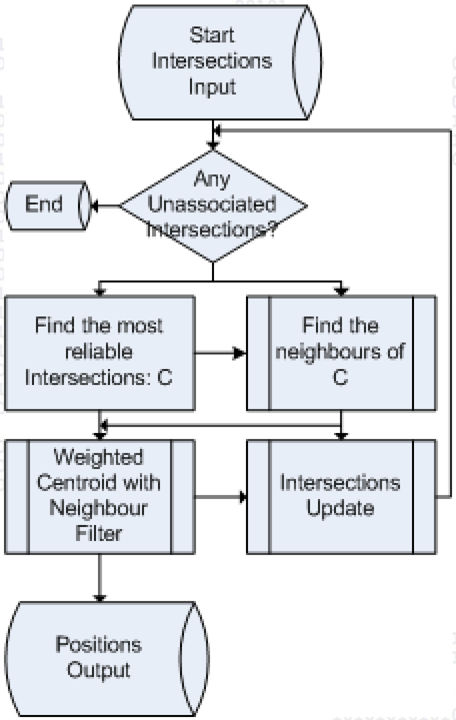
Detection Association

- Multiple cameras mean multiple detections
- Some erroneous detections due to false alarms and occlusion
- Need to associate detections within subjects



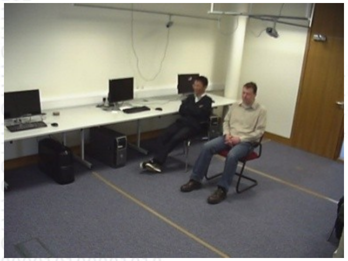
Detection Association

- Reliability based upon
  - Cosine of angle
  - Min Dis
  - Combination of both: If min Dis <T then cos
- Neighbours are detections within radius r of C
- Associated intersections removed from list



“Bus Journey”

- 10 min sequence of 4 subjects entering, sitting, standing and exiting
- Corresponding to 2210 frames captured by each camera
- No. of subjects in each frame, summed over whole sequence, was 3870



Evaluation

$$T = \sum_i T_i w_i$$

$$T_i = \begin{cases} 1 & \text{if } A_i > S_i \\ A_i / S_i & \text{otherwise} \end{cases}$$

$$F = \sum_i F_i w_i$$

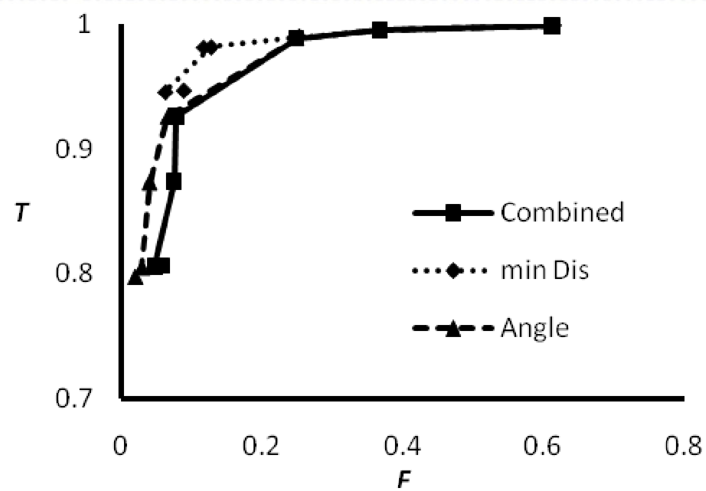
$$F_i = \begin{cases} \dfrac{A_i - S_i}{I_i - S_i} & \text{if } A_i > S_i \\ 0 & \text{otherwise} \end{cases}$$

$$w_i = \frac{S_i}{\sum_i S_i}$$

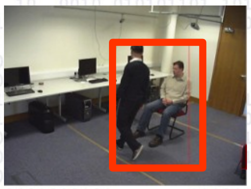
- T_i is true positive rate
- S_i is the actual no. of subjects
- A_i is no. of subjects output from detection association algorithm
- F_i is false positive rate
- I_i is input no. of intersections

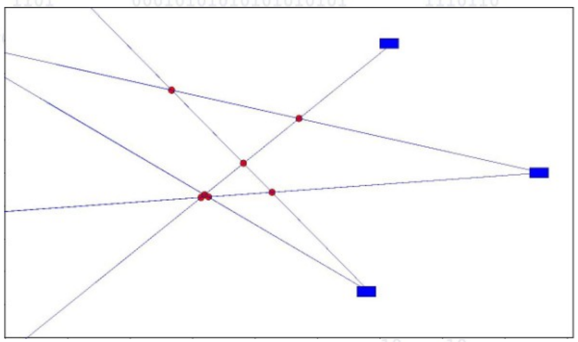
Evaluation

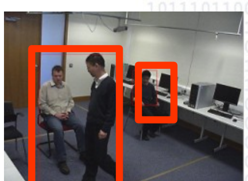
- Tis a measure of subject underestimation
- F is a measure of overestimation
- Vary radius *r* to obtain ROC curves
- *r* varies from 10cm to 2m left to right

