# Visual Tracking & Particle Filters

Patrick Pérez

Irisa, Feb. 2012

technicolor

# What?

- Definition attempt : on-line or off-line extraction, from an image sequence, of state trajectories that characterize, either in image plane or in real world, some aspects of one or several target objects

- Types of targets (by increasing level of prior)
  - Picked objects: video object manually selected, interest points (corners, blobs), moving entities
  - Objects from a given category: cars, faces, people, etc.
  - Objects from a narrower category: moving cars, walking people, talking heads, face of a given person, a specific object

- Appearance models and inference machineries
  - Depend on tracking task
  - Heavily influenced by current trends in image processing and analysis

# Why?

Elementary or principal tool for multiple CV applications

A very large range of application domains, including

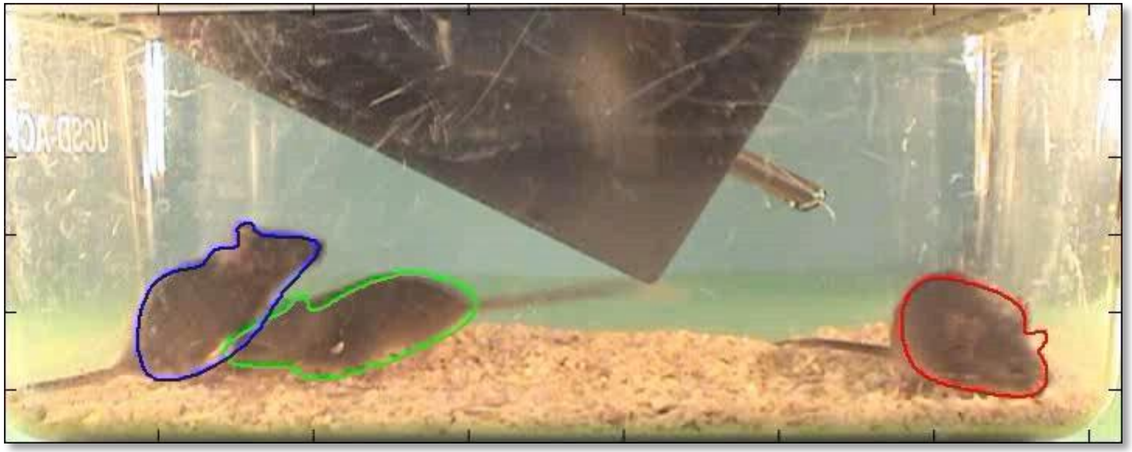
- Other sciences (neuroscience, ethology, biomechanics, sport, medicine, biology, fluid mechanics, meteorology, oceanography)
- Defense, surveillance, safety, monitoring, control, assistance
- Human-Computer Interfaces

Camera as a sensor (video content as a means)

- Video content production and post-production (compositing, augmented reality, editing, re-purposing, stereo-3D authoring, motion capture for animation, clickable hyper videos, etc.)
- Video content management (indexing, annotation, search, browsing)

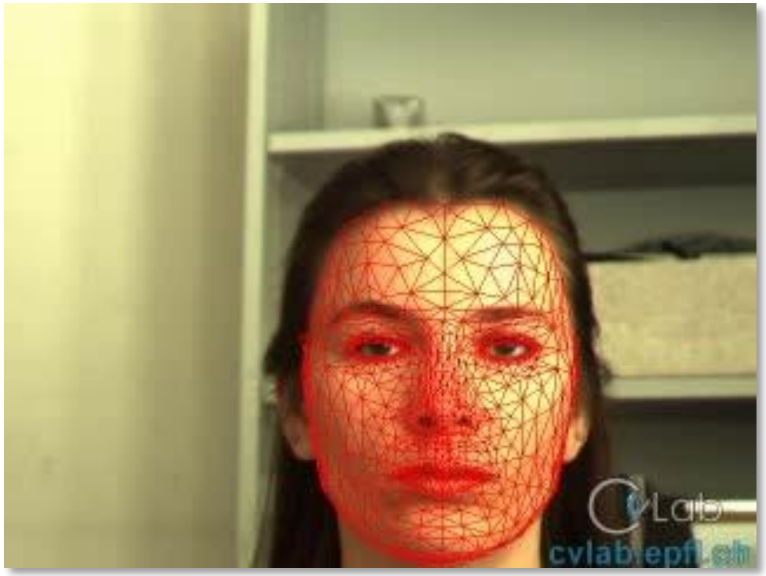
Video content as central object

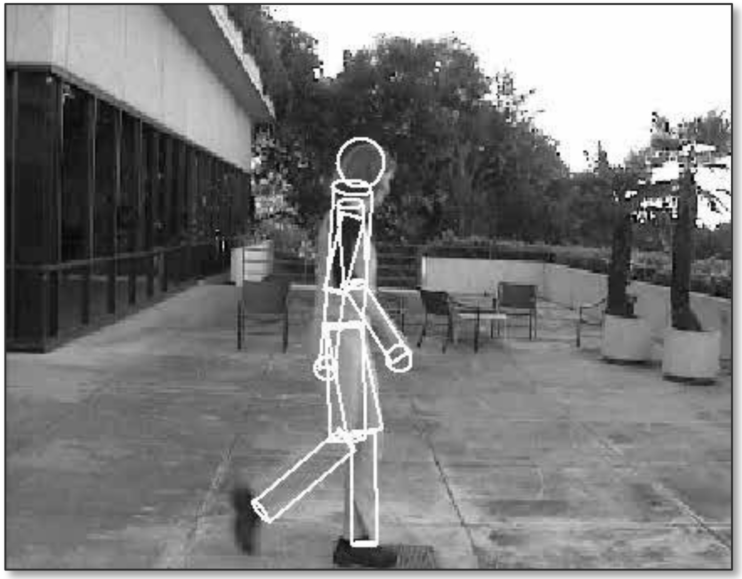
# With 2D dynamic shape prior

http://www2.imm.dtu.dk/~aam/tracking/
http://vision.ucsd.edu/~kbranson/research/cvpr2005.html

