Video object tracking

Andrea Cavallaro

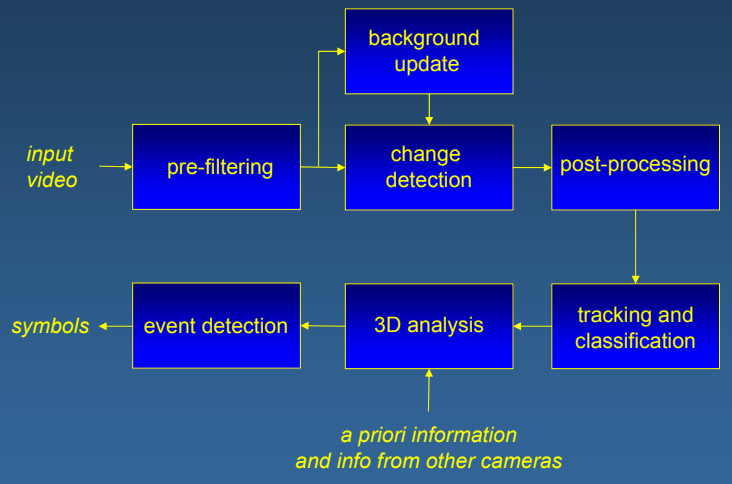
Queen Mary, University of London

andrea.cavallaro@elec.qmul.ac.uk
http://www.elec.qmul.ac.uk/staffinfo/andrea

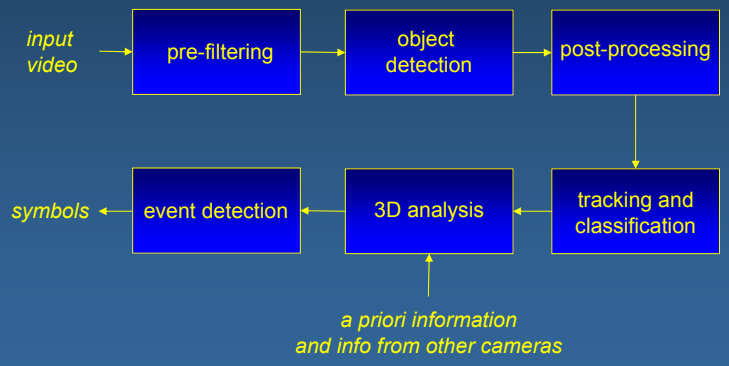
Outline

- Introduction
- Tracking algorithms
  - Mean-shift
  - Particle filtering
  - Graph-matching
- Integration of detection and tracking
- Integration of audio and video
- Event detection