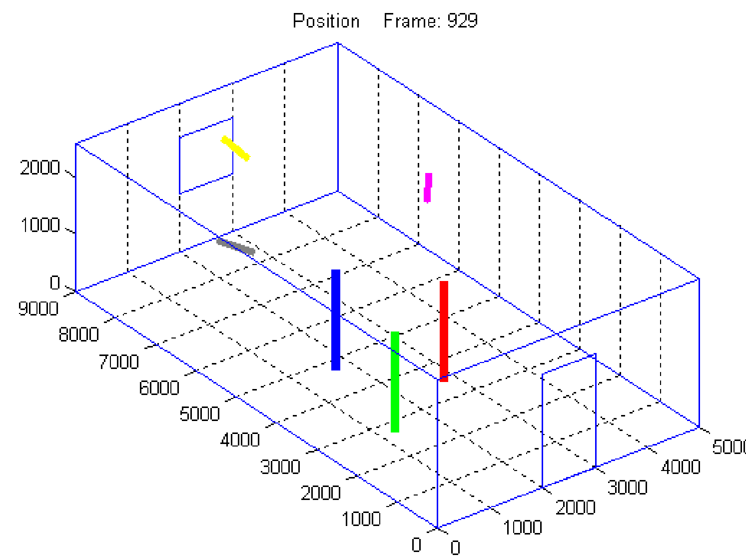


Correct Estimation

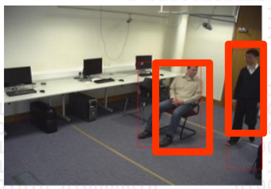


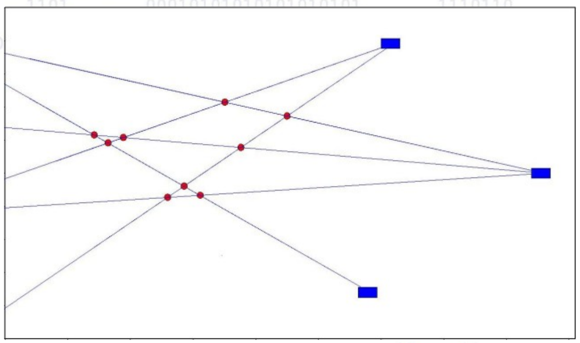


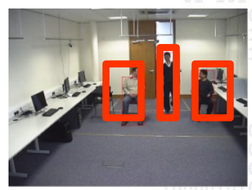


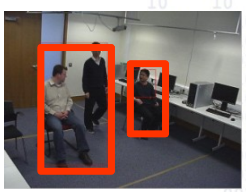


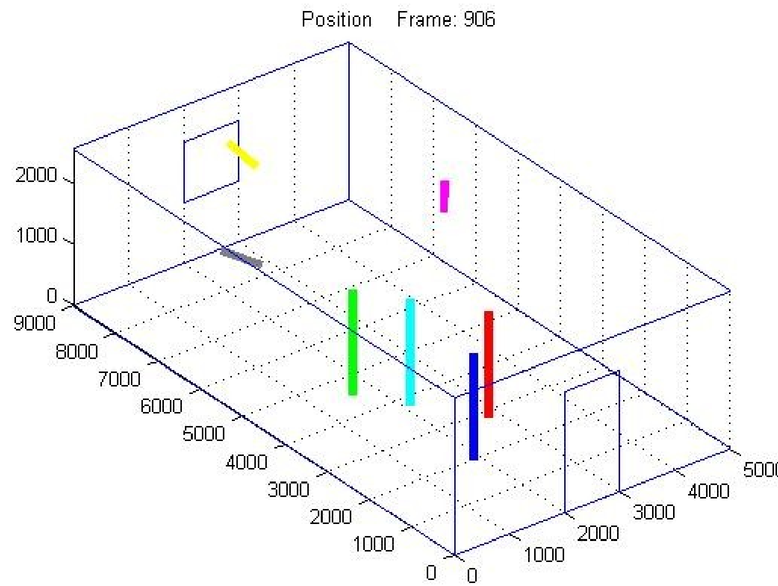
Over Estimation



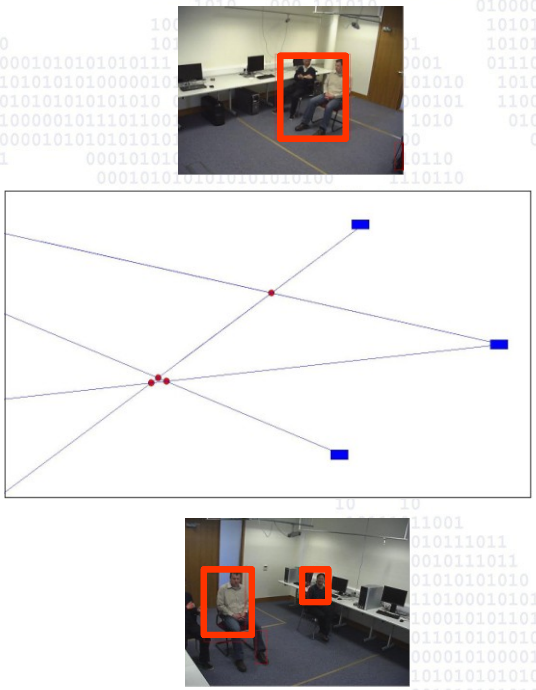


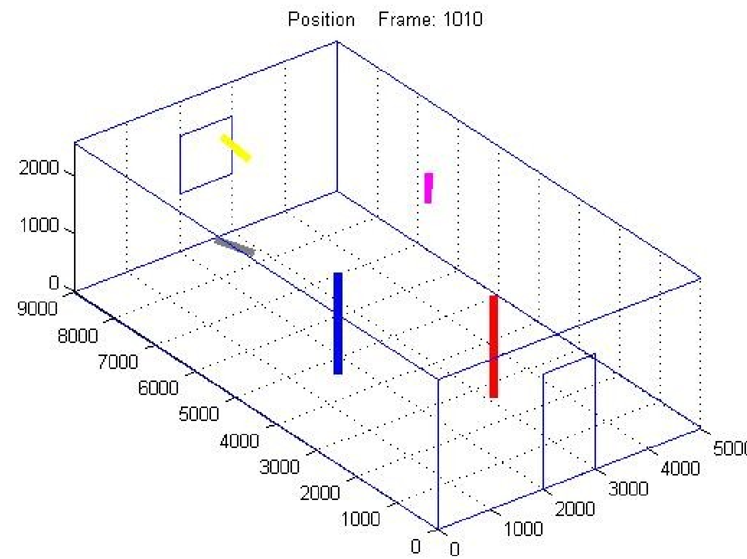






Under Estimation





Tracking

- Bayesian framework
  - Prediction
  - Filter
- Cannot be evaluated for most state-space models

$$p\left(x_t \mid z_{1:t-1}\right) = \int p\left(x_t \mid x_{t-1}\right) p\left(x_{t-1} \mid z_{1:t-1}\right) dx_{t-1}$$

$$p\left(x_t \mid z_{1:t}\right) = \frac{p\left(z_t \mid x_t\right) p\left(x_t \mid z_{1:t-1}\right)}{\int p\left(z_t \mid x_t\right) p\left(x_t \mid z_{1:t-1}\right) dx_t}$$

$$\hat{x}_t = \arg\max_{x_t} p\left(x_t \mid z_{1:t}\right)$$

Particle filter

- Represent posterior by set of random samples and weights
- Weights are updated according to observation likelihood
- Sequential importance sampling

$$p(x_t) \approx \frac{1}{N_p} \sum_{i=1}^{N_p} \delta\left(x_t - x_t^i\right)$$

$$p\left(x_t \middle| z_{1:t}\right) \approx \sum_{i=1}^{N_p} w_t^i \delta\left(x_t - x_t^i\right)$$

