IBM Research

Class 5: Tracking 2

Andrew Senior

aws@andrewsenior.com
http://www.andrewsenior.com/technical

Further tracking

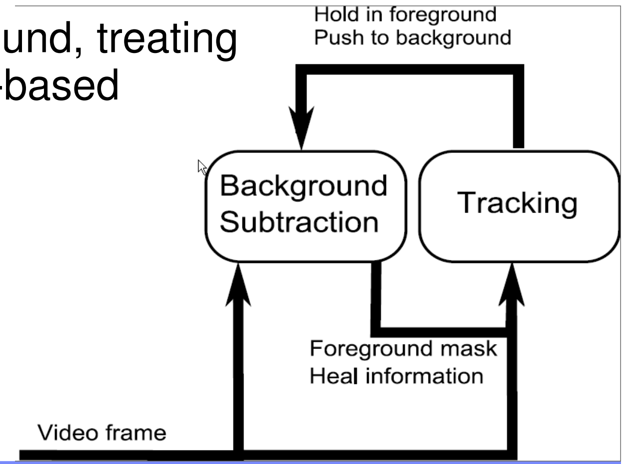
- Interaction of tracking and background subtraction
- Multi-cue FG/snake/head tracker
- 3D tracking
- Condensation
- Articulated body tracker
- Mean-Shift tracker (Histogram techniques)
- Tracking in crowds
- Tracker-based alerts

Tracking difficulties

- Many other tracking problems:
  - Fragmentation- BGS often fails. An object becomes two regions
    – new fragments are absorbed into nearby tracks untils split by fission
    – “Fusion” class accumulates evidence for nearby objects merging
  - Two objects may enter together and be indistinguishable until later
    – “Fission” class accumulates evidence for splitting object
  - One object leaves as another enters
    – Detect “Relay” tracks
  - One object occludes another for a long period
  - Objects stop and are “learned” by the background model
    – Tracker control over the BGS inhibits adaptation of tracked objects
    – Tracker forces push/pop to background model for truly static objects

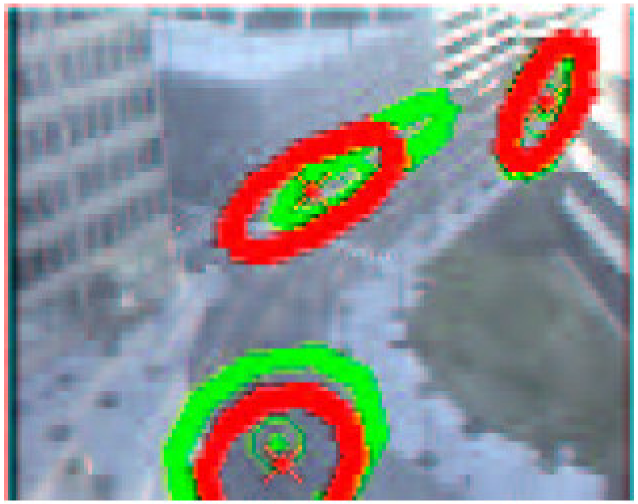
Interaction of tracking and background subtraction

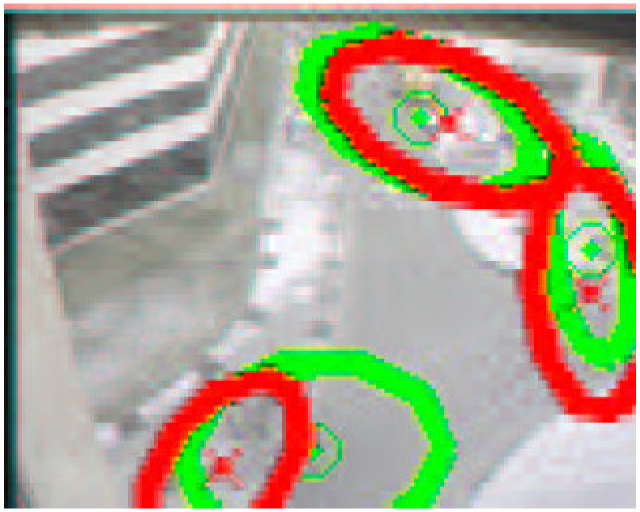
- Often constructed as a modular, feed-forward system
  - Simpler analysis
- Tracking can inform background subtraction
- Object detection
  - BGS is a one-class classification problem
  - With a known object, 2-class classification should be easier
    - Choose ML class of pixel among BG & predicted FG- to give more accurate boundary
- Tracker "understands" "objects"
  - Knows that an object is stopped or moving
  - Tracker can control when objects become part of background, treating them as unitary regions, whereas BGS must rely on pixel-based methods or region heuristics.
  - Inhibit adaptation for verified, temporarily-stopped objects
  - Push known stopped objects into BG



Track Sources and Sinks

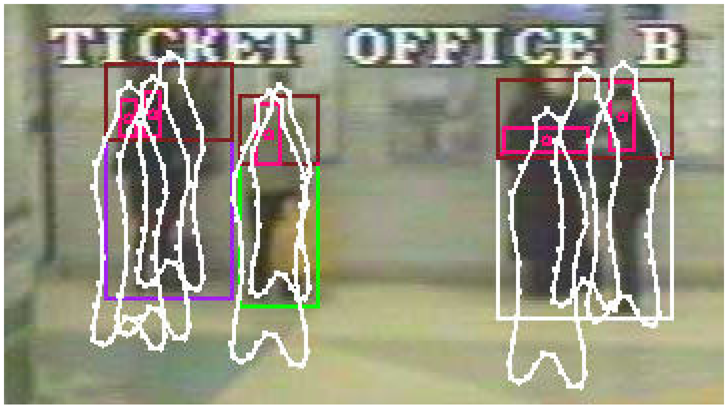
- Hand mark / learn where objects appear and disappear (see behaviour analysis class)
  - Stauffer “Estimating Sources and Sinks”
- Information can be used to distinguish between noise and true observations
  - A new object shouldn’t appear except at a source
  - Objects reaching a sink are likely to disappear

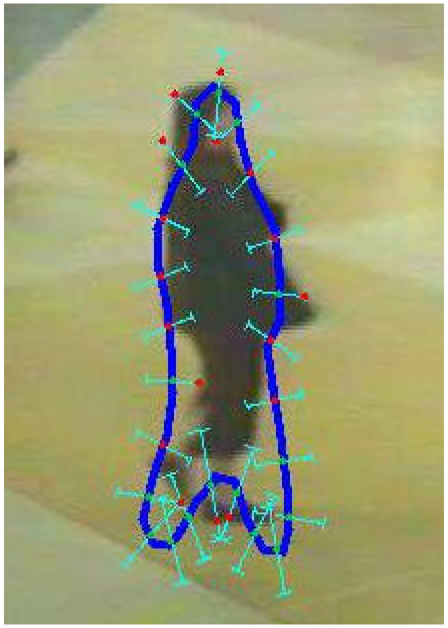




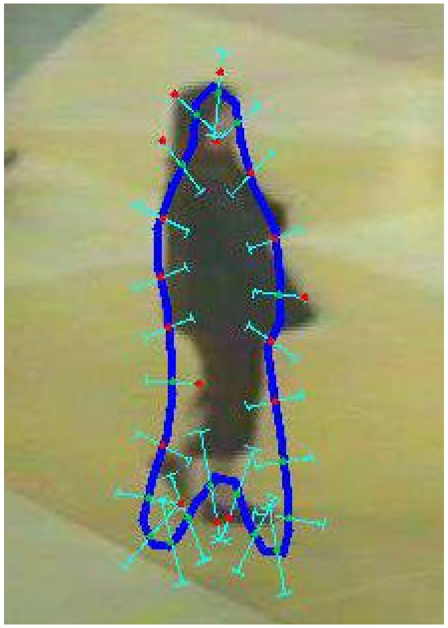
Siebel's "Reading" tracker

- Based on Baumberg 1995 (Leeds tracker)
- Extended by Siebel 2002
- Detection by BGS
- Tracking of regions
- Modelling people by snakes
  - Size based on calibration
  - Hypotheses based on head&shoulders (cf W4)