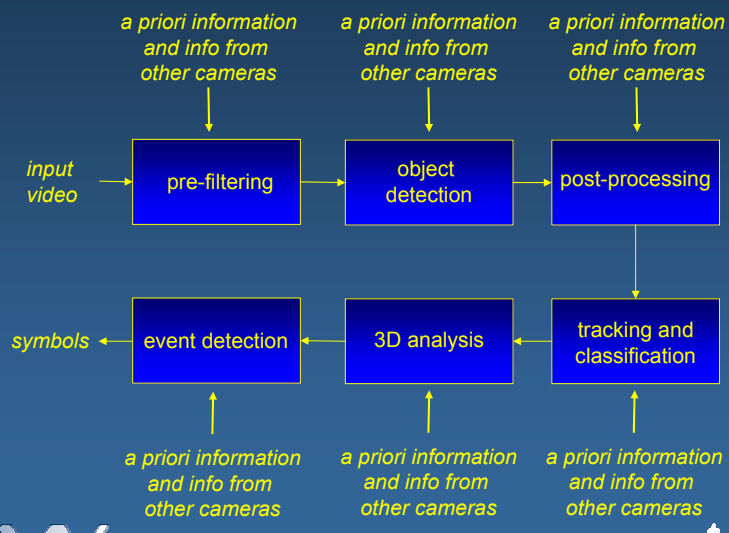
Framework



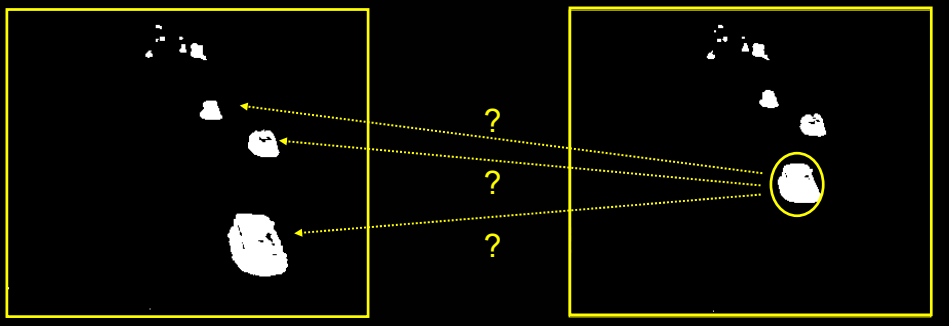
Framework



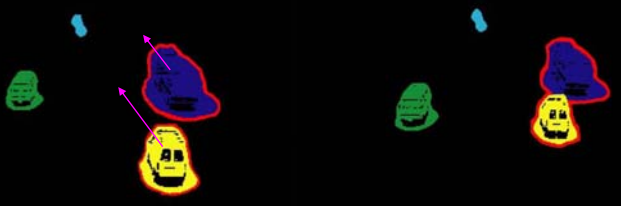
Framework



Why object detection is not enough?



Object tracking



Object tracking: examples

Outline

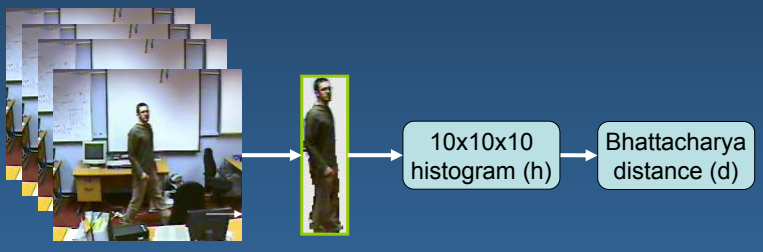
- Introduction
- Tracking algorithms
  - Mean-shift
  - Particle filtering
  - Graph-matching
- Integration of detection and tracking
- Integration of audio and video
- Event detection

Problem statement

- Objective
  - To predict the target state over time → Position, Shape
- Problems
  - Changes in pose and illumination
  - Partial and total occlusions
  - Clutter and targets with similar appearance
- Steps
  - Target representation → Normalised colour histogram
  - Likelihood of a candidate → Based on Bhattacharyya coefficient
  - Tracking algorithms
    - Mean shift (MS)
    - Particle filter (PF)

Likelihood

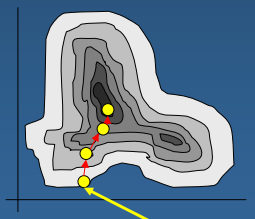
- Likelihood
  - Color → RGB space → 3D color histograms



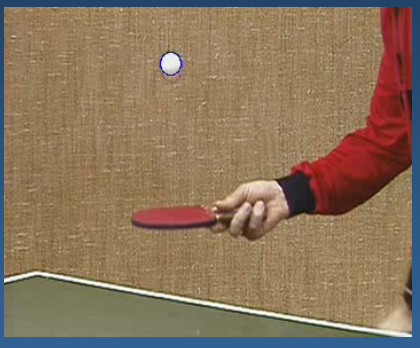
$$p(C \mid X_t) = e^{-\left(\frac{d(h,h_{ref})}{\sigma}\right)^2}$$

Mean shift: description

- Mean shift
  - Deterministic non-parametric approach
  - Iterative procedure
  - Kernel-based
  - Gradient-based approach
    - If the distance function is smooth (kernel) → effective



Mean shift: example



Particle filter: description

- State $\mathbf{x}_k = \mathbf{f}_k(\mathbf{x}_{k-1}, \mathbf{u}_k)$
- Observation $\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k, \mathbf{n}_k)$
- Objective
  - to estimate unknown state $\mathbf{x}_k$ based on a sequence of observations $\mathbf{z}_k, k = 0,1,\ldots$
  - find the posterior distribution

$$p(\mathbf{x}_k \mid \mathbf{z}_{1:k}) \approx \sum_{i=1}^{N} w_k^i \, \delta(\mathbf{x}_k - \mathbf{x}_k^i)$$