# With 3D shape prior



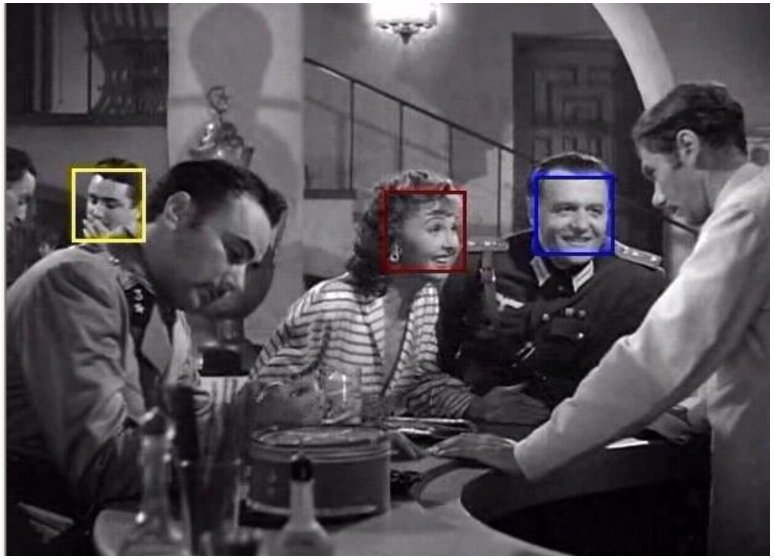
http://cvlab.epfl.ch/research/completed/realtime_tracking/
http://www.cs.brown.edu/~black/3Dtracking.html



technicolor

# With appearance prior

- in form of an object detector combined with on-line learning to distinguish among targets



http://www.cs.washington.edu/homes/xren/research/cvpr2008_casablanca/



technicolor

# With no appearance prior

- Tracking from user selection


http://server.cs.ucf.edu/~vision/projects/sali/CrowdTracking/index.html


technicolor

# With no appearance prior

- Tracking from user selection


http://www.robots.ox.ac.uk/~vgg/research/vgoogle/

# Sources of trouble

- Why is it harder that it might seem?
  - temporal variability of visual appearance
  - low video quality: low contrast, noise, motion blur
  - occlusions (partial to total) and clutter
  - unpredictable motions
  - constraints on computational complexity

http://comaniciu.net/

# Formalizing tracking

Image-based "measurements": $\mathbf{z}_t \in \Gamma$

- Raw or filtered images (e.g., intensities, colors, texture)
- Low-level features (e.g., edgels, corners, blobs, optical flow)
- High-level detections (e.g., face bounding boxes)

Single target "state": $\mathbf{x}_t \in \Lambda$

- Bounding box parameters (up to 6 DoF)
- Segmentation (pixel-wise labeling)
- 3D rigid pose (6 DoF)
- 2D/3D articulated pose (up to 30 DoF)
- 2D/3D deformation modes
- Discrete indices (identity, activity, visibility, expression, appearance examplars, etc.)

# Formalizing tracking

## Sequential tracking

- Given past and current measurements

$$\mathbf{z}_{1:t} := (\mathbf{z}_1 \cdots \mathbf{z}_t)$$

  output an estimate of current hidden state

$$\hat{\mathbf{x}}_t = \text{function}(\mathbf{z}_{1:t})$$

## Batch "tracking"

- Given batch of measurements $\mathbf{z}_{1:T}$
  output an estimate of all hidden states

$$\hat{\mathbf{x}}_t = \text{function}(\mathbf{z}_{1:T}),\ t = 1 \cdots T$$

technicolor

# Deterministic tracking

## Sequential tracking

- Optimization of ad-hoc objective function

$$\hat{\mathbf{x}}_t = \arg\min E(\mathbf{x}_t; \hat{\mathbf{x}}_{t-1}, \mathbf{z}_t)$$

- Or iterative minimization of function $E(\mathbf{x}_t; \mathbf{z}_t)$ initialized at $\hat{\mathbf{x}}_{t-1}$