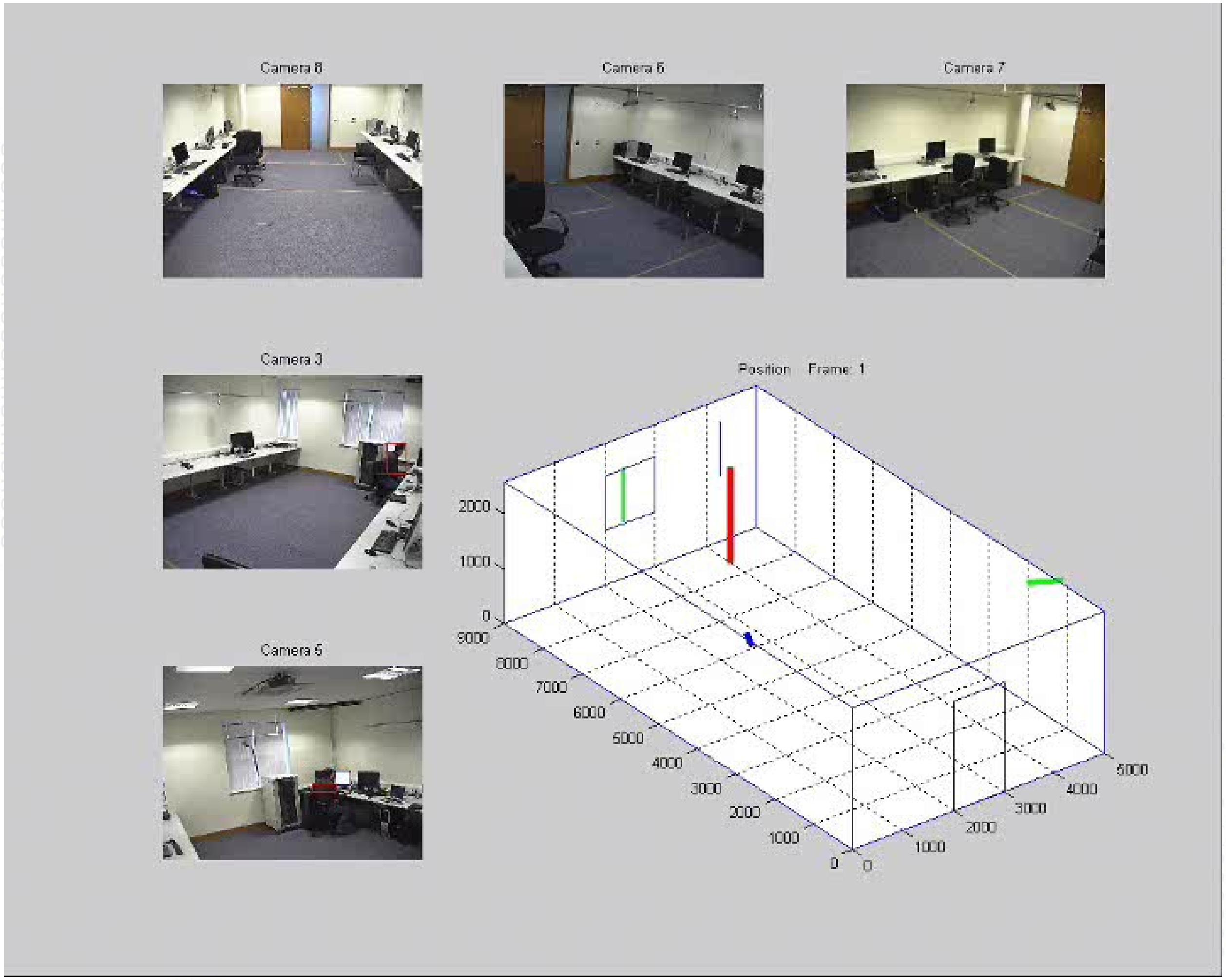
$$w_t^i \propto w_{t-1}^i \frac{p\left(z_t \middle| x_t^i\right) p\left(z_t \middle| x_t^i\right)}{q\left(x_t^i \middle| x_{t-1}^i, z_t\right)}$$

Kalman particle filter

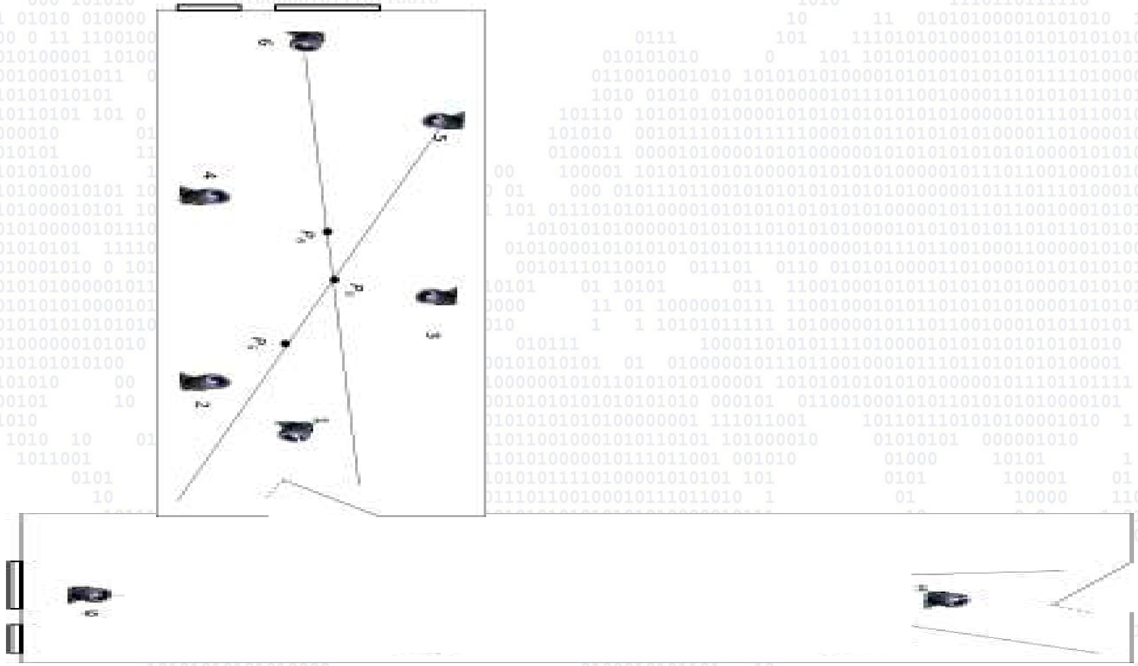
- Kalman filter used to propagate each particle
- Steers particles towards regions with high likelihoods
- Fewer particles, less computation

$$\bar{x}_t = Ax_{t-1} + K\left(y_t - H\right.$$

Tracking



CSIT Sensor Network



Subject Reacquisition

- Camera field of views are non-overlapping requires subject reacquisition for tracking
- Subject reacquisition is identification through applied detection, tracking and learning.
- Association of a current observed object with a previously observed object.
- Time gap could be seconds, hours, days etc.