- Solution (Bayesian)
  - Prediction step
    - Based on state equation
  - Update step
    - Based on likelihood function

State transition model

- Typically
  - Zero-order model
    - *Limitation*: random positioning of the particles

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_k$$

  - First-order model
    - *Limitation*: high manoeuvring targets

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \boldsymbol{\theta}_{k-1} + \mathbf{u}_k$$

- Adaptive state transition model
  - Zero order model with adaptive noise variances

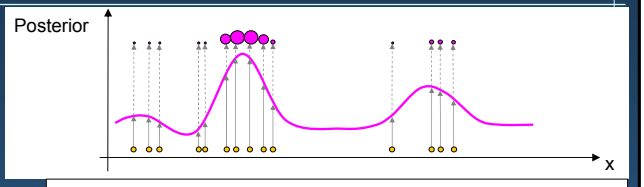
$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{C}_k \mathbf{u}_k$$

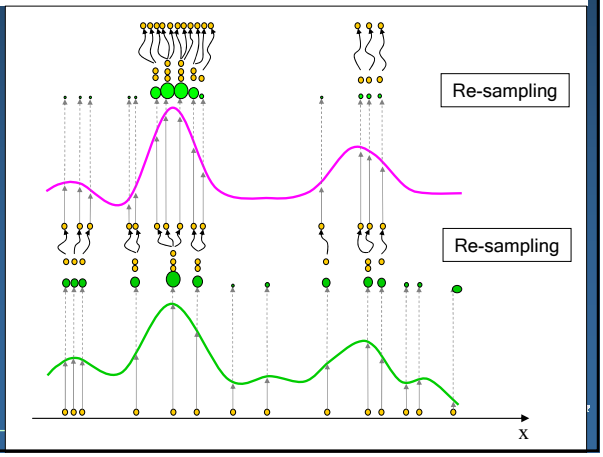
$$\mathbf{C}_k \propto \frac{1}{n} \sum_{t=k-n-1}^{k-1} \left| \mathbf{x}_t - \mathbf{x}_{t-1} \right|$$

average state velocity in the previous *n* frames

Re-sampling

- Problem
  - weight degeneration
- Solution
  - re-sampling (eliminates particles with small weights)





Particle filter: example

Hybrid tracker

re-sample particles

apply state transition

MS for each particle

re-weighting

E[.]

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{C}_k \mathbf{u}_k$$

Adaptive state transition model

Zero-order model with adaptive noise variances

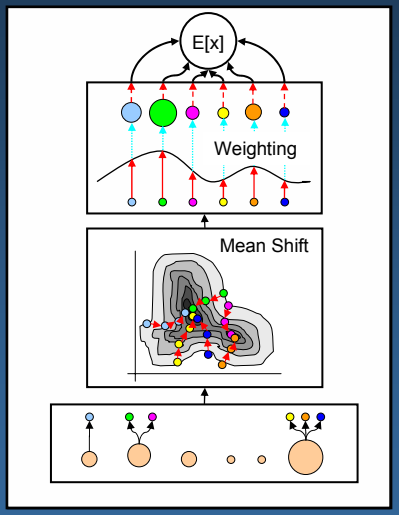
$$\mathbf{C}_k \propto \frac{1}{n} \sum_{t=k-n-1}^{k-1} |\mathbf{x}_t - \mathbf{x}_{t-1}|$$

Average state velocity in the previous *n* frames

The operator Mean Shift acts on the position 2D state space only

$$p(\mathbf{x}_k \mid \mathbf{z}_{1:k}) \approx \sum_{i=1}^{N} w_k^i \, \delta(\mathbf{x}_k - MS(\mathbf{x}_k^i))$$

Hybrid tracker



- Advantages

  - After MS → each particle is near a local maximum of the filtered posterior (position 2D sub-space)