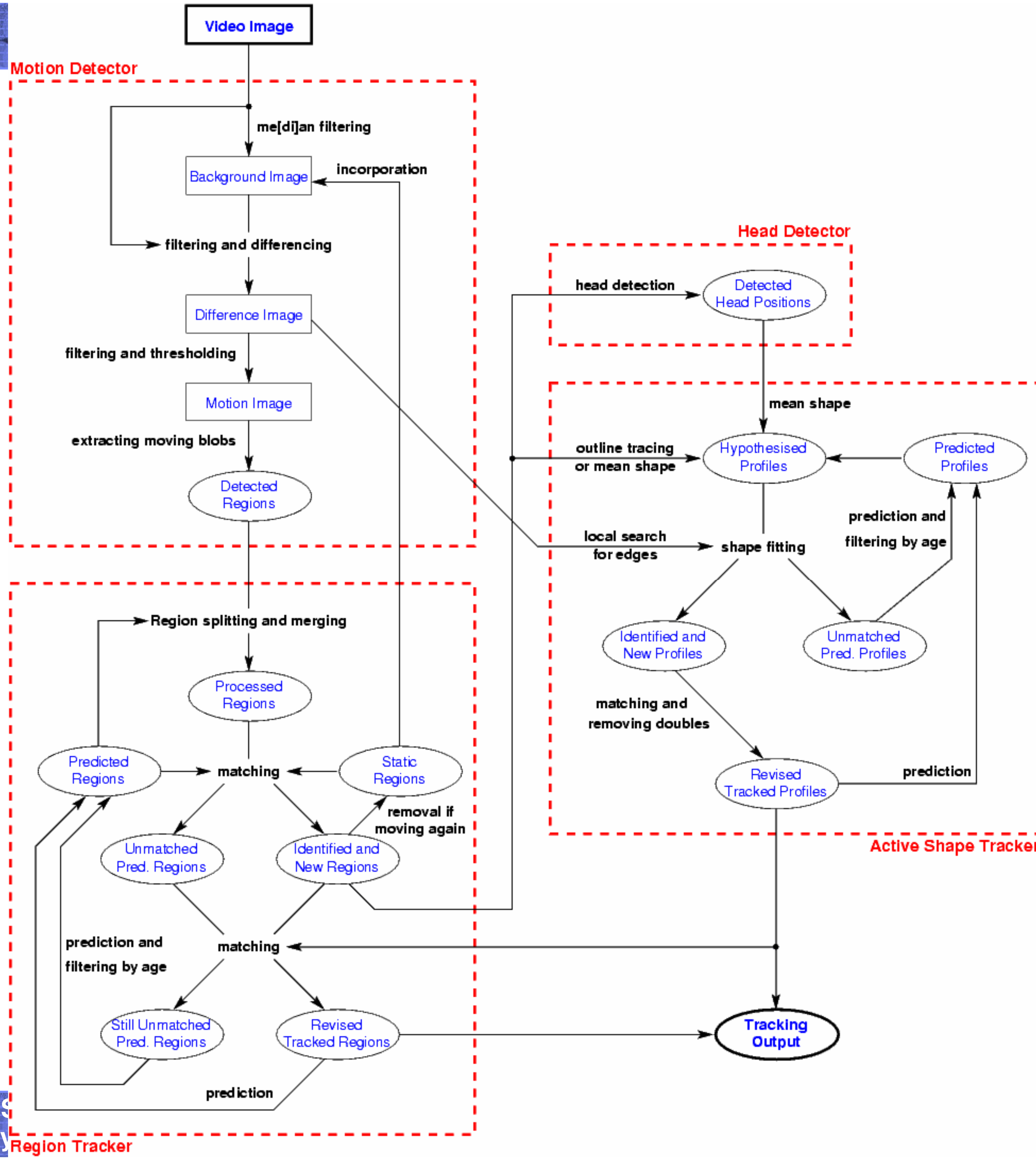




Snakes (Active Shapes)– a common model-based tracking approach



- Mark outline of training set of objects
  - Database of pedestrian silhouettes
- Fit curves e.g. B-Spline to contour
- Control points of splines concatenated into a vector x
- Find mean and Covariance matrix S of {x}
  - Hence find principal components $\{v_i\}$
- Track shape
  - At sample points on contour, find edge in perpendicular search direction
  - Find control point displacement to fit edge displacements
  - Project into principal components to ensure fit to model.
- Result- shape that matches observed contour, while still similar to training set exemplars


Video Image
Motion Detector
me[di]an filtering
Background Image
incorporation
filtering and differencing
Difference Image
filtering and thresholding
Motion Image
extracting moving blobs
Detected Regions
Head Detector
head detection
Detected Head Positions
mean shape
outline tracing or mean shape
Hypothesised Profiles
Predicted Profiles
prediction and filtering by age
local search for edges
shape fitting
Identified and New Profiles
Unmatched Pred. Profiles
matching and removing doubles
Revised Tracked Profiles
prediction
Active Shape Tracker
Region splitting and merging
Processed Regions
Predicted Regions
matching
Static Regions
removal if moving again
Unmatched Pred. Regions
Identified and New Regions
prediction and filtering by age
matching
Still Unmatched Pred. Regions
Revised Tracked Regions
prediction
Tracking Output
Region Tracker

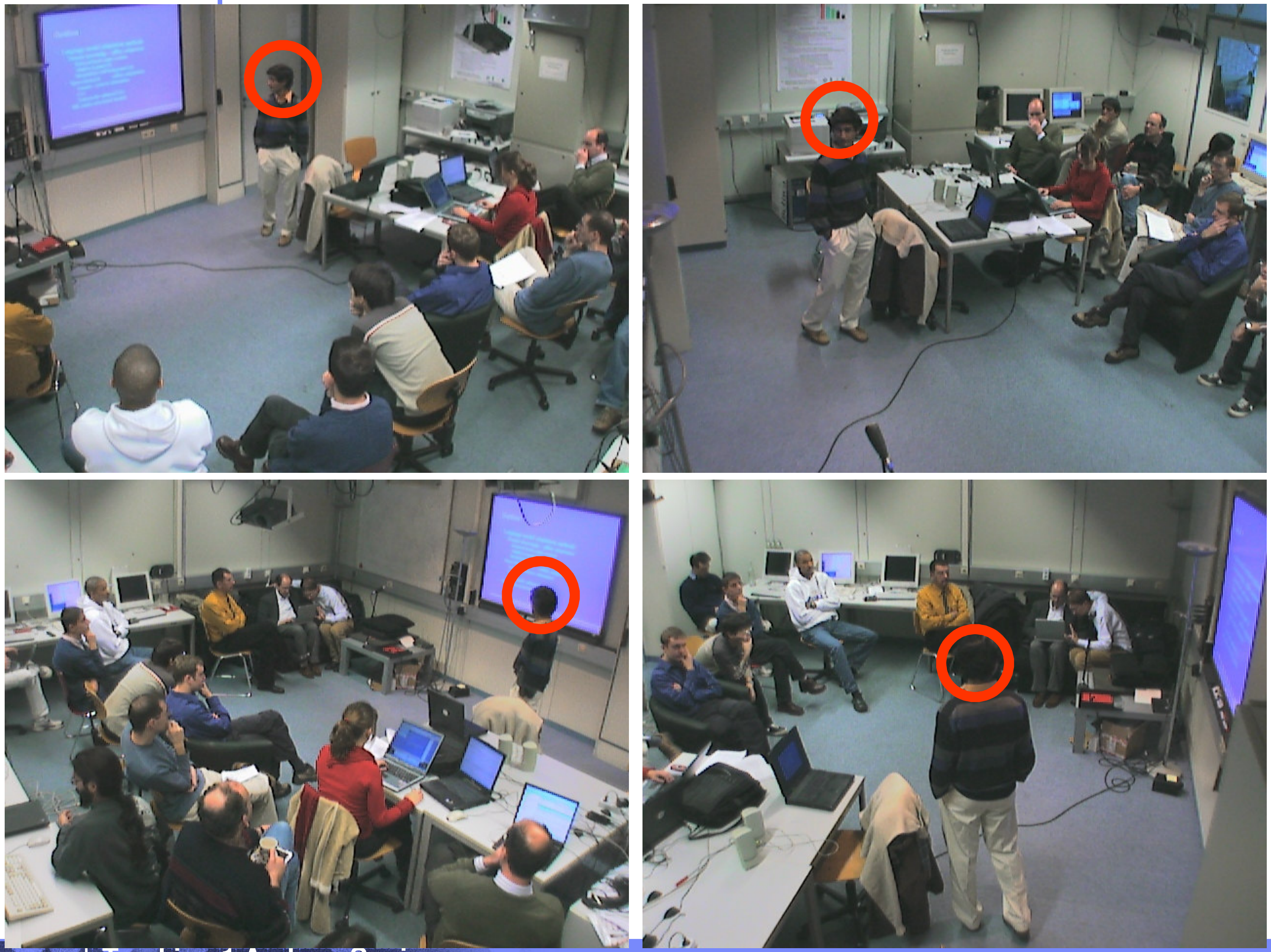
IBM Research

Tracking for seminar understanding
The "CHIL" project

Tracking for seminar understanding

- CHIL (Computers in the Human Interaction Loop) project:
  - EU 6th Framework consortium for mining seminar data
  - Similar to AMI (focusing on “meeting mining”. Now AMIDA)
- Understand speech and participant actions
  - For indexing, summarization, live status
- Speaker location important for
  - Role & activity understanding
  - Steering of resources
    - Microphone arrays: for improved speaker ID, speech reco
    - PTZ cameras: For face reco, gesture, AV speech
- Goal: Joint Audio-Visual tracking

CHIL data- speaker head location

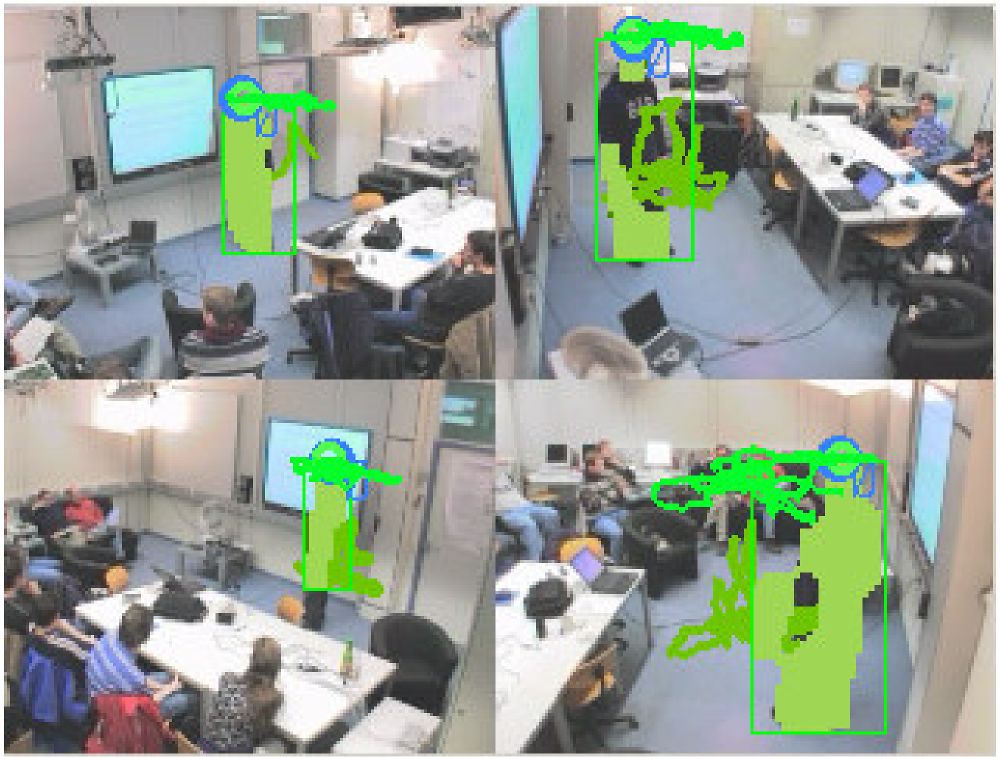


2D Tracking

- Track independently in each 2D camera view using IBM Smart Surveillance engine
  - Tracking through occlusions using probabilistic appearance models.
  - Relies on adaptive background subtraction
  - Background initialized from automatic backgrounds from ground truth sequences
  - Use “region of uninterest” to mask out non-speaker foreground areas in each camera (roughly estimated)
  - At 320x240, 4 cameras

3D Localization

- Triangulate "top of head" positions
  - y=upper row of object model bounding box
  - x=centroid of uppermost pixels
  - Each detection in a 2D image specifies a 3D ray.
  - Hypothesize closest approach of ray pairs as head locations



3D Tracking

- Use Viterbi search through 3D triangulation points
  - Beam search (50 candidates)
  - Find least distance path through 3D points
  - Extra penalties (start, end and skipped frames)
- Assumes exactly one “speaker”
  - No speaker location prior
  - Does not exploit 2D tracking
  - Points are sparse- linear interpolation for comparison to ground truth
- Speed
  - C++, 4 cameras ~27fps on 3.0GHz machine
  - ~Linear in #pixels