## Batch "tracking"

- Optimization of ad-hoc compound objective function

$$\hat{\mathbf{x}}_{1:T} = \arg\min E(\mathbf{x}_{1:T}; \mathbf{z}_{1:T})$$

# Probabilistic tracking

## Hidden Markov chain/dynamic state space model

- Evolution model (dynamics), typically 1st-order Markov chain

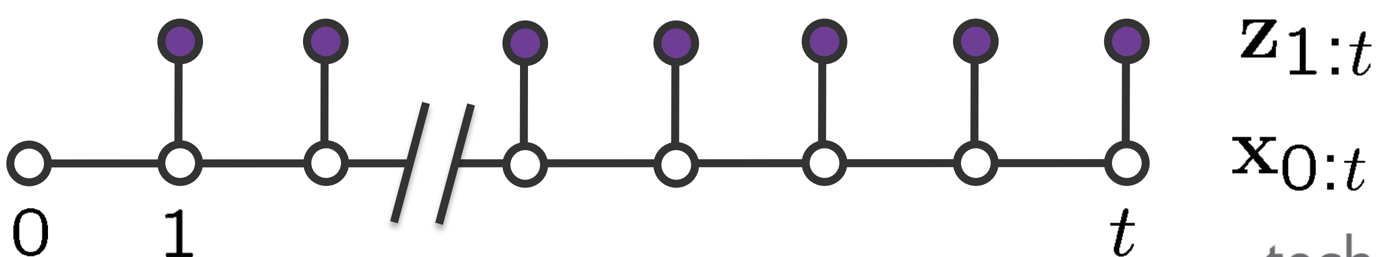
$$p(\mathbf{x}_i|\mathbf{x}_{1:i-1}) = p(\mathbf{x}_i|\mathbf{x}_{i-1})$$

- Observation model

$$p(\mathbf{z}_i|\mathbf{z}_{1:i-1}, \mathbf{x}_{0:i}) = p(\mathbf{z}_i|\mathbf{x}_i)$$

- Joint distribution

$$p(\mathbf{x}_{0:t}, \mathbf{z}_{1:t}) = p(\mathbf{x}_0) \prod_{i=1}^{t} p(\mathbf{x}_i|\mathbf{x}_{i-1}) p(\mathbf{z}_i|\mathbf{x}_i)$$



# Probabilistic tracking

## Sequential tracking

- Sequential MAP estimate: $\hat{\mathbf{x}}_t = \arg\max p(\mathbf{x}_t | \hat{\mathbf{x}}_{t-1}, \mathbf{z}_t)$

- Computation of the *filtering* pdf $p(\mathbf{x}_t | \mathbf{z}_{1:t})$, and point estimate:

$$\hat{\mathbf{x}}_t = \arg\max p(\mathbf{x}_t | \mathbf{z}_{1:t}) \text{ or } \mathbb{E}[\mathbf{x}_t | \mathbf{z}_{1:t}]$$

## Batch “tracking”

- Joint MAP estimate: $\hat{\mathbf{x}}_{1:T} = \arg\max p(\mathbf{x}_{1:T} | \mathbf{z}_{1:T})$

- Computation of *smoothing* pdf $p(\mathbf{x}_t | \mathbf{z}_{1:T})$, and point estimates:

$$\hat{\mathbf{x}}_t = \arg\max p(\mathbf{x}_t | \mathbf{z}_{1:T}) \text{ or } \mathbb{E}[\mathbf{x}_t | \mathbf{z}_{1:T}], \; t = 1 \cdots T$$

# Probabilistic filtering

- Various forms
  - (approximately) linear Gaussian: Kalman filters and variants
  - General case: sequential Monte Carlo approximation (*particle filter*)

- Pros: transports full distribution knowledge
  - Takes uncertainty into account (helps with clutter, occlusions, weak model)
  - Provides some confidence assessment
  - Allows more powerful parameter estimation

- Cons
  - More computations
  - Curse of dimensionality

# Limitation of KF for visual tracking

- Strong limitations on observations model
  - Measurements must be of same nature as (part of) state, e.g. detected object position
  - Measurement of interest must be identified (data association problem)

- In visual tracking, especially difficult
  - State specifies which part of data is concerned (actual measurement depends on hypothesized state)
  - Clutter is frequent

- Variants of KF (extended KF, unscented KF) can help, to some extent

# Useful deterministic trackers with no prior

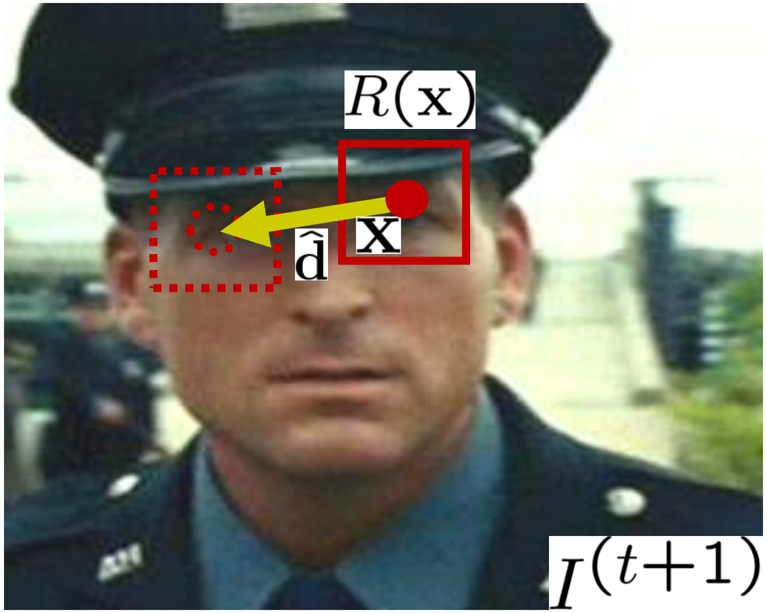
- "Point" (fragment) tracking
- Fragment-based object tracking
- Statistics-based object tracking

## Common denominators

- Iterative search initialized at previous estimate (static camera)
  - Successive linearizations