  - The efficiency of the particles is increased

  - Multi-modality of the posterior is maintained

  - Extra computation is compensated by less particles

Results

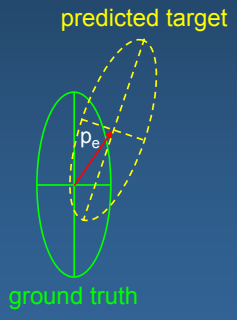
- Initialisation
  - Ground-truth initialisation of the target
- Parameters
  - Histograms: 10x10x10 (RGB)
  - MS: 5 times with different kernel sizes (+/- 10%)
  - PF, HT: 3D state model (to compare with MS): position; target size
  - Transition model $\sigma_x = \sigma_y = 14$; $\sigma_h = 0.013$; $k_s = 5$; $k_p = 10$
  - PF: 150 samples; HT: 30 samples
- Presentation of results
  - Videos
  - Sample frames & objective measure

Evaluation

- Subjective evaluation
  - Side-by-side visual comparison of tracking results
- Objective evaluation
  - Deviation from the ground-truth
    - APE: average position error (p_e)
    - ASE: average size error

$$ASE = \sqrt{W^2 + H^2}$$

W: width error
H: height error



Results: *highway*

MS

PF

Proposed

Evaluation: *highway*

	MS	PF	Proposed
APE	0.95	12.8*	0.88
ASE	2.74	22.3*	3.58

MS

PF

Proposed

Results: *soccer*

MS

PF

Proposed

Evaluation: *soccer*

	MS	PF	Proposed
APE	242*	3.9	3.2
ASE	18.2*	10.8	9.8

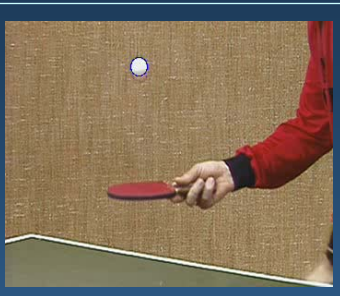
MS

PF

Proposed

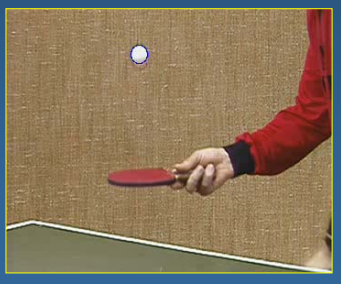
Results: *table tennis*

MS



PF

Proposed



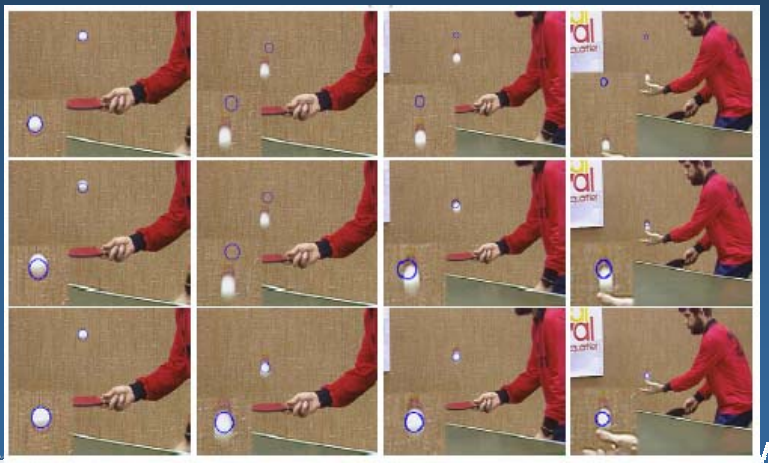
Evaluation: *table tennis*

	MS	PF	Proposed
APE	43.2*	24.1*	2.0
ASE	6.7*	3.3*	2.8

MS

PF