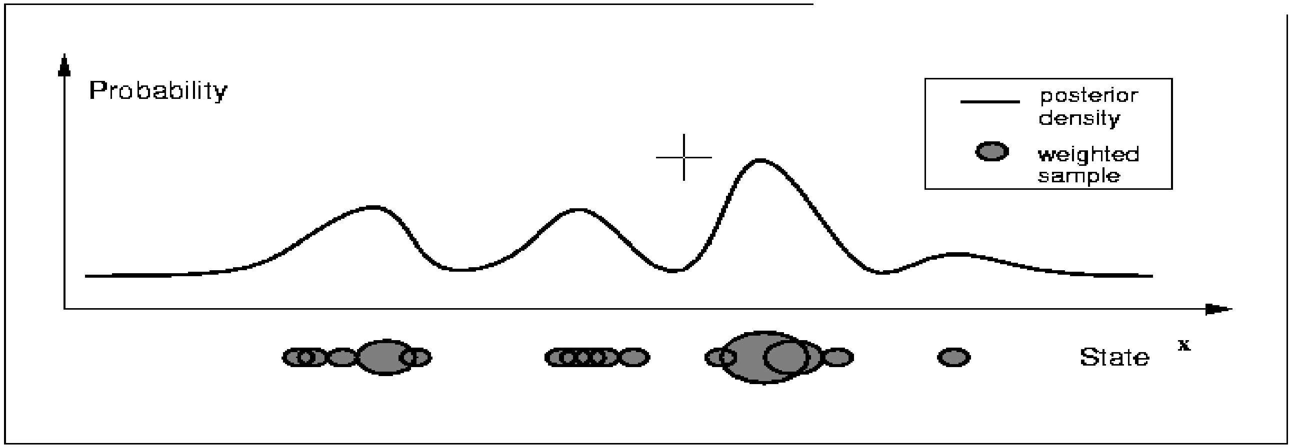
Condensation based tracking: Particle filtering

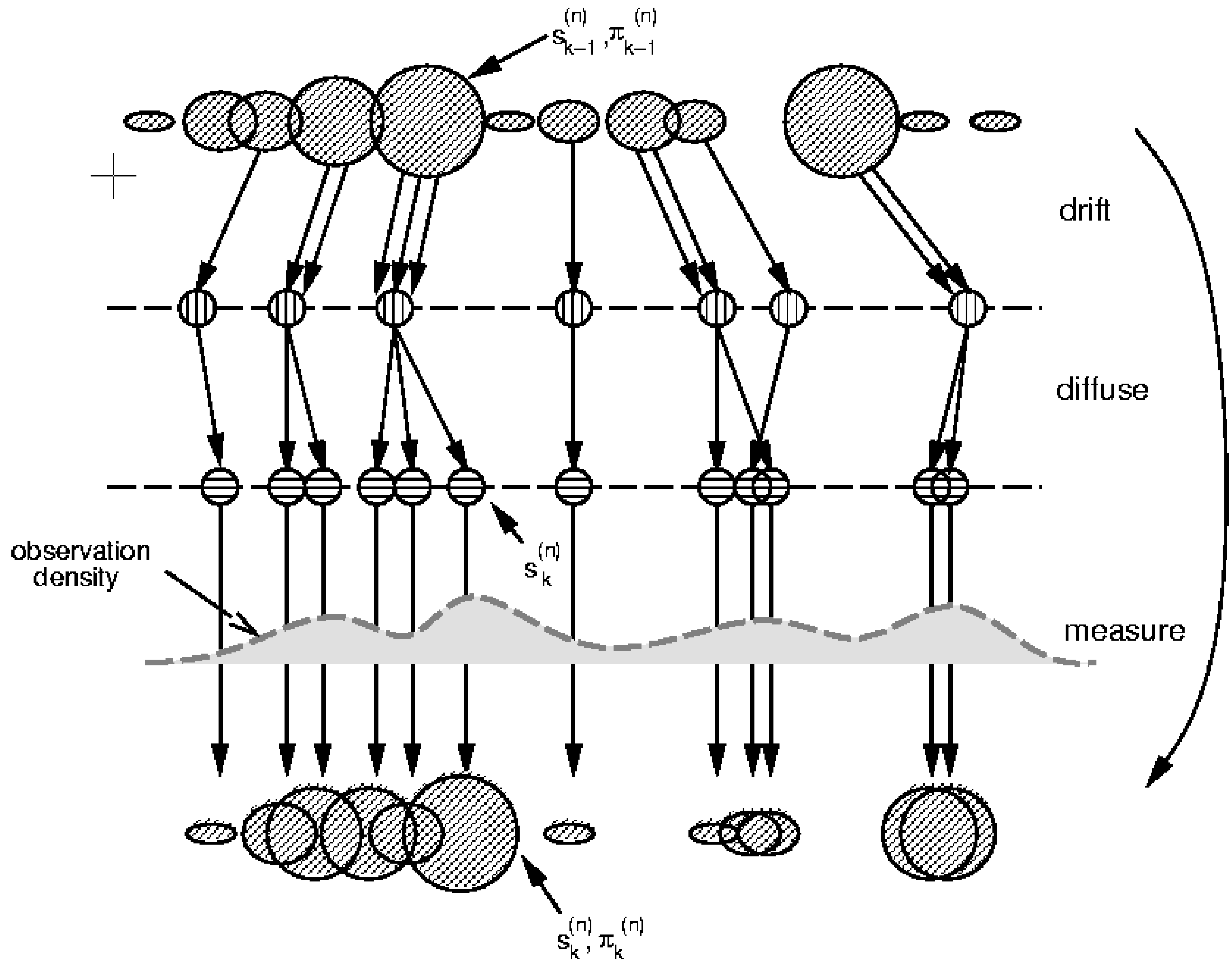
- Particle filter models multiple hypotheses as “particles”
  - Particles represent parameters of a hypothesis and are weighted with prior of the hypothesis
  - At each iteration particles are propagated / perturbed
  - Tracking, possibly random variation
  - Evaluate particles to determine their relative likelihood
  - Resample the particles by weight to give new distribution
- Need hundreds of particles for even a few dimensions ~5
- Curse of dimensionality: more dimensions means many more particles
- Scoring/fitting have to be fast or very effective for so many hypotheses

Condensation

- Bayes rule, on state x and observations z:
  - $p(x|z) = k\, p(z|x)p(x)$
- Particles are sampled according to the prior p(x)
- Reweighted according to the evidence p(z|x)
  - Results in a distribution p(x|z) (after normalization)
- Iterate for subsequent frames using $p(x_{t-1}|z_{t-1} \ldots z_1)$ instead of p(x)
- This is based on prediction:

$$p(\mathbf{x}_t|\mathcal{Z}_{t-1}) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathcal{Z}_{t-1})$$




$s_{k-1}^{(n)}, \pi_{k-1}^{(n)}$
drift
diffuse
observation density
$s_k^{(n)}$
measure
$s_k^{(n)}, \pi_k^{(n)}$

Condensation

- Representative value?
  - Mode? Weighted mean?
- Tracking over time?
- Surveillance example

Particle Filter CHIL Tracker

- Inspired by Nickel *et al.*
- Particles are speaker location hypotheses in 2D space
- Particles reweighted according to image evidence: based on **image differencing**
    - Fast evaluation
    - Avoid background subtraction.
- Find mode & resample particles

Particle Filter Tracking

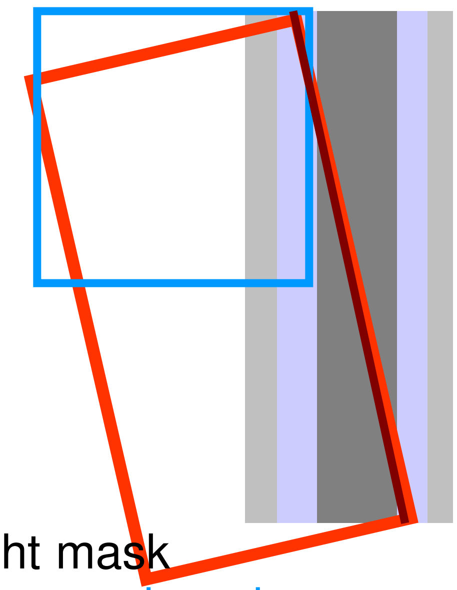
Hypothesis locations in green (height is weight)
Red rectangle is (cylindrical) object position projection

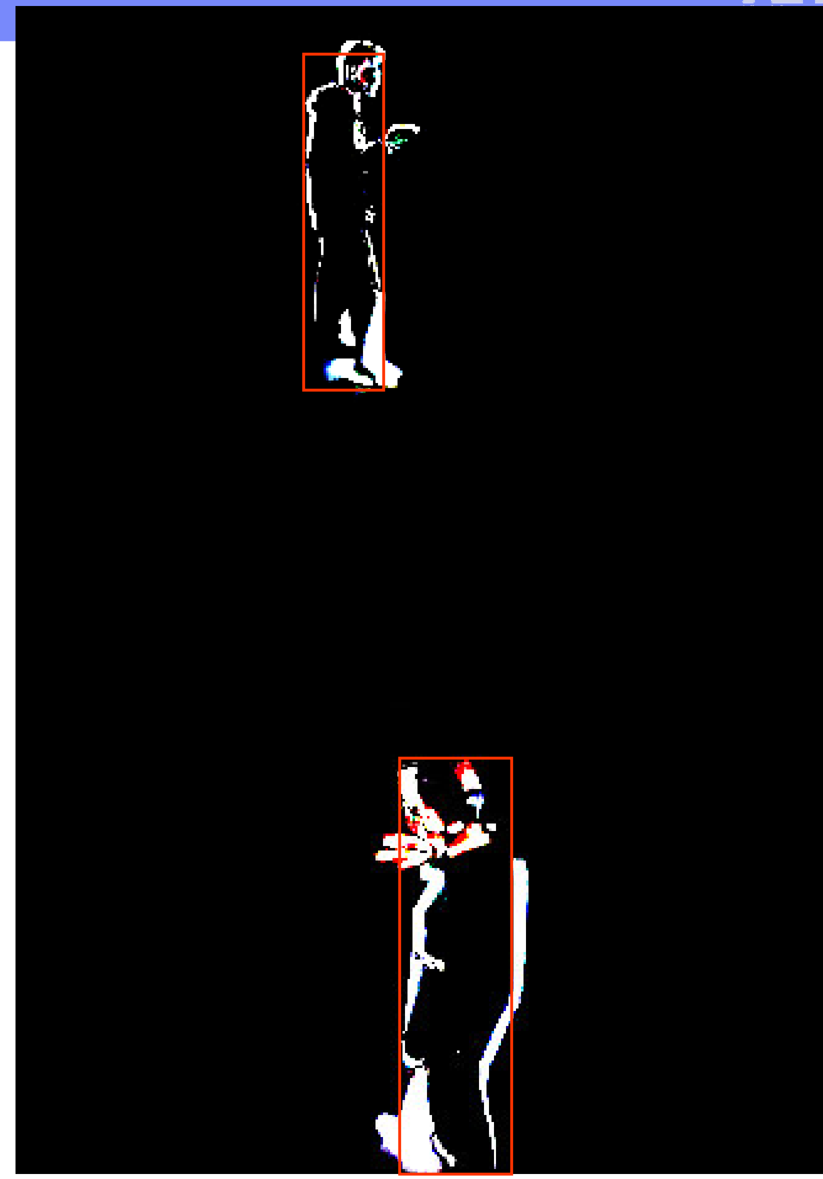
Particle scoring

- At each hypothesis project vertical edges of cylindrical object into each view.
- Evaluate particle according to sum of weighted frame difference around object edges.