Subject Reacquisition

- Online principal component analysis
- Ten components are learnt
- Temporal voting used for reacquisition

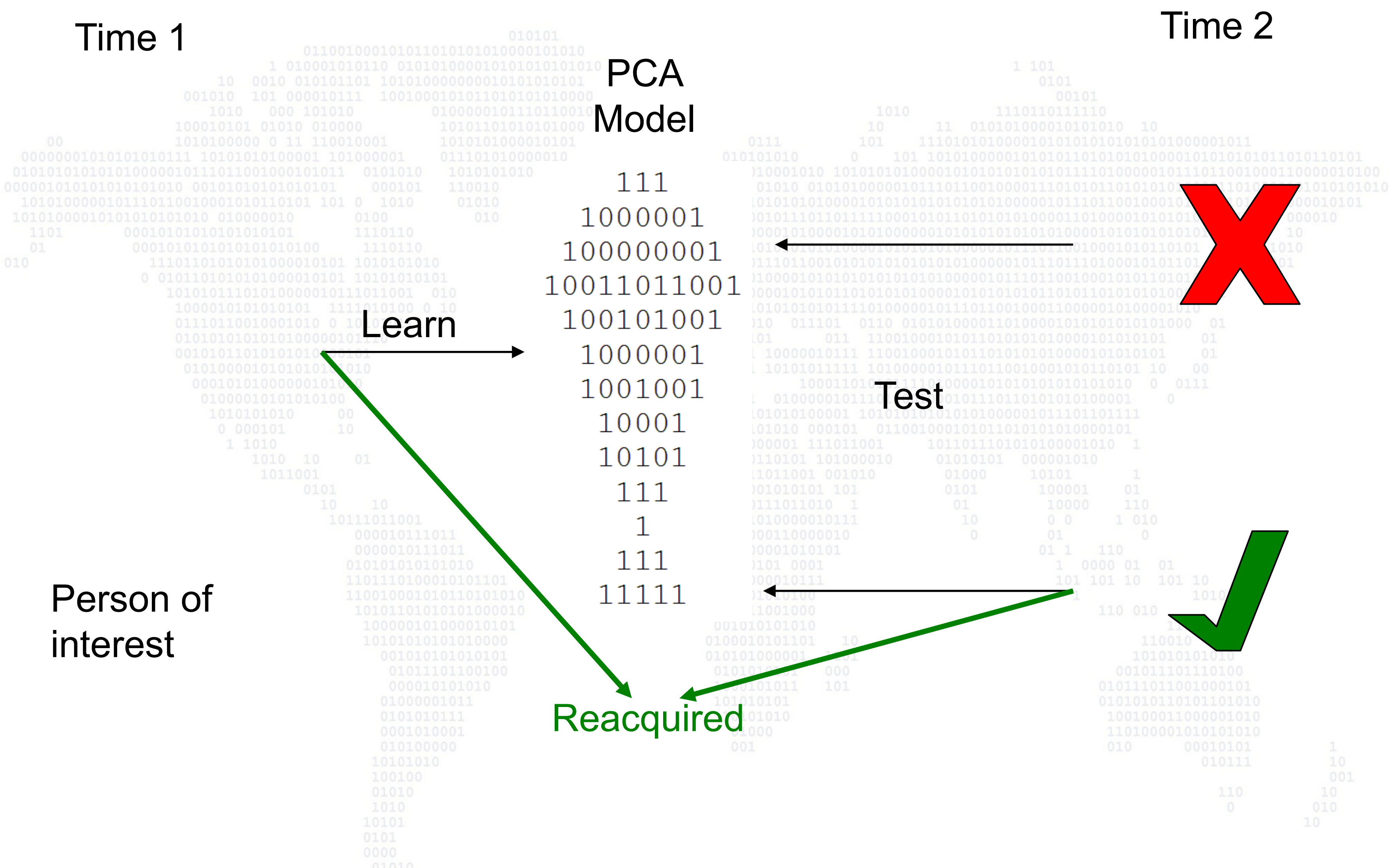
$$\mathbf{v}(0)=\mathbf{x}(1)$$

$$\mathbf{v}(t)=\frac{n-1}{n}\mathbf{v}(t-1)+\frac{1}{n}\mathbf{x}(t)\mathbf{x}^{T}(t)\frac{\mathbf{v}(t-1)}{\left\|\mathbf{v}(t-1)\right\|}$$

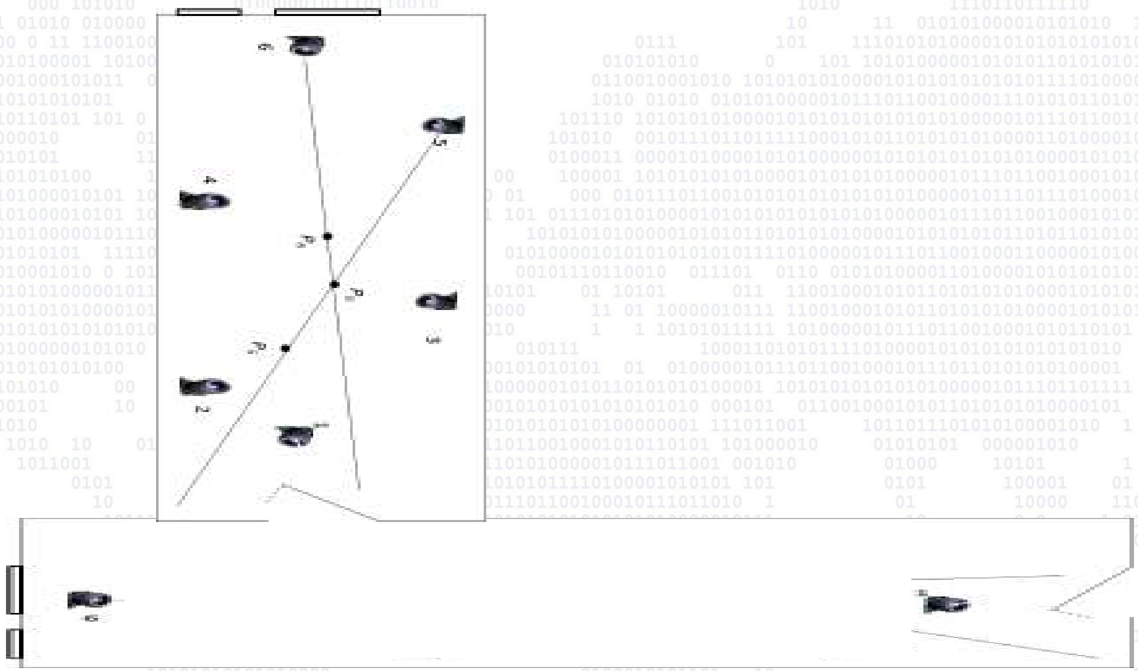
$$\mathbf{x}_2(t)=\mathbf{x}(t)_1-\mathbf{x}_1^{T}(t)\frac{\mathbf{v}_1(t)}{\left\|\mathbf{v}_1(t)\right\|}$$

$$S(\mathbf{x}_t)=\|\tilde{\ }\qquad {}^{M}_{i=1}\qquad {}^{2}_{i}$$

Subject Reacquisition



CSIT Sensor Network



Reacquisition Video

- 8 subjects at first camera
- Same 8 subjects at second camera
- Followed by 14 new subjects

True Reacquisition (TP)	False 'Unknown' Decision (FN)	True 'Unknown' Decision (TN)	False Reacquisition (FP)
6 (75%)	2 (25%)	9 (64%)	5 (36%)

Overview

- Introduction & Motivation
- Tracking over a sensor network
- Gender Profiling
- Multi agent surveillance architecture
- Conclusion & Summary

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Face-based Gender Profiling

Current image → Face Detection → Principal Component Analysis → Support Vector Machine