# Fragment tracking

- Problem: tracking "key points" (SIFT, SURF, STAR, RIFF, FAST), or random image patches, as long as possible
  - Input: detected/chosen patches
  - Output: *tracklets* of various life-spans




$$\hat{\mathbf{d}} = \arg\min_{\mathbf{d}} \underbrace{\sum_{\mathbf{p}\in R(\mathbf{x})} |I^{(t+1)}(\mathbf{p}+\mathbf{d}) - I^{(t)}(\mathbf{p})|^2}_{\text{SSD}}$$

technicolor

# Fragment-based tracking of arbitrary objects

- Track in next frame fragments from current bounding box
- Terminate weak tracklets
- Infer global motion of bounding box
- Select new fragments if necessary
- In effect: *part-based adaptive appearance model*

$$\hat{\mathbf{x}}_{t+1} = \hat{\mathbf{x}}_t + \text{robust average}(\mathbf{d}_1 \cdots \mathbf{d}_{n_t})$$

# Fragment-based tracking of arbitrary objects

- Can work really well (and fast), with accurate positioning

- Until
  - It drifts (due to partial occlusion, out-of-plane rotation)
  - It breaks down (diverging drift, total occlusion)

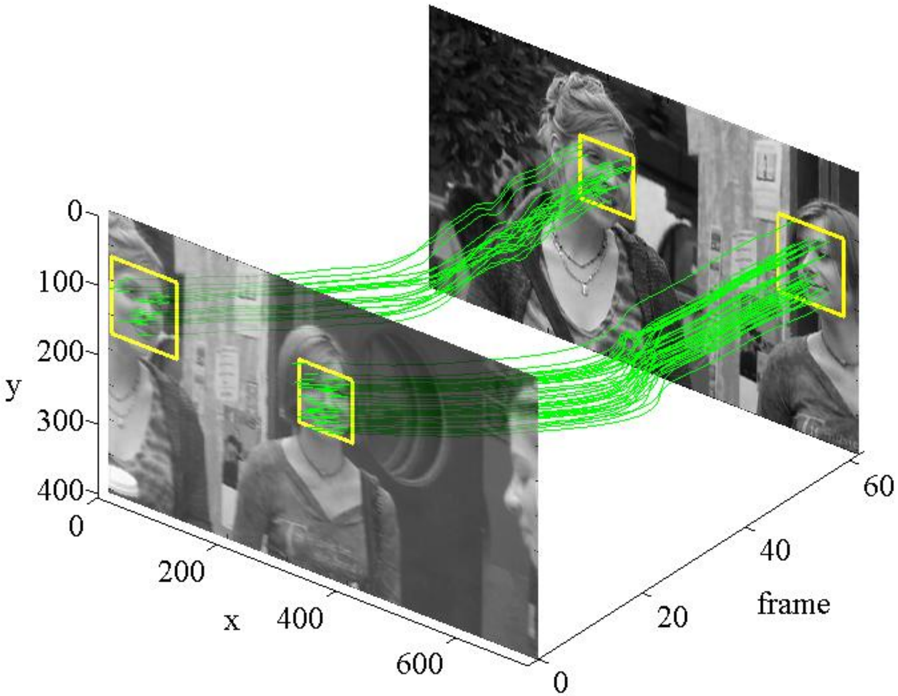


technicolor

# Face grouping

## For face recognition in movies and TV series

- *Detect* faces in each frame
- Connect faces traversed by sufficient fraction of tracklets





http://www.robots.ox.ac.uk/~vgg/research/nface/

# Statistics-based tracking of arbitrary objects

- Instead of pixel-wise appearance modeling, model appearance via global or semi-local statistics
- Examples
  - Texture statistics
  - Color and intensity distributions, possibly part-based
  - Intensities co-occurrences and co-variances
- Archetypical example: tracking with color histograms

http://comaniciu.net/



technicolor

# Color-based tracking and meanshift

- Global description of tracked region: color histogram
- Reference histogram with *B* bins

$$\mathbf{q}^* = (q_u^*)_{u=1\cdots B}$$

set at track initialization

- Candidate histogram at current instant

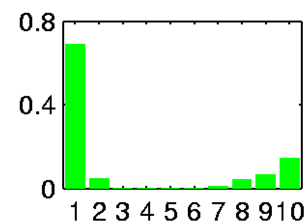
$$\mathbf{q}(\mathbf{x}) = (q_u(\mathbf{x}))_{u=1\cdots B}$$

gathered in region $R(\mathbf{x})$ of current image.

- At each instant

$$\hat{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}} \mathrm{dist}(\mathbf{q}^*, \mathbf{q}(\mathbf{x}))$$

  - searched around $\hat{\mathbf{x}}_t$
  - iterative search initialized with $\hat{\mathbf{x}}_t$






technicolor

# Color-based tracking and meanshift

- Global description of tracked region: color histogram
- Reference histogram with *B* bins

$$\mathbf{q}^* = (q_u^*)_{u=1\cdots B}$$

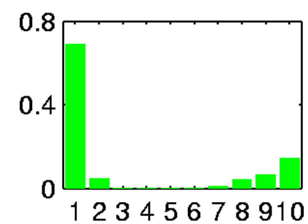
  set at track initialization
- Candidate histogram at current instant