- Optionally, apply face detector (slow)

$$\omega(p) = \sum_{v} \sum_{x} \delta I(x) w_p(x)$$

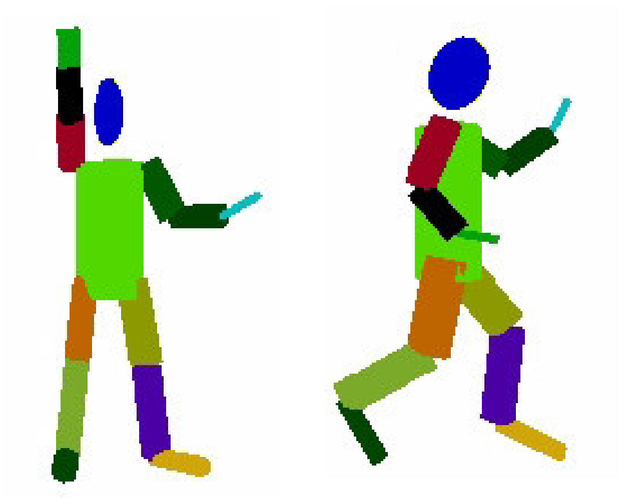


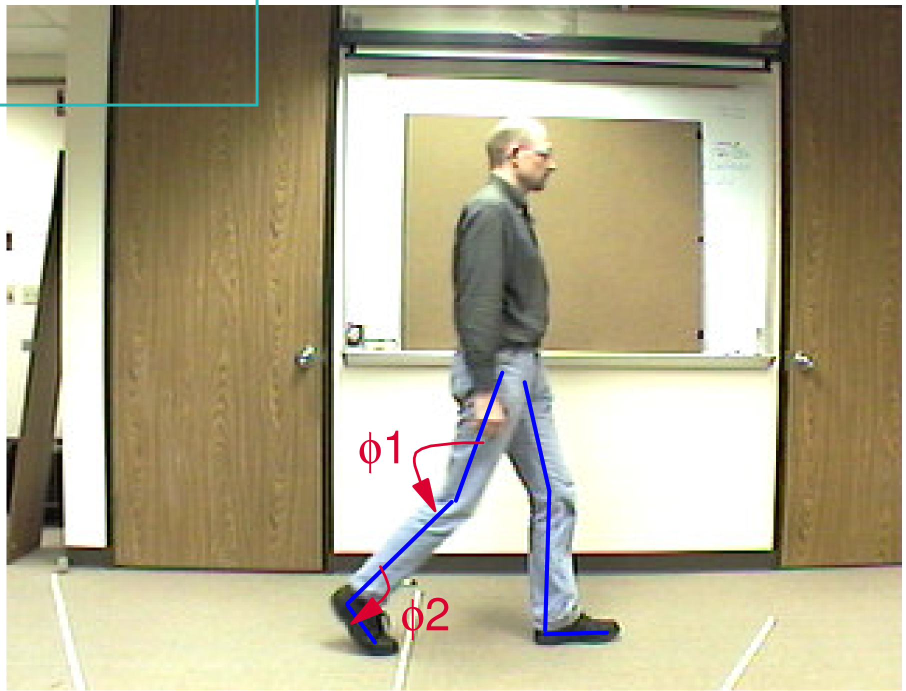


Object edges and weight mask for right edge, with face search region

Body Pose: Articulated Human Body Tracking

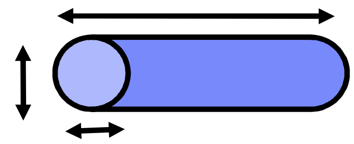
- Track articulations of human body, in real time
  - Track legs for gait analysis
  - Track arms/head for human-computer interaction
    - Gesture recognition
    - Gaze direction
- Iterative fitting of a 3D model

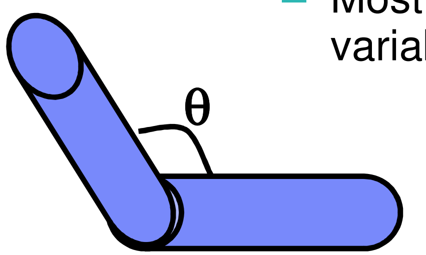




Model

- Up to 14 element model of generalized cylinders and ellipsoids
  - Coded in OpenGL- renderable as an image with limb labels.
  - Joints parametrized as twists in a kinematic chain
- Parameters
  - Static: joint lengths, diameters, limits
  - Dynamic: joint angles
    - Most dynamic parameters are held fixed to limit number of free variables.

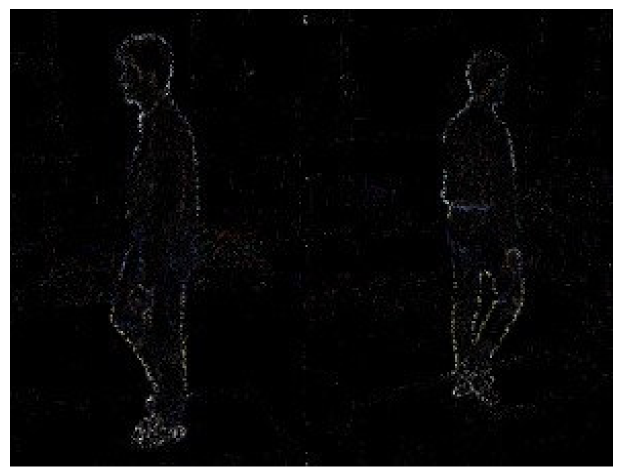




Features

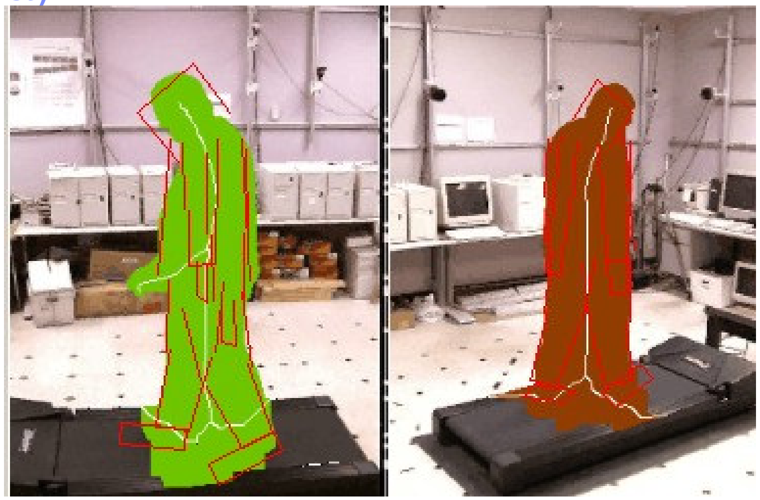
- Silhouette features
  - Extract silhouette of moving objects using background subtraction
    - Provided with CMUMobo data
    - Otherwise calculated with an adaptive version of the Horprasert algorithm
    - Multi-object edges lost without pixel-level segmentation

- Edge features
  - Calculated edges (Sobel operator) and difference with a background edge map
    - Sign of edges unknown
    - Internal edges

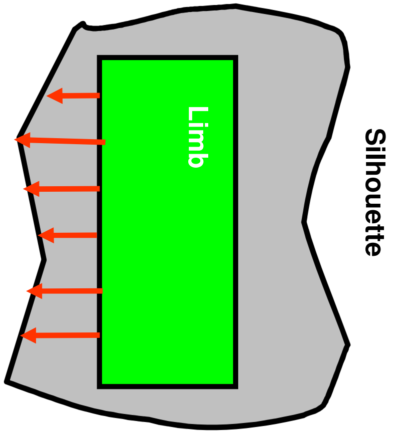


Fitting
(After Bregler & Malik, Drummond & Cipolla)

- Generate model occluding contour in each view
- Project model into each view using current parameters $\underline{\theta}$



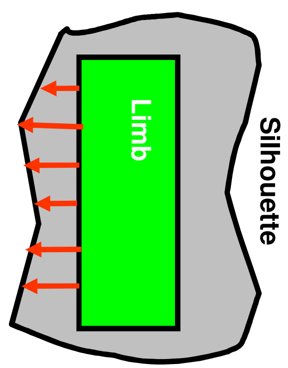
- For each contour element, search perpendicular for matching silhouette edge
- Gives many local displacements dx,dy
  - Even currently occluded edges might become disoccluded

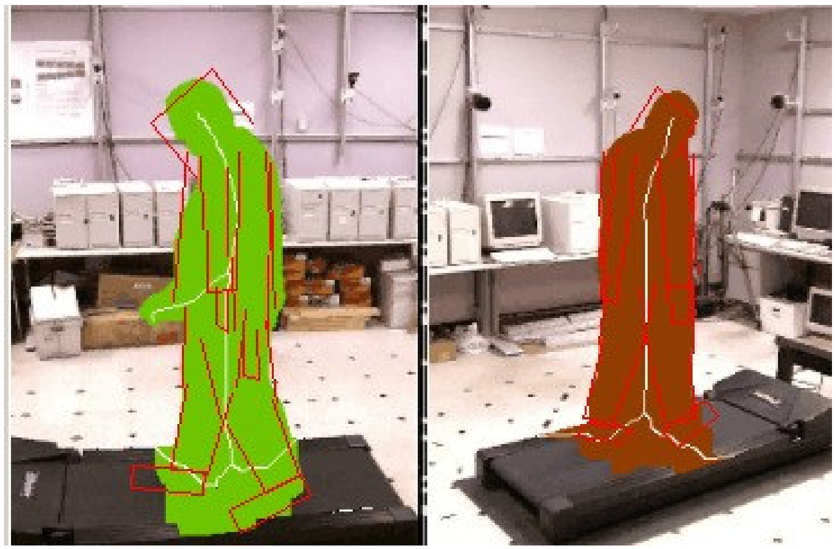


- Bregler & Malik used area textures (LK tracker)
- Drummond & Cipolla used image edges
- Framework supports all three simultaneously