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

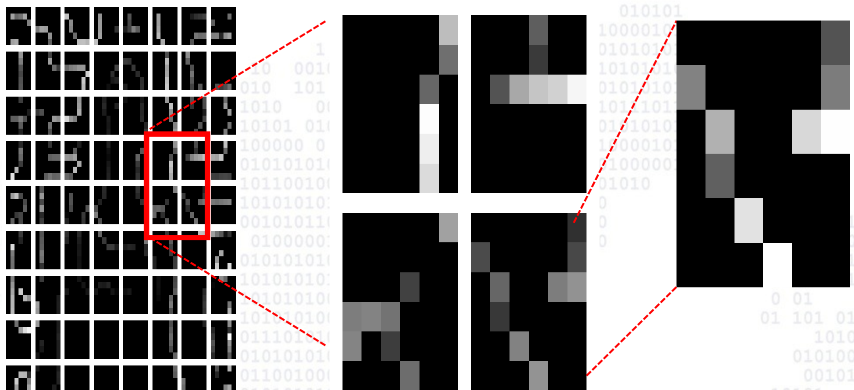
Body-based Gender Profiling

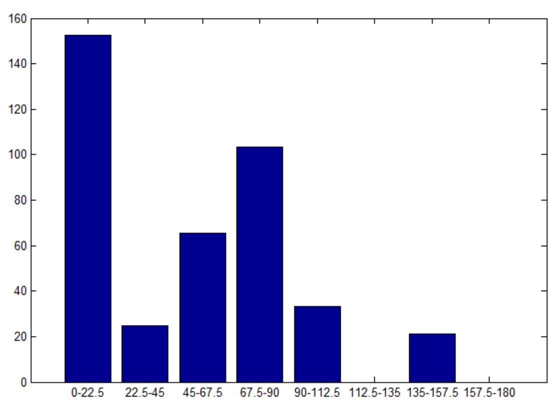
- Profiling at-a-distance
- Full body
- Based on clothes colour

Body-based Gender Profiling

- 128x48 pixel image
- Calculate Canny edge image
- Calculate gradient image then multiply with Canny

Canny Histogram of Gradients (CHoG)

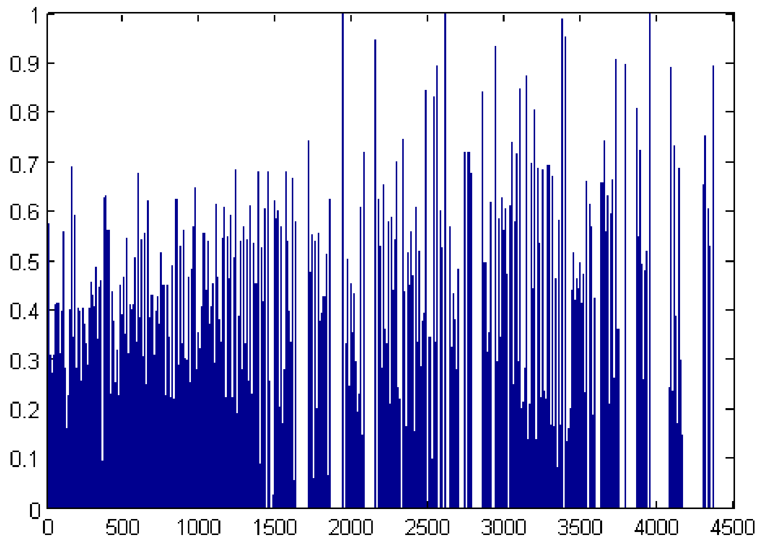




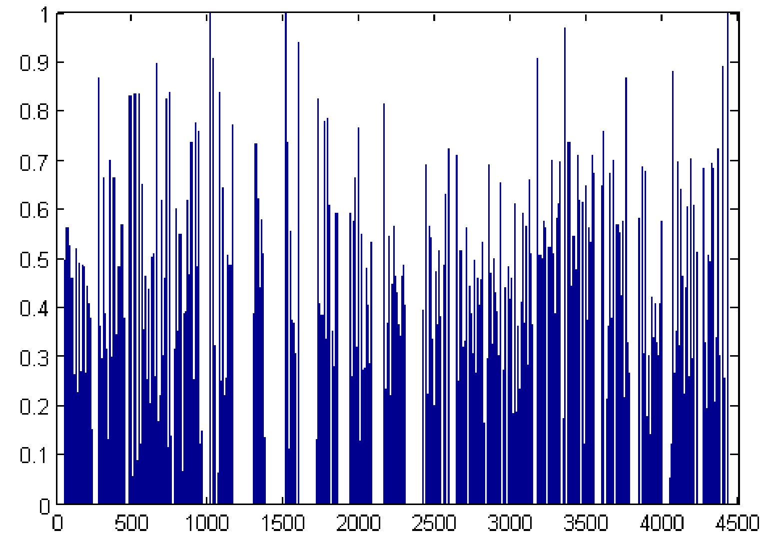
- 21x8 grid of cells of 6x6 pixels
- Histogram of cell edge orientations with 8 bins
- Bin value is sum of intensity gradients
- Cells grouped into overlapping blocks of 2x2
- Blocks overlap to give140 (20x7) blocks.
- Concatenated HoGs from each block gives 4480 element feature vector