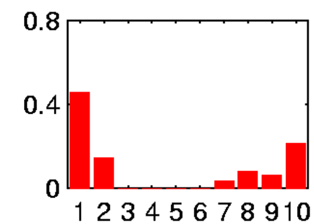
$$\mathbf{q}(\mathbf{x}) = (q_u(\mathbf{x}))_{u=1\cdots B}$$

  gathered in region $R(\mathbf{x})$ of current image.
- At each instant

$$\hat{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}} \mathrm{dist}(\mathbf{q}^*, \mathbf{q}(\mathbf{x}))$$

  - searched around $\hat{\mathbf{x}}_t$
  - iterative search initialized with $\hat{\mathbf{x}}_t$



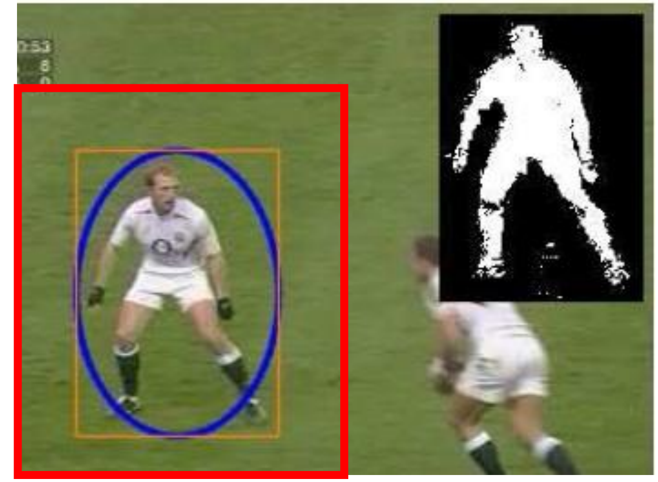
# Pros and cons

- Low computational cost (easily real-time)
- Surprisingly robust
  - Invariant to pose and viewpoint
  - Often no need to update reference color model

- Invariance comes at a price
  - Position estimate prone to fluctuation
  - Scale and orientation not well captured
  - Sensitive to color clutter (e.g., teamates in team sports)
- Deterministic local search challenged by
  - abrupt moves
  - occlusions

# Variants

- Remove background corruption in reference
  - Simple segmentation based on surrounding color at initialization
  - Re-estimation of foreground model
  - Amounts to zero bins for colors more frequent in surrounding than in selection



- Scale/orientation estimation
  - Originally: greedy search around current scale/orientation
  - Afterwards: incorporate loose spatial layout (via multiple spatial kernels or spatial partionning with sub-models)
- Robustness to camera movement
  - Robust estimation of dominant apparent motion
  - Start search at previous position displaced according to dominant motion

# Particle filtering

- Monte Carlo based on sequential importance sampling

- History
  - Gordon 1993, *Novel approach to non-linear/non-Gaussian Bayesian state estimation*
  - Kitagawa 1996, *Monte Carlo filter and smoother for non-Gaussian nonlinear state space models*
  - Isard et Blake 1996, *CONDENSATION: CONditional DENSity propagATION for visual tracking*

- Reason of success in CV
  - Visual tracking often implies multimodal filtering distributions
  - PF maintains multiple hypothesis: more robust to occlusion and temporary loss
  - Easy to implement and little restrictions on model ingredients

# Generic synopsis

- Given $\{(\mathbf{x}_{0:i-1}^{(m)}, \pi_{i-1}^{(m)})\}_{m=1\cdots M}$
- One step proposal

$$\tilde{\mathbf{x}}_i^{(m)} \sim q(\mathbf{x}_i|\mathbf{x}_{i-1}^{(m)}, \mathbf{z}_i),\ m = 1 \cdots M$$

- Weights update

$$\tilde{\pi}_i^{(m)} \propto \pi_{i-1}^{(m)} \frac{p(\mathbf{z}_i|\mathbf{x}_i^{(m)})p(\mathbf{x}_i^{(m)}|\mathbf{x}_{i-1}^{(m)})}{q(\mathbf{x}_i^{(m)}|\mathbf{x}_{i-1}^{(m)}, \mathbf{z}_i)} \text{ avec } \sum_{m=1}^{M} \tilde{\pi}_i^{(m)} = 1$$

- Resampling
  - If $\sum_{m=1}^{M} \tilde{\pi}_i^{(m)2} > M_{\text{seuil}}^{-1}$

$$\forall m,\ a_m \sim \sum_{k=1}^{M} \tilde{\pi}_i^{(k)} \delta_k,\ \mathbf{x}_{1:i}^{(m)} = (\mathbf{x}_{1:i-1}^{(a_m)}, \tilde{\mathbf{x}}_i^{(a_m)}) \text{ et } \pi_i^{(m)} = \frac{1}{M}$$

  - Otherwise

$$\forall m,\ \mathbf{x}_{1:i}^{(m)} = (\mathbf{x}_{1:i-1}^{(m)}, \tilde{\mathbf{x}}_i^{(m)}) \text{ et } \pi_i^{(m)} = \tilde{\pi}_i^{(m)}$$

- Monte Carlo approximation

$$\mathbb{E}[f(\mathbf{x}_i)|\mathbf{z}_{1:i}] \approx \sum_{m=1}^{M} \pi_i^{(m)} f(\mathbf{x}_i^{(m)})$$