Proposed



Results: *emilio*

MS

PF

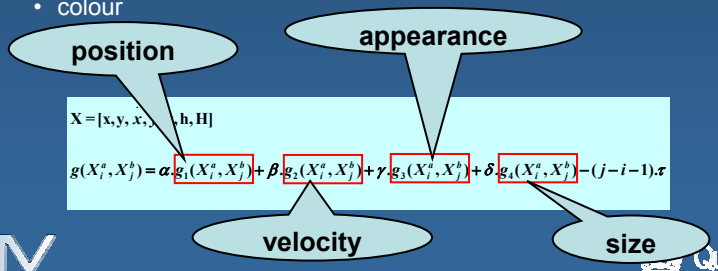
Proposed

Single vs. multiple target tracking

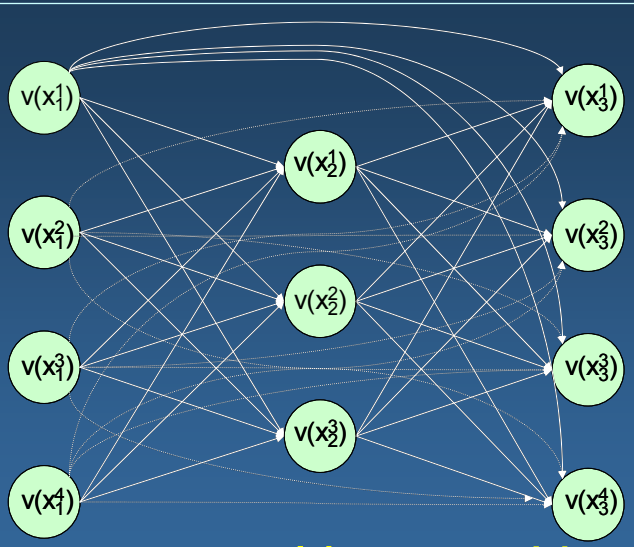
- Single target tracking
  - Hybrid mean shift / particle filter tracker
    - faster and more accurate than particle filter
    - more reliable than mean shift with fast targets
  - Adaptive transition model
    - Deal with highly manoeuvring targets
  - Cope with camera motion
- How about multiple targets?
  - Need to consider target ‘interactions’
  - NP problem
  - Complexity grows exponentially with n. of targets (PF)

Multiple object tracking

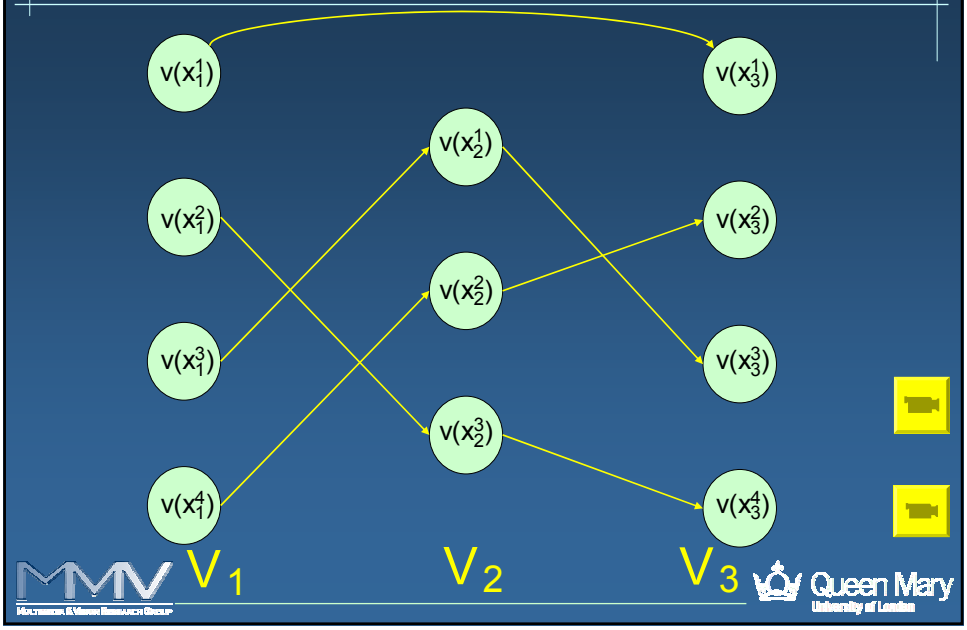
- Graph matching using weighted features
  - Data association verified throughout several frames to validate the correctness of the tracks
  - Support track recovery in occlusion scenarios
  - Features
    - centre of mass
    - velocity
    - bounding box
    - colour



Graph matching: full graph



Graph matching: max path cover



Detection vs. tracking

- Detection
  - Usually frame-based
    - Can be improved with temporal features (e.g., pedestrians)
  - Trained classifier
    - Choice of training set
    - Choice of negative examples
    - Choice of poses covered in the training set
- Tracking
  - Propagates the initialisation information
    - Model: template, statistical representation, parts, …
  - Should update the model
  - Should self-initialise

→ Integration!