CHoG

Average Female



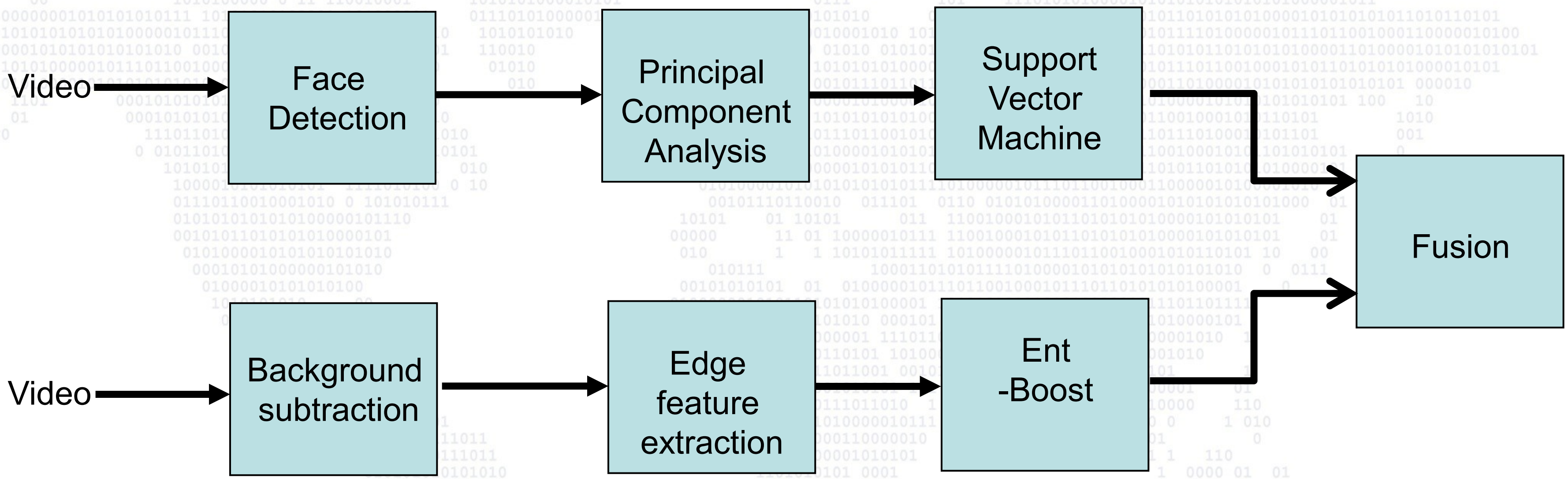
Average Male



Classifier

- Dataset of 413 images for each gender
  - MIT pedestrian recognition set
  - Viewpoint invariant pedestrian recognition set (VIPeR)
- Entropy Boost classifier
- Five-fold validation (80% training, 20% testing)
- 81% correct recognition

Gender Profiling



Experimental Results

- Five sequences consisting of one or more subjects.
- Video resolution is 640×480 and frame rate is 10 fps (Panasonic WV-LZ62 camera).

Types	Total images	Pedestrians	Genders	Positive detections
Single-1	763	1	Female	672
Single-2	854	1	Male	807
Single-3	1013	1	Male	755
Single-4	1015	1	Female	870
Multiple-1	805	2	Female/male	730
Multiple-2	623	2	Male/male	571

Experiments

- EB-Fusion approach is compared against:
  - Face based gender classification using PCA coefficients with SVM (FACE-PCA)
  - Full body based gender classification using HOG features with SVM (BODY-HOG)
  - Concatenated HOG features of face and full body components with SVM (CP-FB) .

Types	EB-FUSION	CP-FB	FACE-PCA	BODY-HOG
Single-1	8.2%	16.4%	67.6%	17.8%
Single-2	9.1%	14.4%	43.2%	16.3%
Single-3	9.5%	13.2%	95.6%	14.0%
Single-4	10.1%	14.1%	94.4%	15.5%
Multiple-1	11.6%	15.4%	95.2%	16.3%
Multiple-2	9.0%	13.3%	95.9%	14.2%

Experimental Results

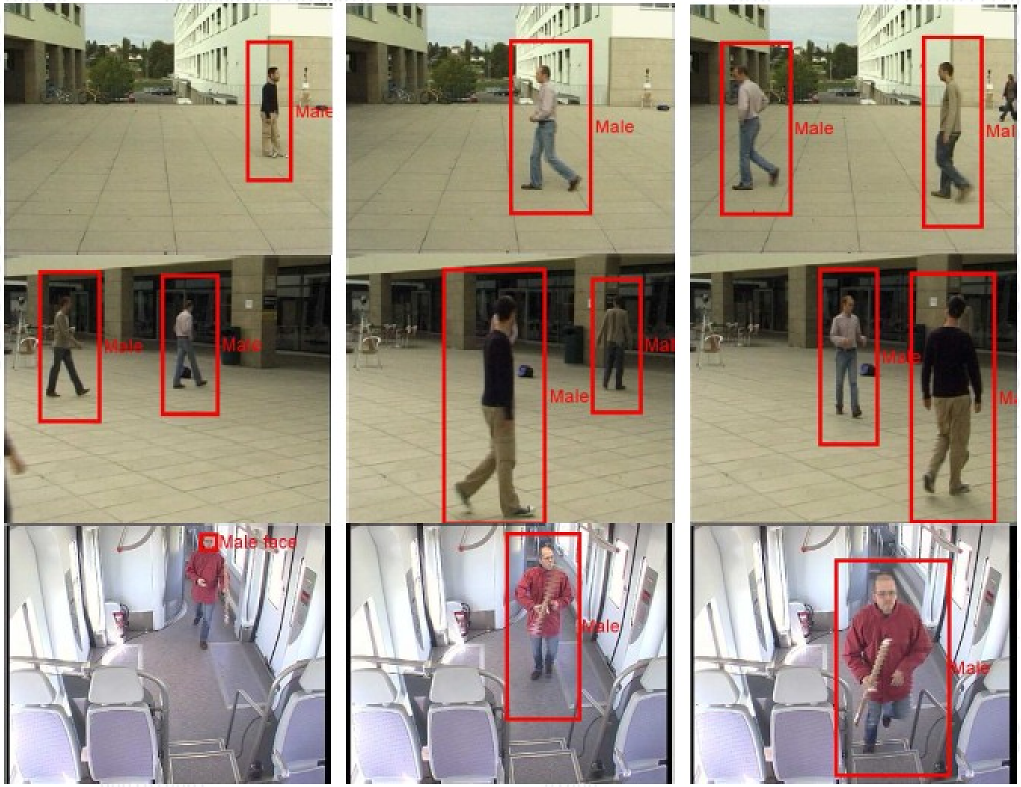
Table 1 General information of the publicly accessible videos in Experiment 3.

Types	Total images	Pedestrians	Gender	Positive detections
Campus-1	2000	> 3	Male/female	1613
Campus-2	2000	> 3	Male/female	1554
Train-1	1626	> 3	Male/female	1019

Table 2 Error statistics of gender classification in Experiment 3.

Types	EB-FUSION	CP-FB	FACE-PCA	BODY-HOG
Campus-1	16.4%	18.2%	99.6%	19.8%
Campus-2	15.1%	17.4%	99.2%	19.1%
Train-1	14.7%	19.8%	76.3%	19.6%