# Proposal density

- Optimal density

$$q(\mathbf{x}_i|\mathbf{x}_{i-1},\mathbf{z}_i) = p(\mathbf{x}_i|\mathbf{x}_{i-1},\mathbf{z}_i) = \frac{p(\mathbf{z}_i|\mathbf{x}_i)p(\mathbf{x}_i|\mathbf{x}_{i-1})}{p(\mathbf{z}_i|\mathbf{x}_{i-1})}$$
$$\Rightarrow \pi_i^{(m)} \propto \pi_{i-1}^{(m)}\, p(\mathbf{z}_i|\mathbf{x}_{i-1}^{(m)}) \text{ with } \textstyle\sum_{m=1}^{M} \pi_i^{(m)} = 1$$

  usually not accessible

- Bootstrap filter: classic for its simplicity (but often confused with general SIS)

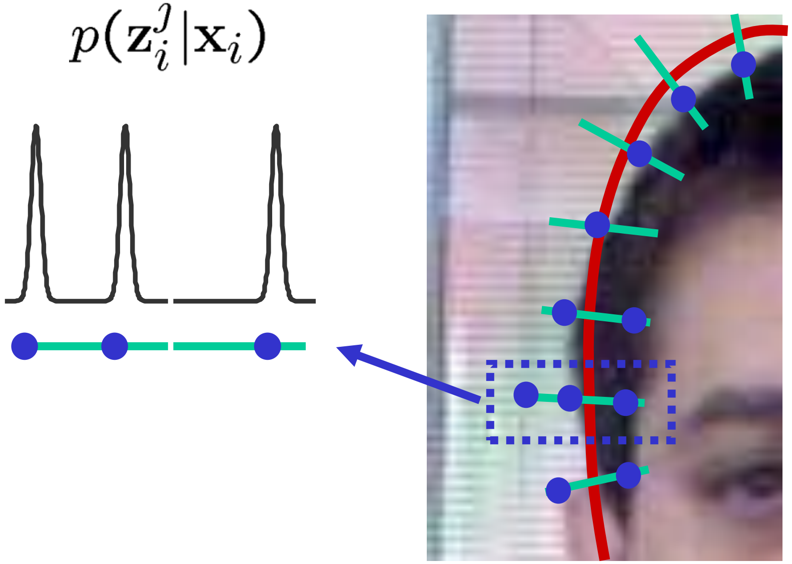
$$q(\mathbf{x}_i|\mathbf{x}_{i-1},\mathbf{z}_i) = p(\mathbf{x}_i|\mathbf{x}_{i-1})$$
$$\Rightarrow \pi_i^{(m)} \propto \pi_{i-1}^{(m)}\, p(\mathbf{z}_i|\mathbf{x}_{i}^{(m)}) \text{ with } \textstyle\sum_{m=1}^{M} \pi_i^{(m)} = 1$$

- In-between: try and use current data for better efficiency

# "CONDENSATION"

- State: active shape model with autoregressive dynamics
- Observation model: based on edgels near hypothesized silhouette
- Bootstrap filter: proposal and dynamics coincide

$$p(\mathbf{z}_i^j|\mathbf{x}_i)$$



[Isard and Blake, IJCV 29(1), 1998]



technicolor

# Color-based PF

- Based on color histogram similarities, following Comaniciu's idea
- Bootstrap filter and data model $p(\mathbf{z}_t|\mathbf{x}_t) \propto \exp \lambda\rho[\mathbf{q}(\mathbf{x}_t), \mathbf{q}^*]$

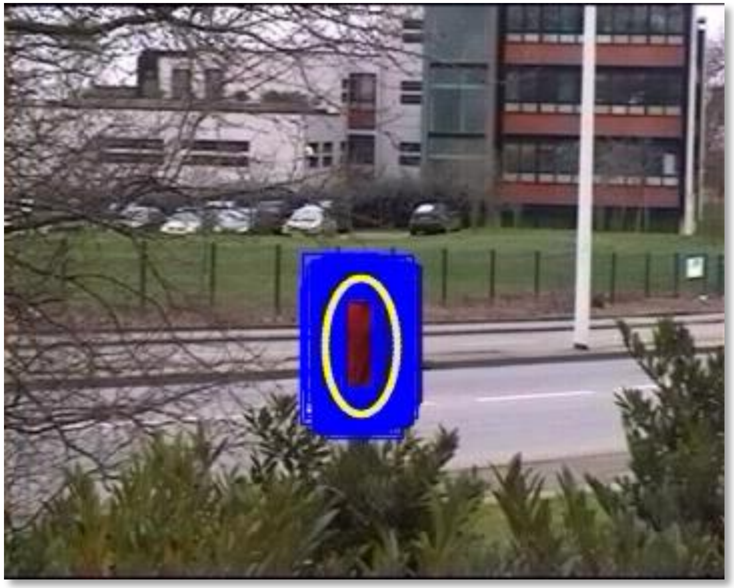
[Perez et al. ECCV'02]



technicolor

# Color-based PF

- Based on color histogram similarities, following Comaniciu's idea
- Bootstrap filter and data model $p(\mathbf{z}_t|\mathbf{x}_t) \propto \exp \lambda\rho[\mathbf{q}(\mathbf{x}_t), \mathbf{q}^*]$



[Perez et al. ECCV'02]