Outline

- Introduction
- Tracking algorithms
  - Mean-shift
  - Particle filtering
  - Graph-matching
- Integration of detection and tracking
- Integration of audio and video
- Event detection

Integration of detection and tracking

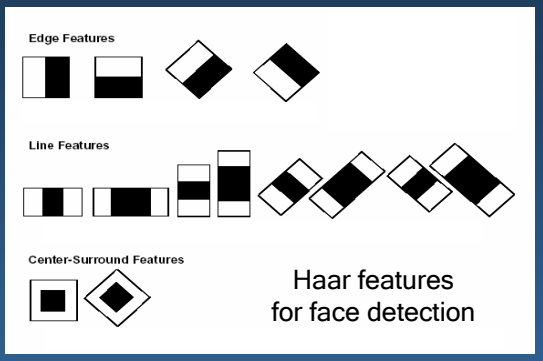
- Problem
  - Detecting objects (e.g., faces) in clutter
  - Tracking multiple object (e.g., faces) under occlusions

→ Integration of Adaboost face detector and Bayesian tracker

Face classifier

- Approach
  - Cascade of classifiers
  - Integral image
  - Training
  - Set of scales
- Output
  - Few false negatives
  - Many false positives …

→ Need additional evidence
→ Fusion of color analysis (chromaticity segmentation) and face classification



Haar features for face detection

Filtering through chromaticity segmentation

face detector only

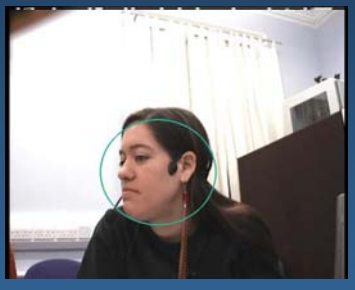
face detector with chromaticity segmentation

Detection and tracking

- Use particle filtering to track *between* detections
  - Initialization
    - detection away from current particles → candidate track
    - candidate track → activated after successive detections (confidence)
  - Filtering
    - if two tracks overlap → keep that with highest confidence score
      - number of tracked frames
      - frequency of detections
  - Termination
    - segmentation cue (*skin*)
    - detection cue (*classifier*)
    - size cue (*ratio* and *area*)

Detection and tracking

Removal of overlapping tracks



Particle (temporal) filtering for face tracking

- Particle filtering integrated with face detector
  - Link candidates from prediction (particles) with candidates from detection → connected detection (CD)
- Particle spread (*temporal prediction*)
  - If no CD → zero-order motion model
  - If CD → particles are partially spread in the detection area
- Object model (*color histogram*)
  - If no CD → no update
  - If CD → partially update (e.g., by 25%)