Fitting

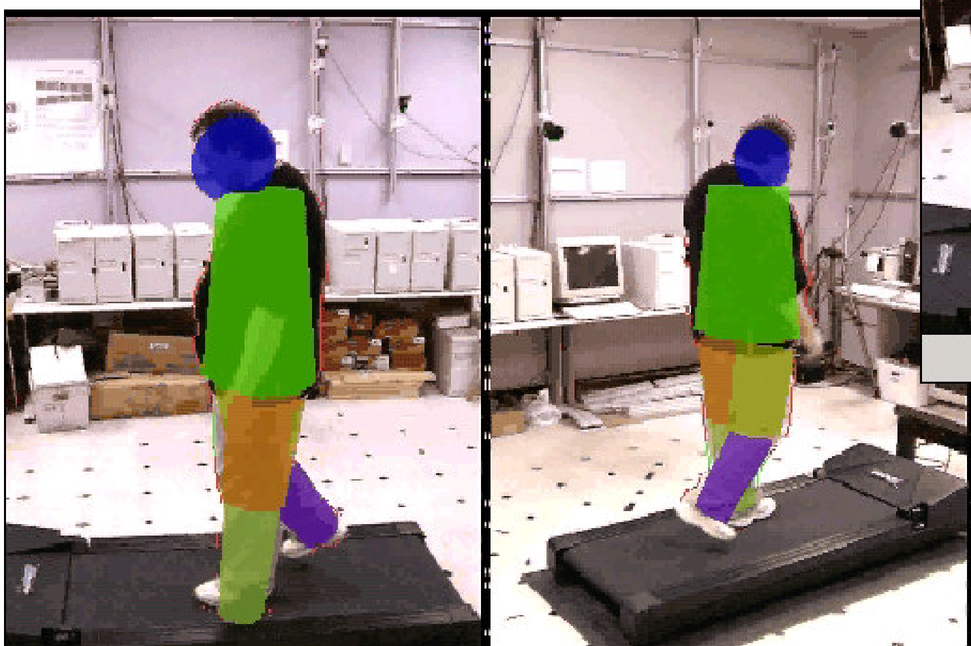
- Generate an equation in parameter changes $d\underline{\theta}$ to produce desired displacement dx, dy
  - Twist formulation for kinematic chain gives dx, dy in terms of $d\theta$: $H_x.d\underline{\theta}+d\underline{x}=0$
- Simultaneously solve all equations with nonlinear least squares.
  - After one iteration recompute edge correspondances
  - Iterate coarse-to-fine
- Apply penalty terms when joint angles go out of bounds.





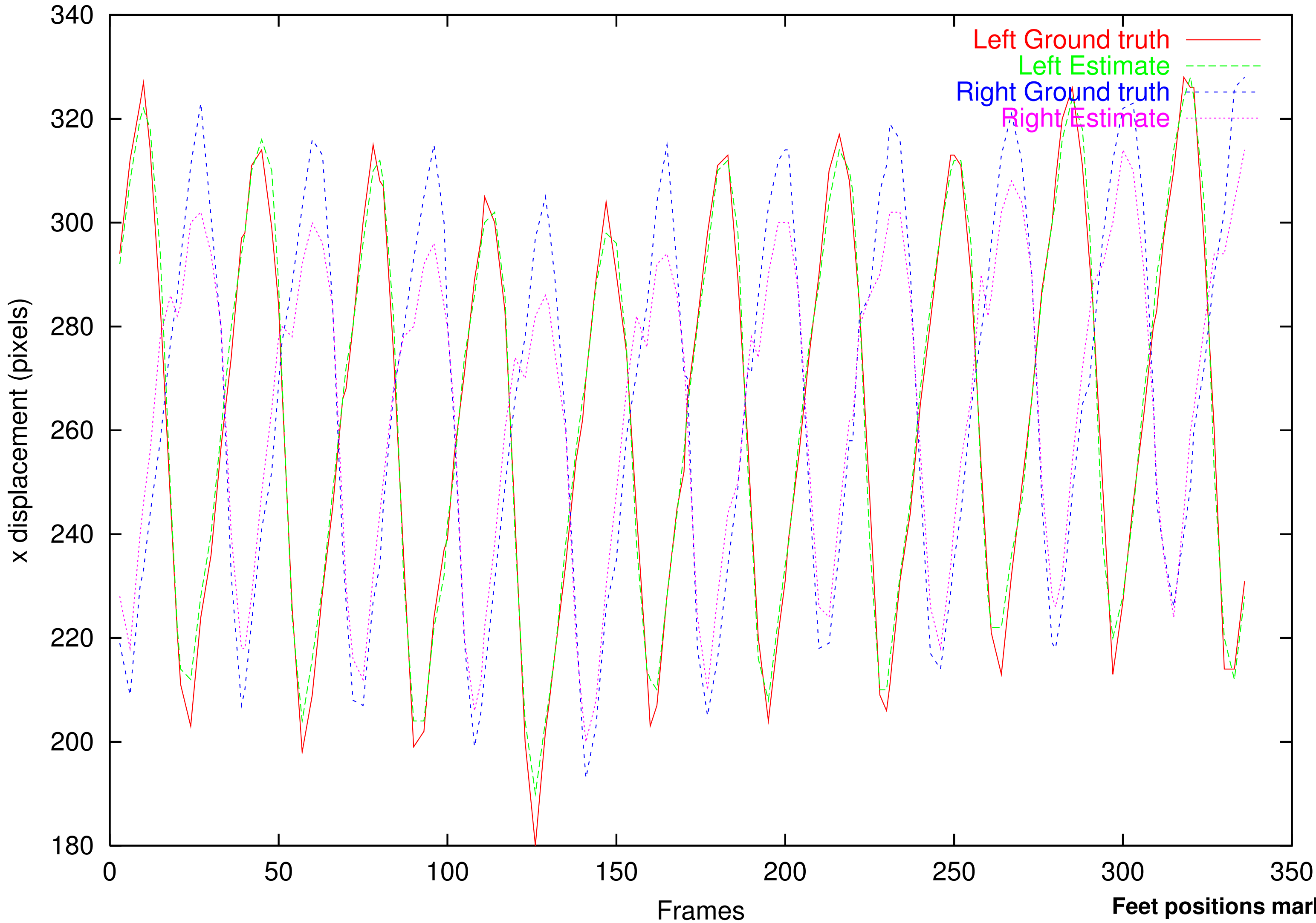
Articulated body tracking

04006_slowWalk_0313_05Mar2003_144245.avi



04002_slowWalk_0313_05Mar2003_143854.avi

Tracking Left & Right feet



Feet positions marked every 3 frames

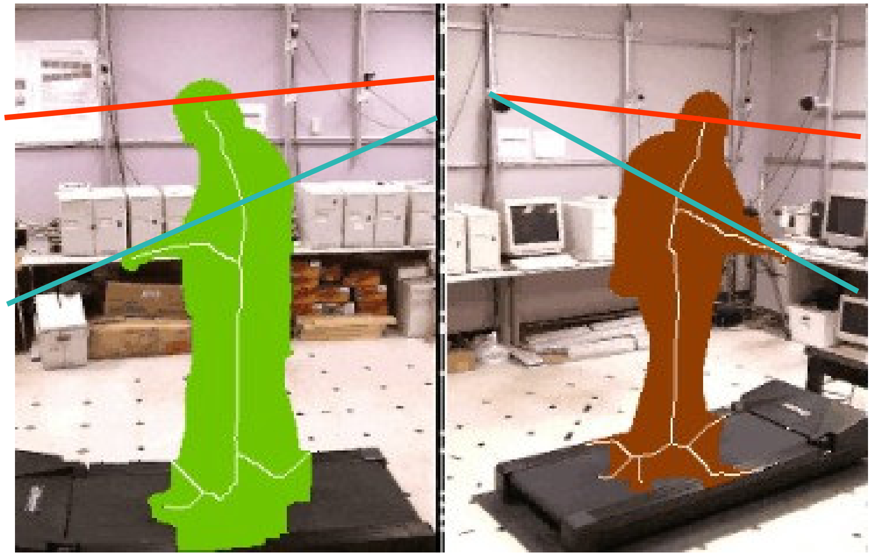
Speed

Views	Iterations	Time (ms)
2	3	31
3	3	38
4	2	33
4	3	44
4	5	62

- Video is 30 fps (33ms/frame)
- Dual 2.8GHz Pentium
- Ambiguity deweighting contributes 10%

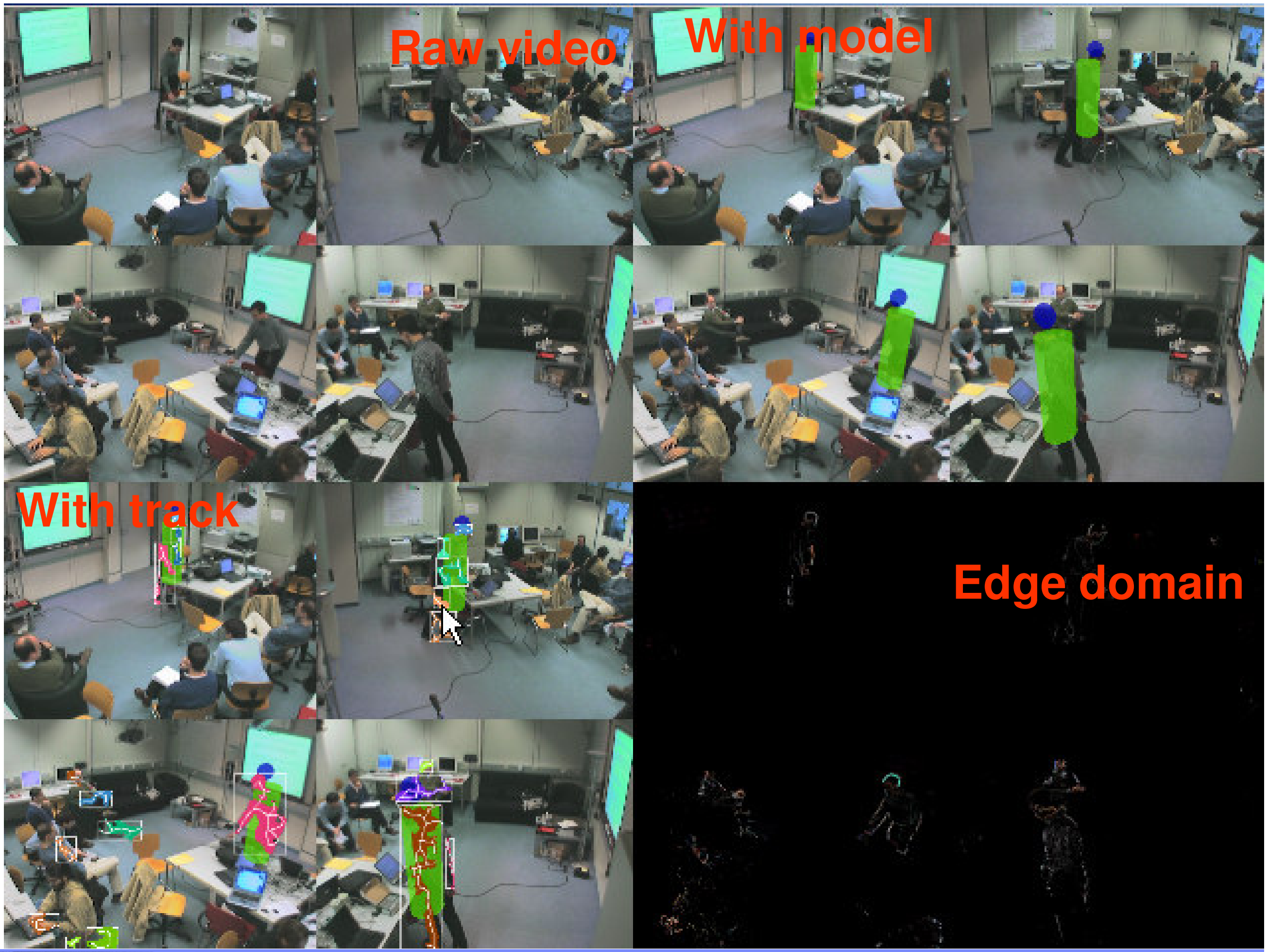
Initialization

- Triangulate skeleton end points using epipolar constraint
- Retain consistent hypotheses
- Simple heuristics to label candidate hands, feet, head in "simple" poses
- Displacements of identified points fit into optimization framework, solving inverse kinematics