# Proposal densities

- Exploit current (even future) data
  - Get close to optimal density with approximation or iterative search (beware though!)
  - Use *detection* in proposal: mixture centered on detections and modification of dynamics to allow jumps

$$q(\mathbf{x}_i|\mathbf{x}_{i-1},\mathbf{z}_i) = \beta p(\mathbf{x}_i|\mathbf{x}_{i-1}) + \frac{1-\beta}{D_i}\sum_{d=1}^{D_i}\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_d,\Gamma_d)$$

$$p(\mathbf{x}_i|\mathbf{x}_{i-1}) = \nu p_{\mathsf{smooth}}(\mathbf{x}_i|\mathbf{x}_{i-1}) + (1-\nu)\mathcal{U}_\Lambda(\mathbf{x}_i)$$

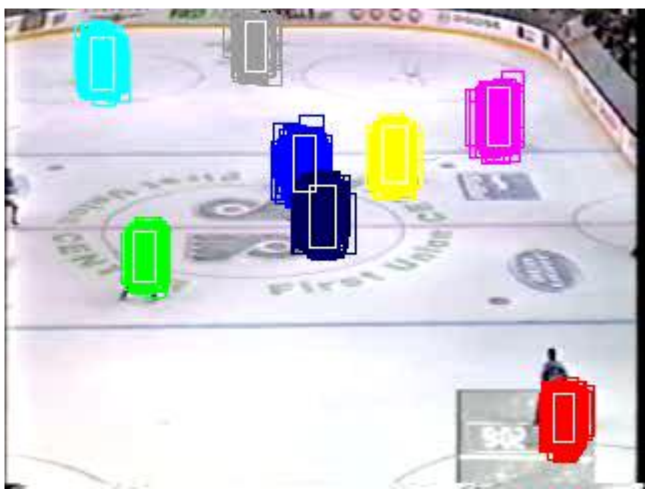
- Exploit model graphical structure (esp. for higher dim.)
  - Rao-Blackwellisation (FP on part of state-space, conditional KF on other)
  - Exploits exact or approximate conditional decoupling between state parts
    - Factored, hierarchical, layered sampling...

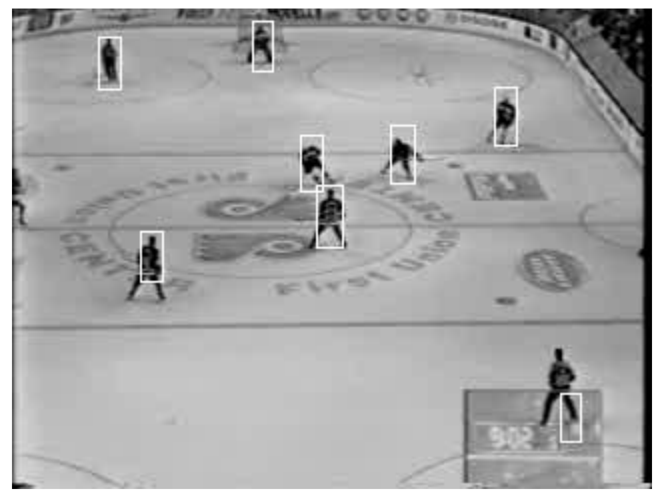
technicolor

# MOT PF and detection-based proposal

- Color-based detection


- Object category detection





[Vermaak et al. ICIP'05][Okuma et al. ECCV'04]

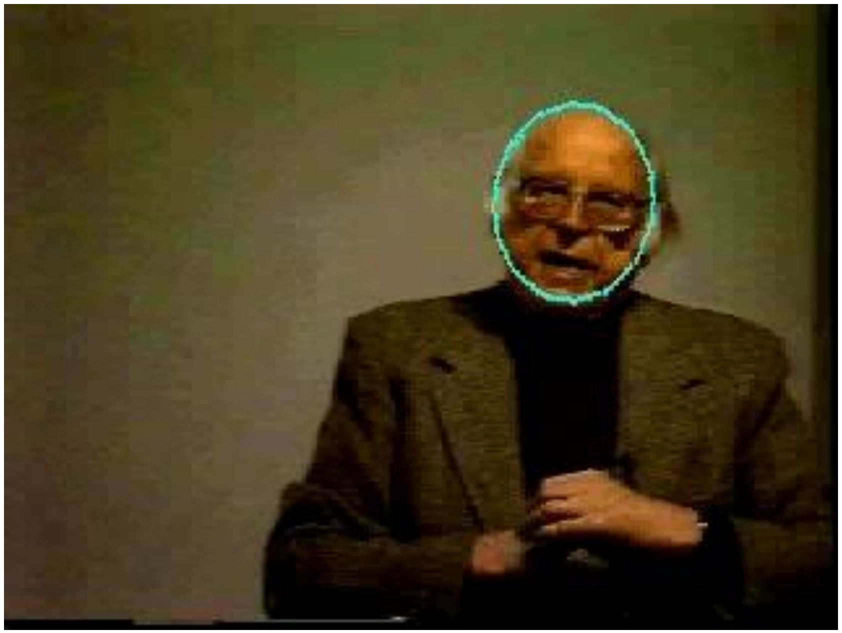
# PF with multiple cues

- Complementary cues for improved robustness
  - Persistent though ambiguous (e.g., color) vs. precise though transient (ex: movement)
  - Sensitivity to different clutter, invariant to different perturbations (e.g., global color, local intensity, contours)
- Bayesian fusion often under conditional independency assumption

$$p(\mathbf{z}_{1,i} \cdots \mathbf{z}_{A,i} | \mathbf{x}_i) = \prod_{a=1}^{A} p(\mathbf{z}_{a,i} | \mathbf{x}_i)$$