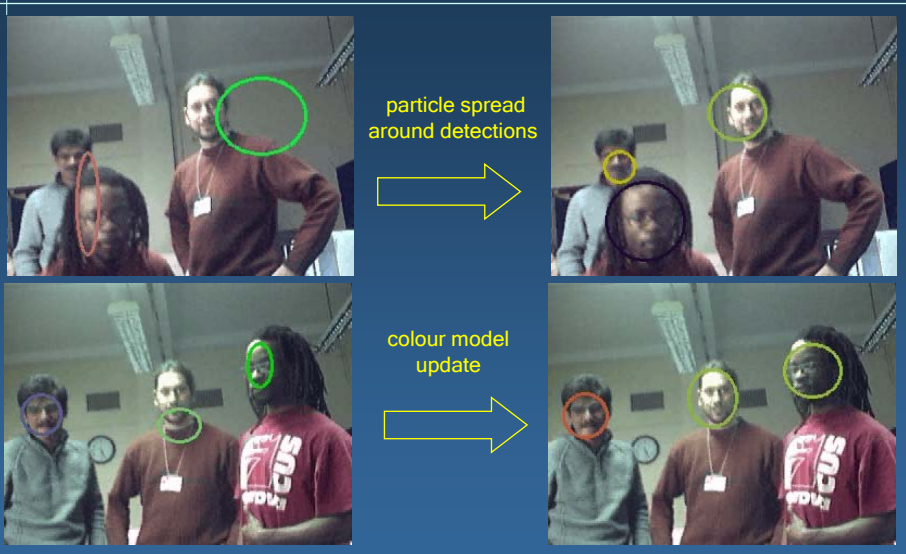
Integrated detection and tracking

without particles around detections

without model update

with integration

Detection and tracking



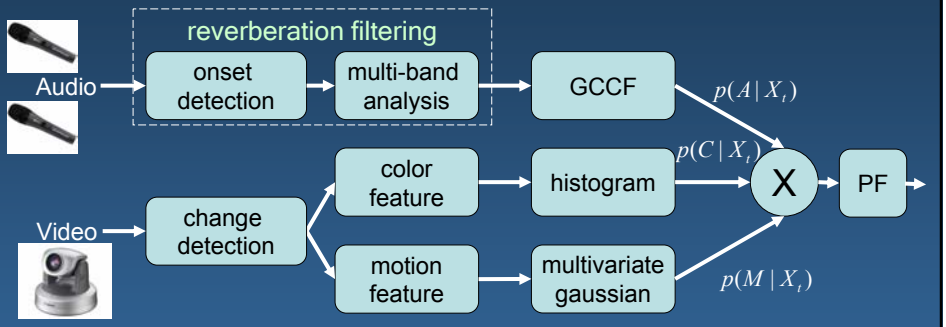
Face detection and tracking

Outline

- Introduction
- Tracking algorithms
  - Mean-shift
  - Particle filtering
  - Graph-matching
- Integration of detection and tracking
- Integration of audio and video
- Event detection

Multi-modal data fusion using particle filter (PF)

$$X_t = \{x, y, width, height\}$$



- Overall likelihood

$$p(O / X_t) = p(M / X_t)\, p(C / X_t)\, p(A / X_t)$$

Audio likelihood

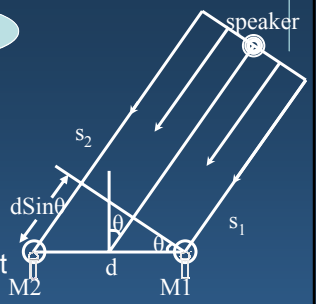
- Time delay of arrival (TDOA)

noise

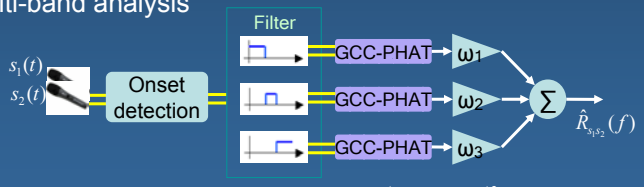
$$s_1(t) = v(t) + n_1(t)$$

$$s_2(t) = \lambda v(t+\tau) + n_2(t)$$

attenuation



- Reverberation filtering
  - Onset detection based on precedence effect
  - Multi-band analysis



$$p(A|X_t) = \frac{1}{\sigma_A \sqrt{2\pi}} e^{-\frac{\left(\hat{\varsigma}_A(\hat{R}_{s_1 s_2}) - x_t\right)^2}{2\sigma_A^2}}$$