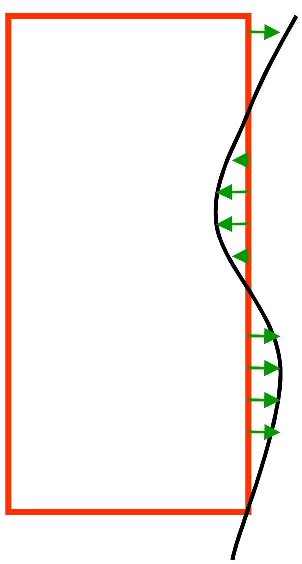
Applied to CHIL scenario: Edge alignment tracker

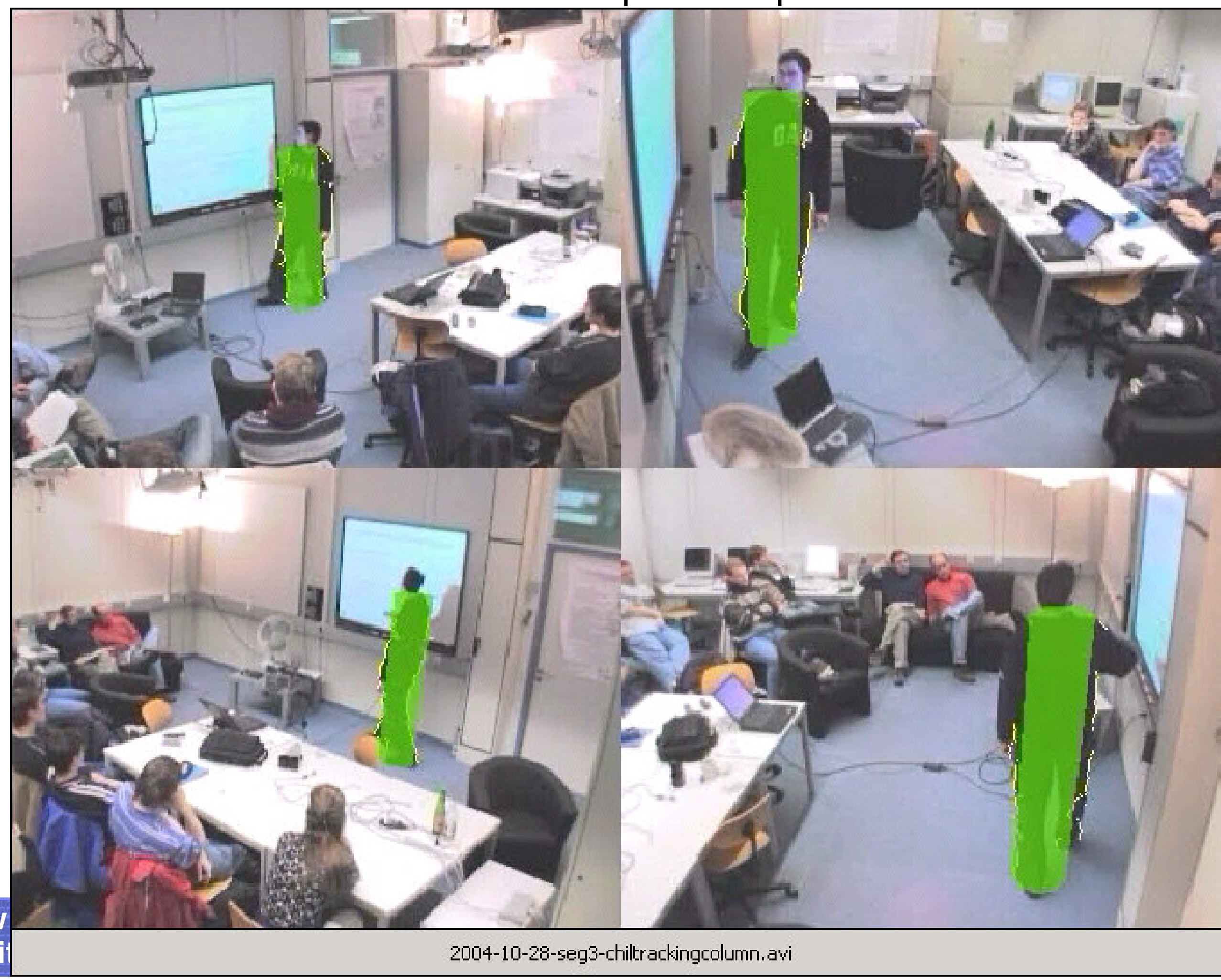
- Articulated body tracker applied with a rigid model
- Edge domain background subtraction
- Align 3D projected model edges with image edges



Fitting edge model

- Cylinder-only model with found edges
- Search perpendicular to model to find edges
- Project image displacements into model coordinates & optimize “pose” (here only 2 dof)





Bayesian Multiple Target Tracker
Narayana & Haverkamp CVPR 2007

- Bayesian model ot associate blobs with prior blobs
- (Not using track model)

	Blob 1	Blob 2	"lost"
Track 3	0.00	0.10	0.90
Track 7	0.00	1.00	0.00
Track 11	0.00	0.39	0.61
Track 12	1.00	0.00	0.00

Bayes belief matrix - frame 0240

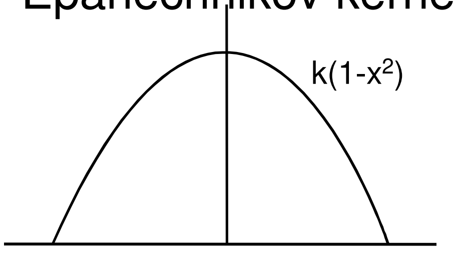
(a)

(b)

Fig. 2. (a) Belief matrix (b) Tracks resulting from Belief matrix for frame 240 of video sequence

Mean Shift Tracking

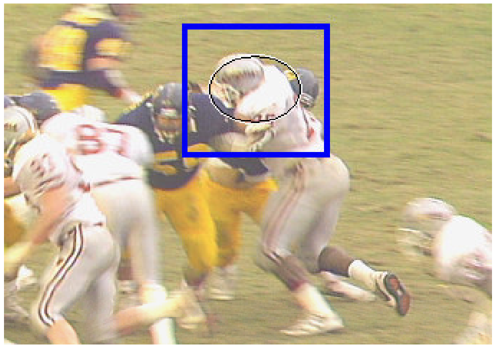
Epanechnikov kernel



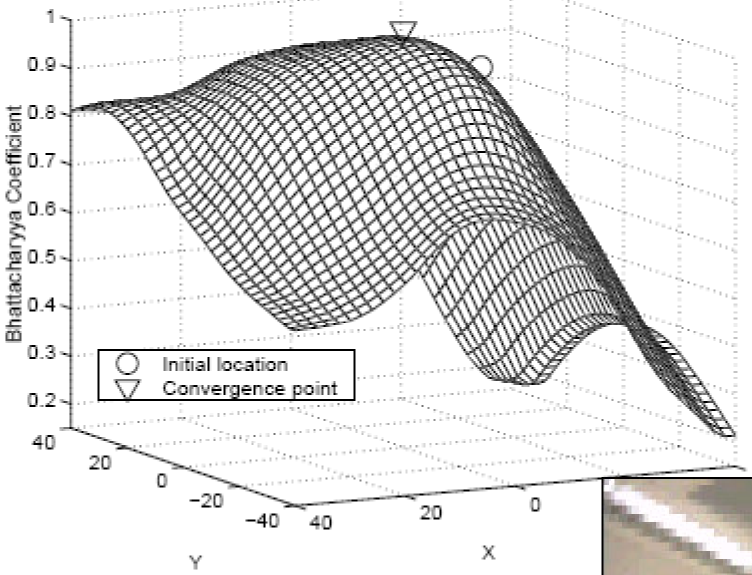
- Comaniciu & Meer
  - Use histogram gradients to track objects
- Given initialization ellipse
  - Compute kernel-weighted histogram q_u
  - Compute displacement of model to maximize Bhattacharrya coefficient
  - $\rho=\Sigma_i (p_i q_i)^{0.5}$
  - $y= \dfrac{\Sigma_j x_j w_j g(||y-x||^2)}{\Sigma_j w_j g(||y-x||^2)}$
  - Simple scale search- try +/- 10% and see which fits best
- 32x32x32 bin histograms 30fps on 600MHz PC
- Contains no spatial information