Results

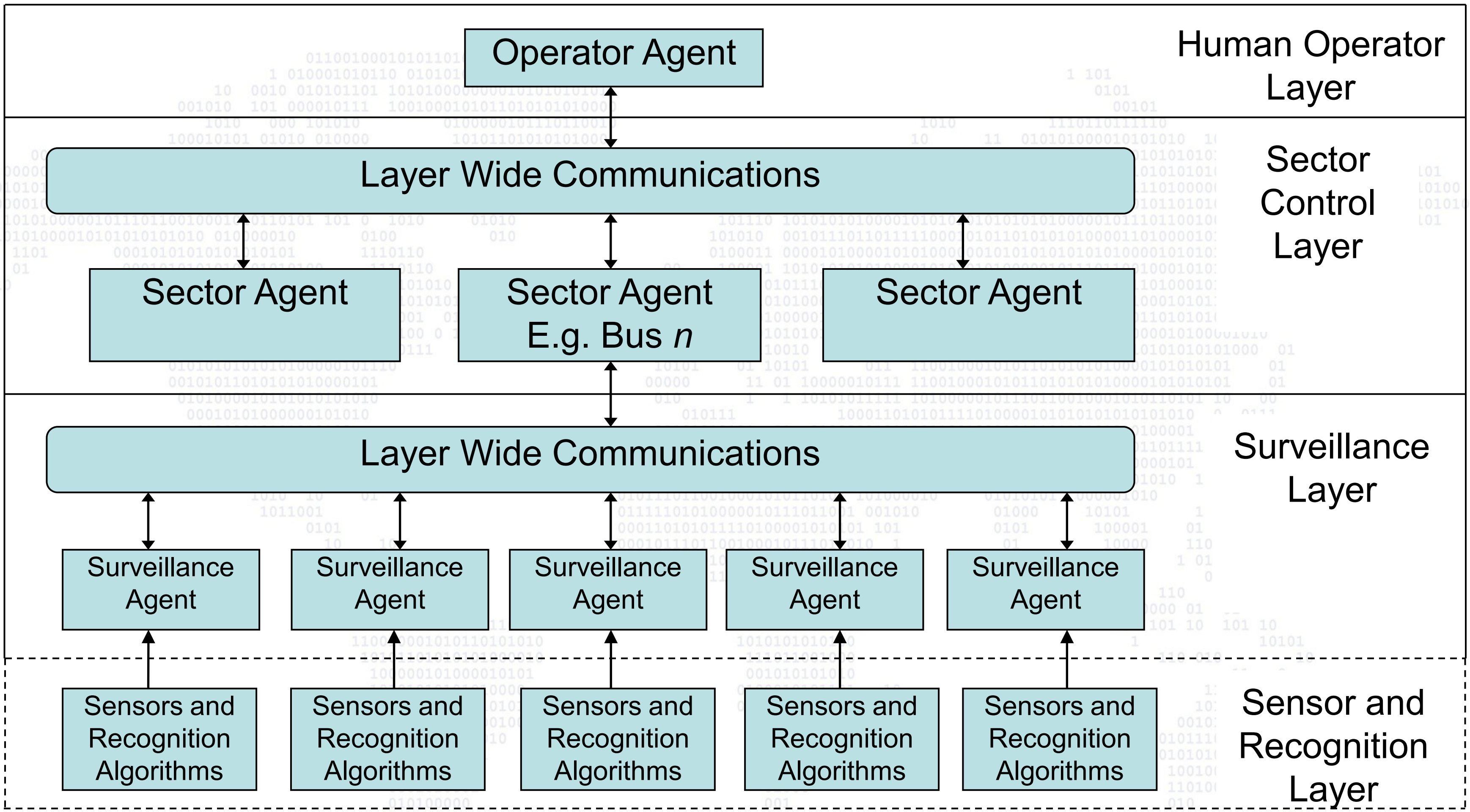


Gender Profiling

Overview

- Introduction & Motivation
- Tracking over a sensor network
- Gender Profiling
- Multi agent surveillance architecture
- Conclusion & Summary

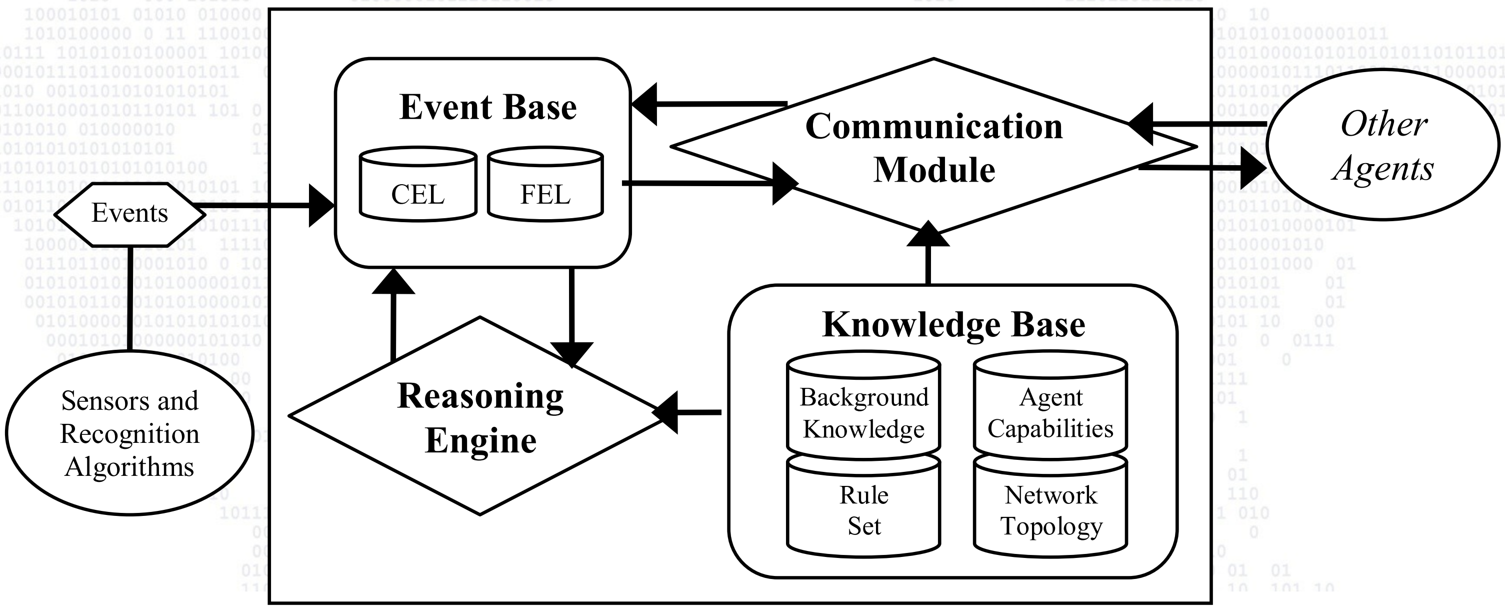
Architecture for Multi-Agent Surveillance



Event Composition

- Third step in threat assessment is to combine the who's in the scene with the where in the scene
- Event management framework
- Implemented using multi-agent architecture

Surveillance Agent



Event Representation

- Passenger boarding
- e_1=(*PBV,21:05:31,1,0.85,0.7,{male}*)
- *PBV* is for event type *Person Boarding Vehicle*
- *21:05:31* is the time of occurrence *occurT*
- *1* denotes the source, in this case a video analytics algorithm
- *0.85* is the source reliability
- *0.7* denotes the significance
- *male* is the value for the gender attribute

Knowledge Base

- Rule 1: Infers the event abusive behaviour towards driver
- $R_1=(LS_1, EType_1, Condition_1, m_1)$
- $LS1=(TPL, TPL+120)$
- *EType1=DA* abbreviated for driver abuse
- *Condition is :*

  $ei.Etype = PL \wedge ej.Etype = PS \wedge ei.location = Drivers\ Cabin$

Communication Protocol

- Centralised reasoning has scalability limitations.
- Subscription-only presents a huge overhead for events that are produced frequently but are rarely needed.
- Communication protocol works on a “need to know” basis.
  - onetime event queries
  - event subscriptions

Onetime queries

- Used where an agent only needs to know about that type of event for one individual event pattern.
- Sets other events in that pattern as a trigger for a onetime request of that event
- Reduces the communication whilst still ensuring the agent has complete knowledge for reasoning.