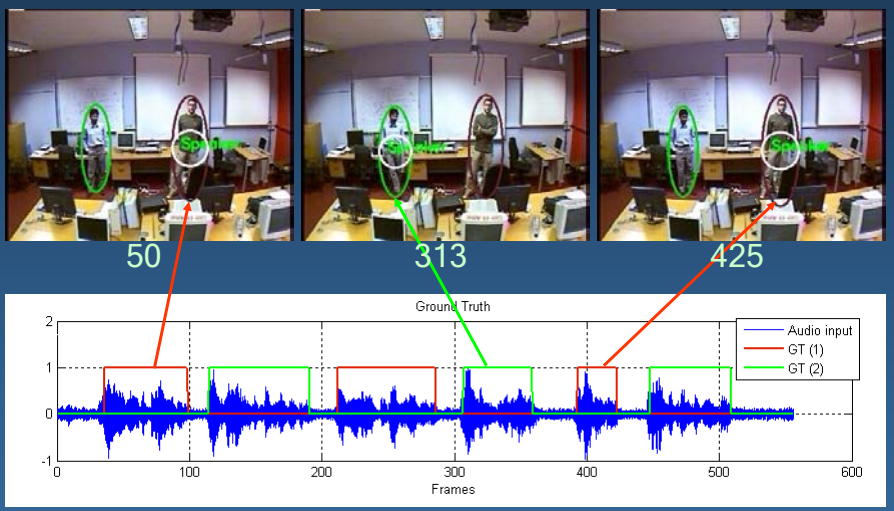
Comparison

Audio only

Video only

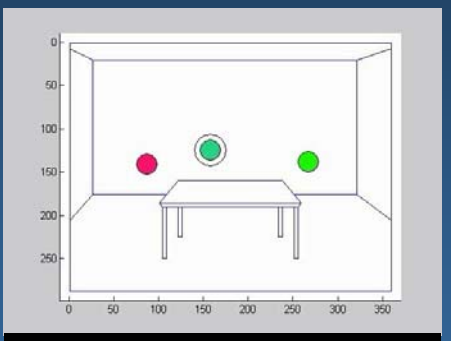
Audio Visual

Results – speaker detection



Results – scene dynamics for teleconferencing

Original video



Abstract representation

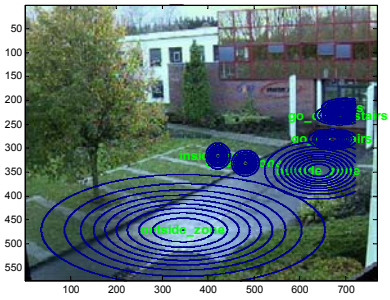
Outline

- Introduction
- Tracking algorithms
  - Mean-shift
  - Particle filtering
  - Graph-matching
- Integration of detection and tracking
- Integration of audio and video
- Event detection

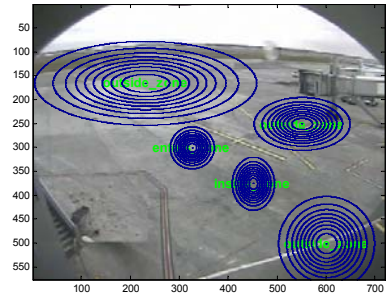
Event detection

Contextual information

- Scene modelling
  - Gaussian for each area of interest
  - outside zone → modelled with multiple Gaussians



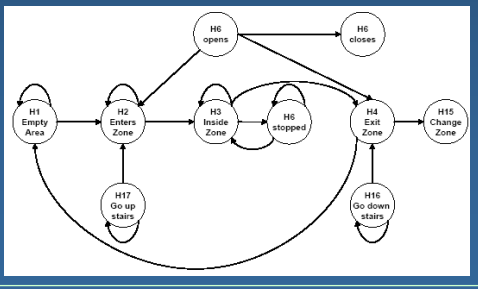
Building entrance - Camera 1



Airport - Camera 4

Object information

- Object detection and tracking
  - Observations O
- Model parameters
  - States $\omega = \{\omega_1, \omega_2, \omega_3, \ldots \omega_N\}$
- Probability
  - State transition probabilities $A = \{a_{ij}\}$
  - Emission probabilities $B = \{b_{jO}\}$
  - Initial state $\omega(0)$

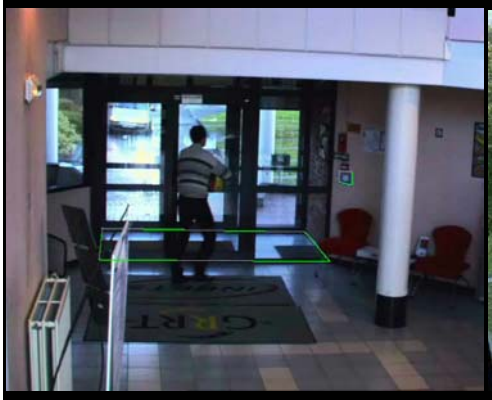


MMV Multimedia & Vision Research Group

Queen Mary University of London

Results

Results



Summary

- Tracking algorithms
  - Mean-shift
  - Particle filtering
  - Graph-matching
- Integration of detection and tracking
- Integration of audio and video
- Event detection

Acknowledgements

Emilio Maggio
Murtaza Taj
Matteo Bregonzio
Huiyu Zhou
Stefan Karlsson

EU FP7 project APIDIS (2008 – 2010)