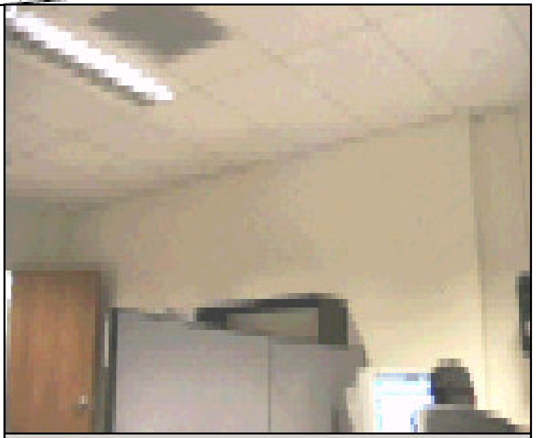
Mean Shift

- Bhattacharyya coefficient over a region of convergence



ComaniciuFootball.avi

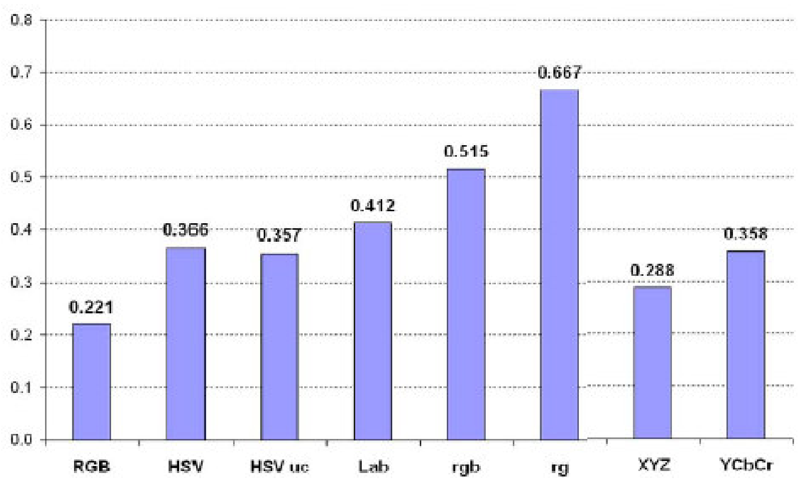


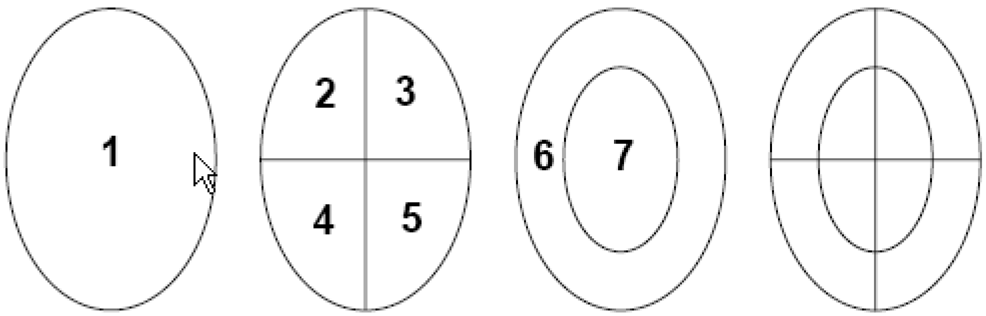


ComaniciuFace.avi

Mean shift

- Widely used, various enhancements
  - e.g. Scale (Collins)
- Multipart e.g. Maggio & Cavallaro
  - Compare colour spaces (RGB works best)

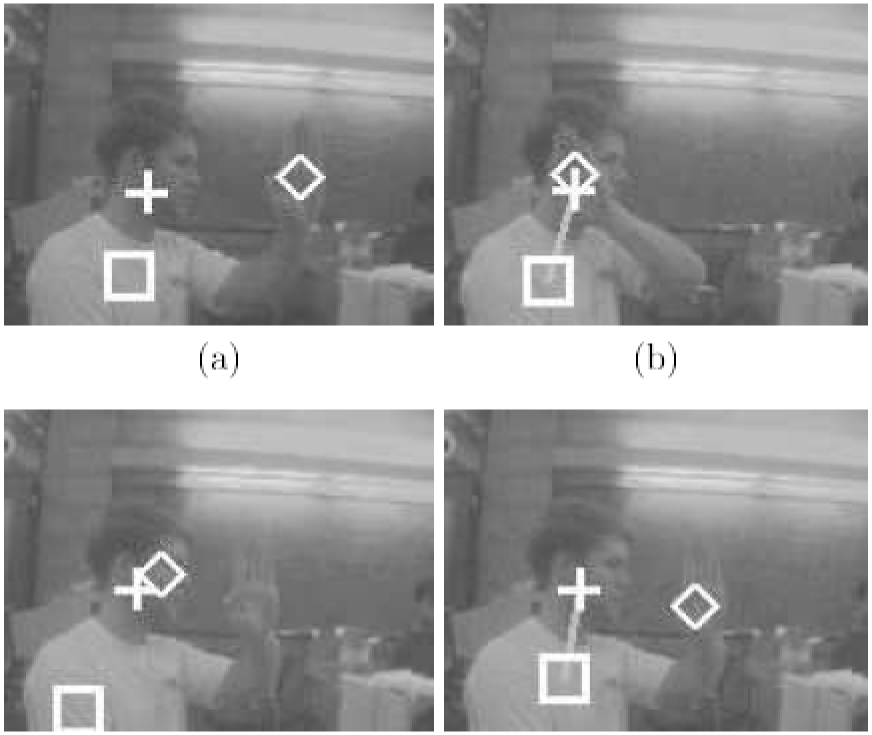




Variable Bandwidth Density-Based Fusion VBDF (Comaniciu '03)

JPDAF

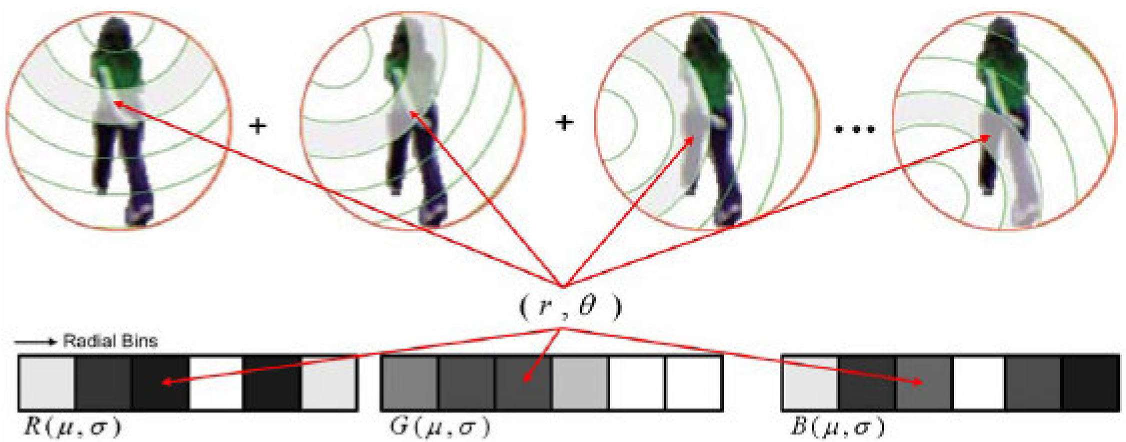
- Joint Probability Data Association Filter
  - Bar-Shalom & Fortmann Tracking and Data Association 1988
- Hager & Rasmussen 98
- Tracking a single object using
- Multiple observations



(a) (b) (c) (d)

Kang et al. Tracking people in crowded scenes

- Kalman filter for predicting position (constant velocity) in both image and ground plane
  - Using calibration
- Mean colour (RGB) representation in each annulus bin around 8 control points
  - Comparison by cross correlation
- Maximize joint probability motion & colour
  - Joint Probability Data Association Filter
- Foreground blobs based on BGS



- Independent tracking of all objects
- 1 fps
- Does not deal with splits & merges
  - seems to require clean initialization

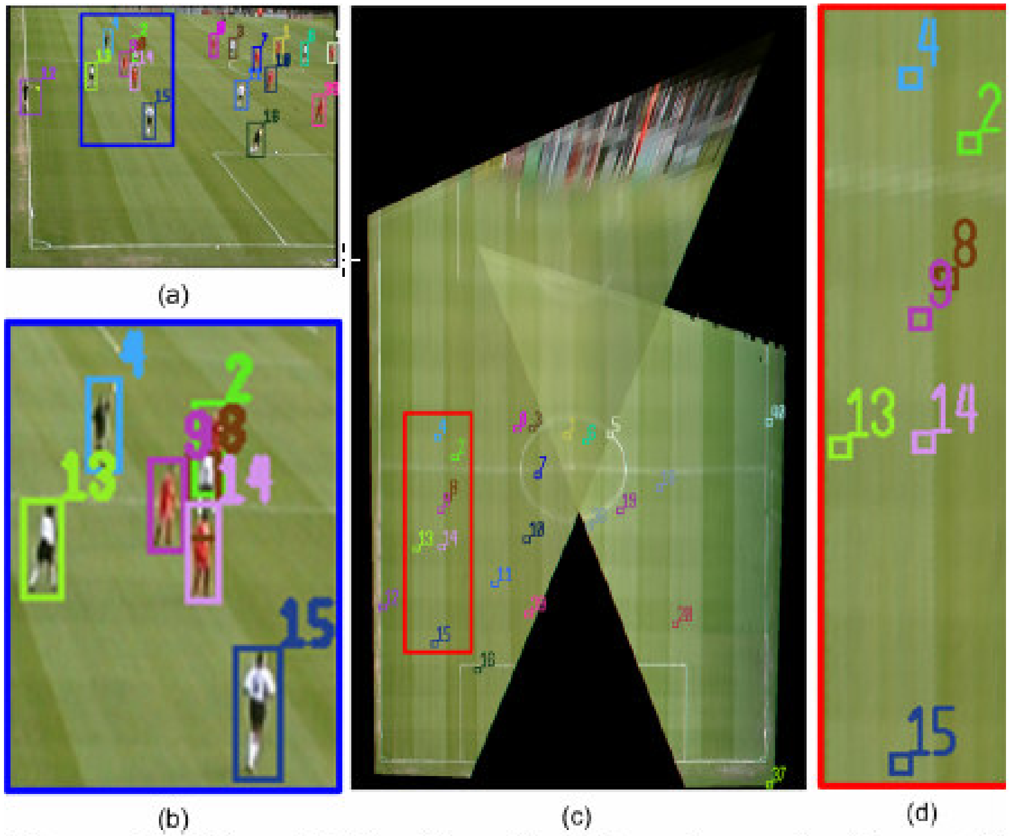
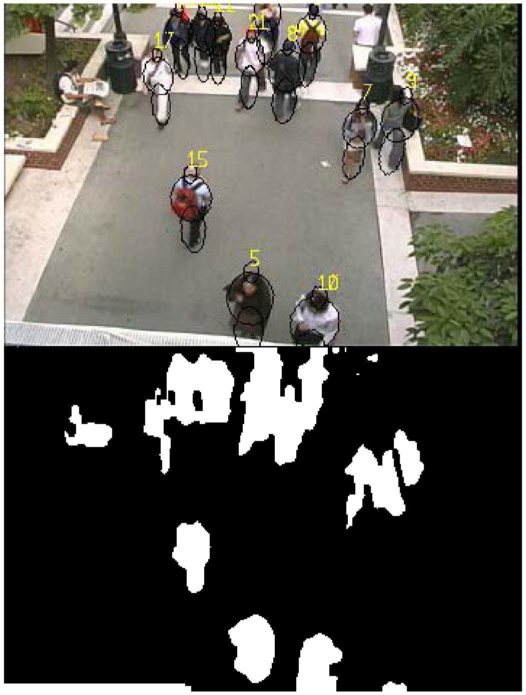


Figure 3. Using 3D for disambiguating cluttered objects. (a) The original frame, (b) Zoom of the most crowded region in the original frame (blue box), (c) The top-down view of the original registered frames, (d) Zoom of the corresponding crowded region from the top-down view (red box).