Experiments

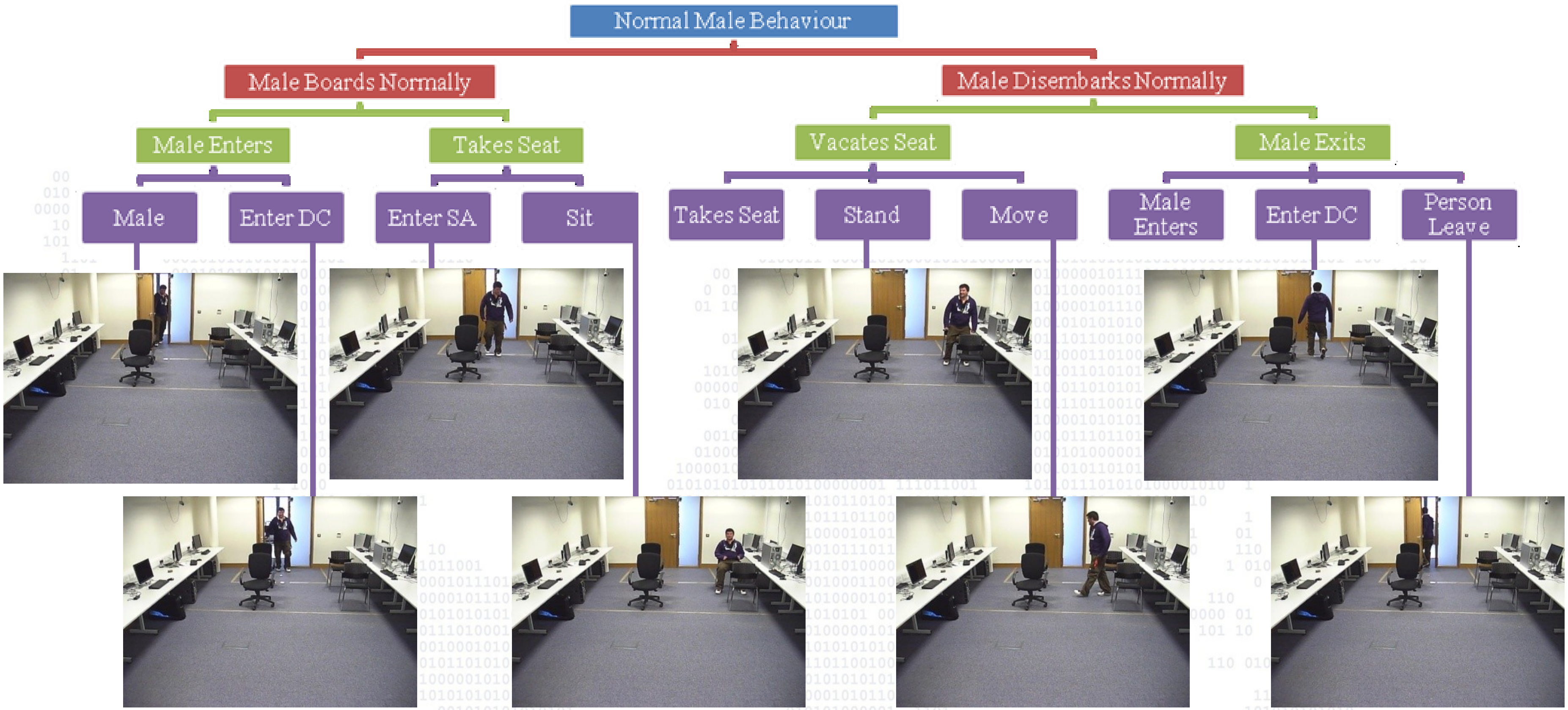
- 55 sequences in total
- 40 of normal passenger behaviour; person boards bus, sits and then exits the bus
- 15 of suspicious behaviour
  - 5 of people loitering in the saloon area
  - 5 of loitering in the driver’s cabin
  - 5 of people obscuring face from cameras

Experiments

- Lab was partitioned out into 2 sections to represent the driver's cabin and the saloon area.
- Area surrounding the door up to the first set of seats is regarded as driver's cabin
- Rest of the floor space and chairs are saloon area.
- Multi Agent System is developed using the agent middle ware JADE.
- Reasoning module for the agents and centralized reasoner developed using PROLOG.

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Normal Male Behaviour



Example of a lab simulated normal male behaviour

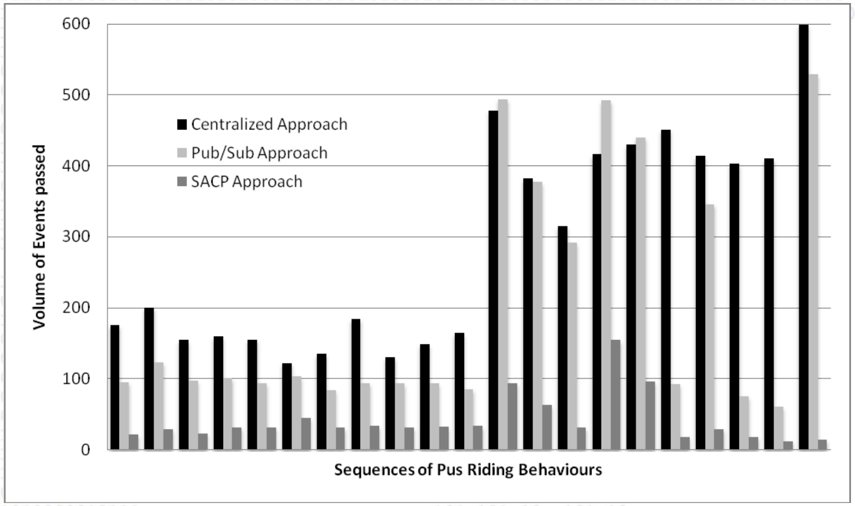
Event Composition Results

- Ground Truth (GT)
- True Positives (TP)
- False Positives (FP)
- False Negatives (FN)
- Sensitivity (S = TP/(TP+FN))
- Precision (P = TP/(TP+FP)).

Compound Event	GT	TP	FP	FN	S (%)	P (%)
Male Normal Behaviour	20	18	2	2	90%	90%
Female Normal Behaviour	20	16	2	4	80%	89%
Loitering Saloon Area	5	4	1	1	80%	80%
Loitering Drivers Cabin	5	4	1	1	83%	80%
Person Obscuring Face	5	4	0	1	80%	80%

Communication Results

- Measured events sent for reasoning
  - centralized
  - agent using publish/subscribe
  - agent using SACP



Conclusion & Summary

- Video analytics for tracking and gender profiling have been demonstrated
- Event recognition and composition for a sensor network demonstrated in laboratory conditions
- Preliminary evaluation of architecture for multi-agent surveillance
- Bus trial for further evaluation