Tracking in crowds



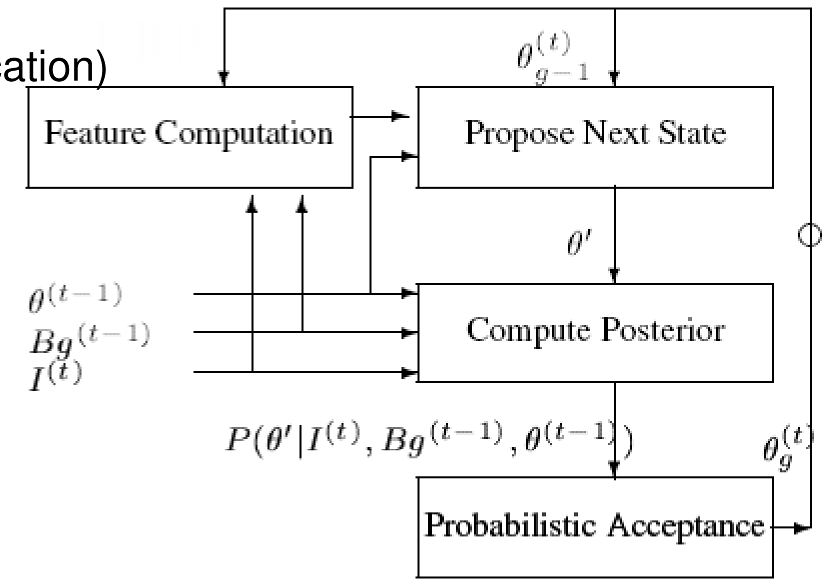
- Use model and calibration for dense scenes
  - Head + Torso + Legs as 3 ellipses
  - Parameterized by 2D head position and height (implying ground plane location) plus thickness and inclination
- Kalman filters for prediction
- Uses sources and sinks
- Image match for a given hypothesis
  - Background exclusion and object attraction

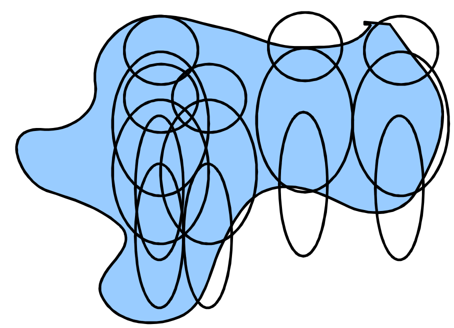
$$P(I^{S_i}|\mathbf{m}_i) \propto \exp\{\underbrace{-\lambda_b|S_i|B(\mathbf{p}_i,\mathbf{d}_i)}_{(1)} + \underbrace{\lambda_f|S_i|B(\mathbf{p}_i,\tilde{\mathbf{p}}_i)}_{(2)}\}$$

  - Mean-shift formulation to predict new location
- Multiple-hypotheses explained with Markov-Chain Monte Carlo

Markov Chain Monte Carlo

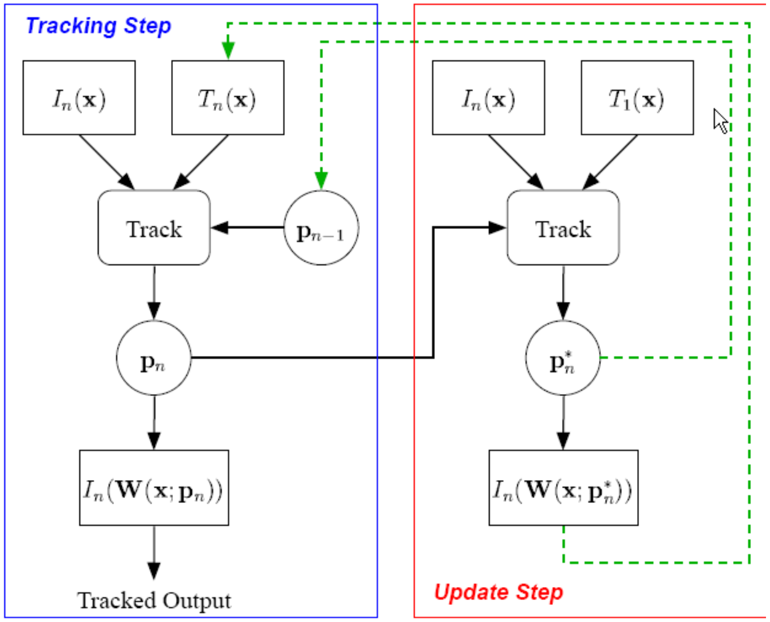
- Current compound state θ
  - (contains all people and locations, together with track history)
- Propose θ' by modifying θ
  - 0.1 Add person (in a sensibly sampled location)
    - Ω head & shoulder curves
    - Unexplained foreground regions
  - 0.1 Remove a person (uniformly)
  - 0.1 Establish correspondence
    - Between new object & dead object
  - 0.1 Break correspondence
  - 0.1 Exchange identity
  - 0.5 Update parameters
    - By mean shift
    - By moving head to head candidate location
- 300 iterations per frame
  - (1000 iterations on isolated frames without history)





Model update problem

- Tracking by fitting model
  - But object appearance changes
  - Lighting, pose, expression, as well as scale, orientation, location
- Constrain by using a general model of class (e.g. Faces, cars)
- Update model
  - Risk of updating model to include tracking errors (drift onto background, other objects
- “The Template Update Problem” Matthews, Ishikawa, Baker PAMI 2004
  - Maintain the original template and align that with the current model
  - Helps to avoid losing track



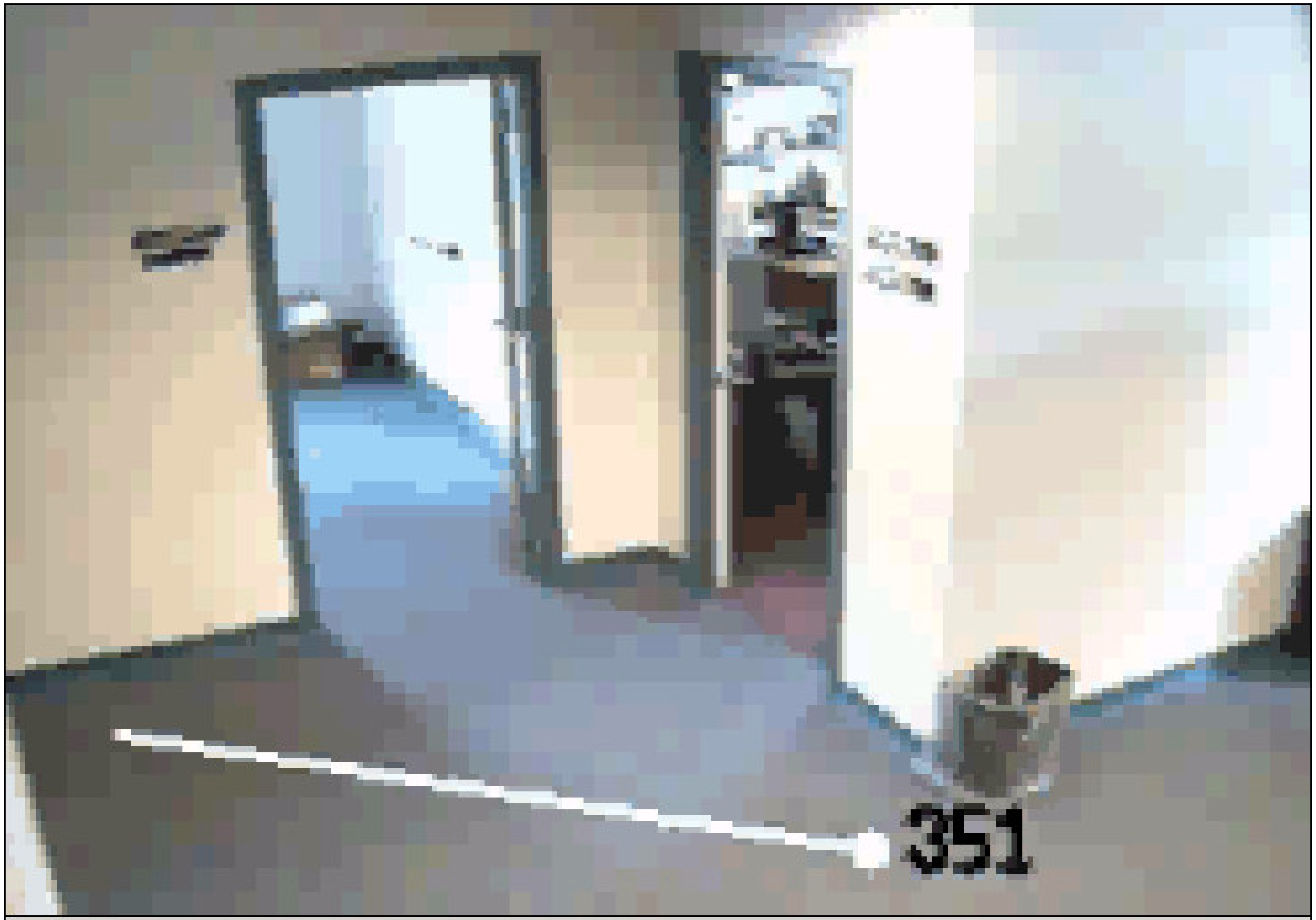
Tracking-based alert detection

- Simple rules on behaviour w.r.t. Geometric primitives
  - Direction of motion
  - Tripwire
  - Region

Tripwire

Tripwires

Directional Motion



MotionDirectionDetectionCrop.avi

References

- Fusion of Multiple Tracking Algorithms for Robust People Tracking Nils T Siebel and Steve Maybank ECCV 2002
- **Real-Time Tracking of Non-Rigid Objects using Mean Shift (2000)** Comaniciu, Ramesh and Meer
- http://www.robots.ox.ac.uk/~ab/abstracts/ijcv98.html **CONDENSATION -- conditional density propagation for visual tracking** Michael Isard and Andrew Blake Int. J. Computer Vision, 29, 1, 5--28, (1998)
- **A Bayesian algorithm for tracking multiple moving objects in outdoor surveillance video** Manjunath Narayana Donna Haverkamp CVPR 2007
- **Joint probabilistic techniques for tracking multi-part objects** Christopher Rasmussen, Gregory D. Hager CVPR 1998
- **Tracking Multiple Humans in Crowded Environment Tao Zhao Ram Nevatia CVPR 2004**
- **MULTI-PART TARGET REPRESENTATION FOR COLOR TRACKING** Emilio Maggio and Andrea Cavallaro ICIP 2005