- Proposal can exploit specificities of different cues

# PF with multiple cues



[Wu and Huang, ICCV’01][Gatica-Perez *et al.*, 2003]

# On-line adaptation and learning

- Goal: update/expand appearance model on the fly for robustness to unexpected changes (on target and/or environment), esp. if no off-line knowledge on appearance

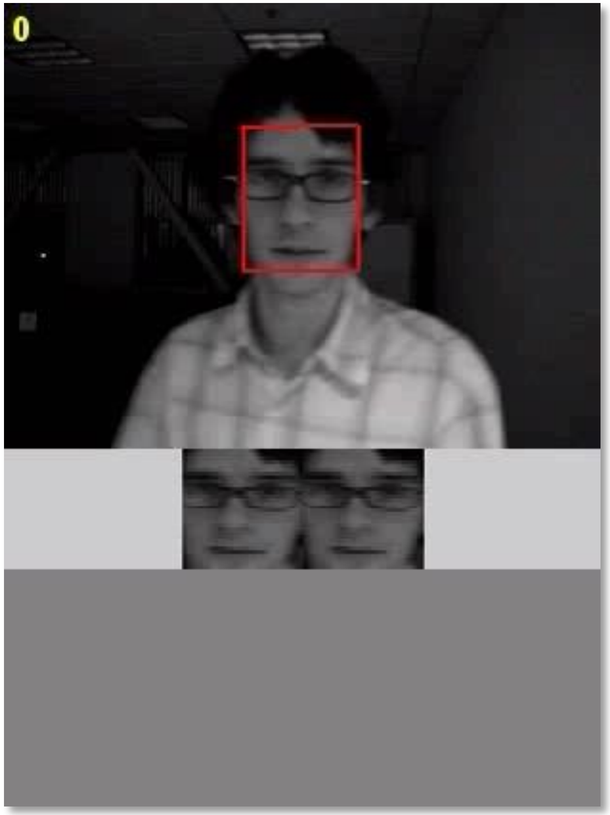
[Jepson et al. PAMI 25(10), 2003]

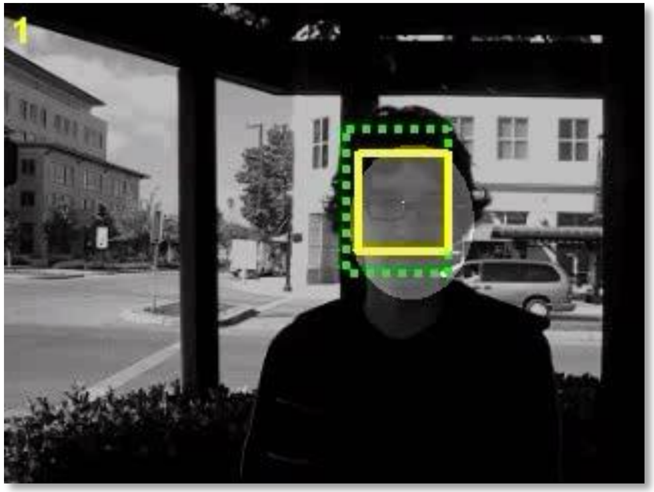
- Problem: *drift* if adaptation too rapid, esp. during occlusions
- Some (insufficient) solutions
  - Tunable learning rate
  - Adaptation conditioned on global monitoring
  - Adaptation on one type of measurement if others suffice for tracking (anchoring)

technicolor

# On-line subspace learning

- One example: Ross *et al.*
  - Constant time PCA update with new data, with *learning rate* $\alpha \sim 0.02$
  - Robust metric to account for background corruption
  - Tracking with particle filter



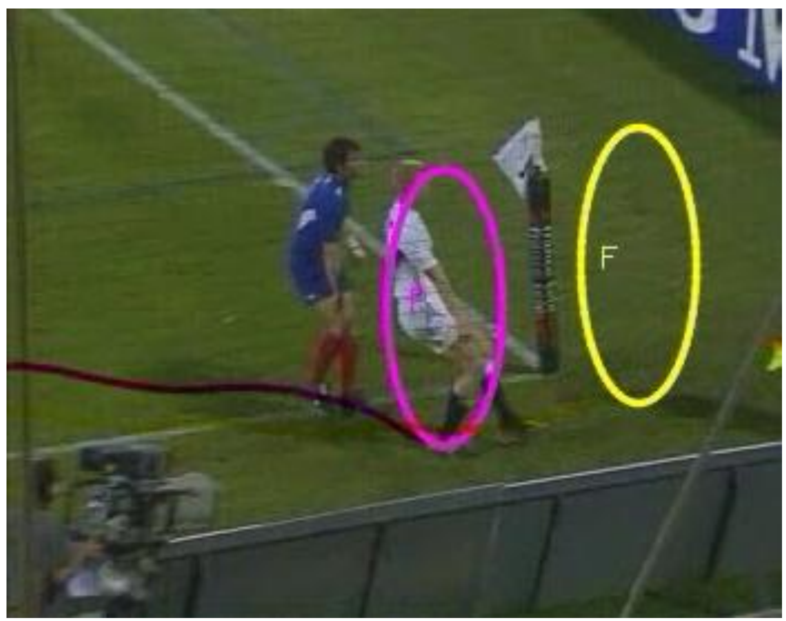




http://www.cs.toronto.edu/~dross/ivt/

technicolor

# Adaptation with external monitoring

- Color model update during zooms

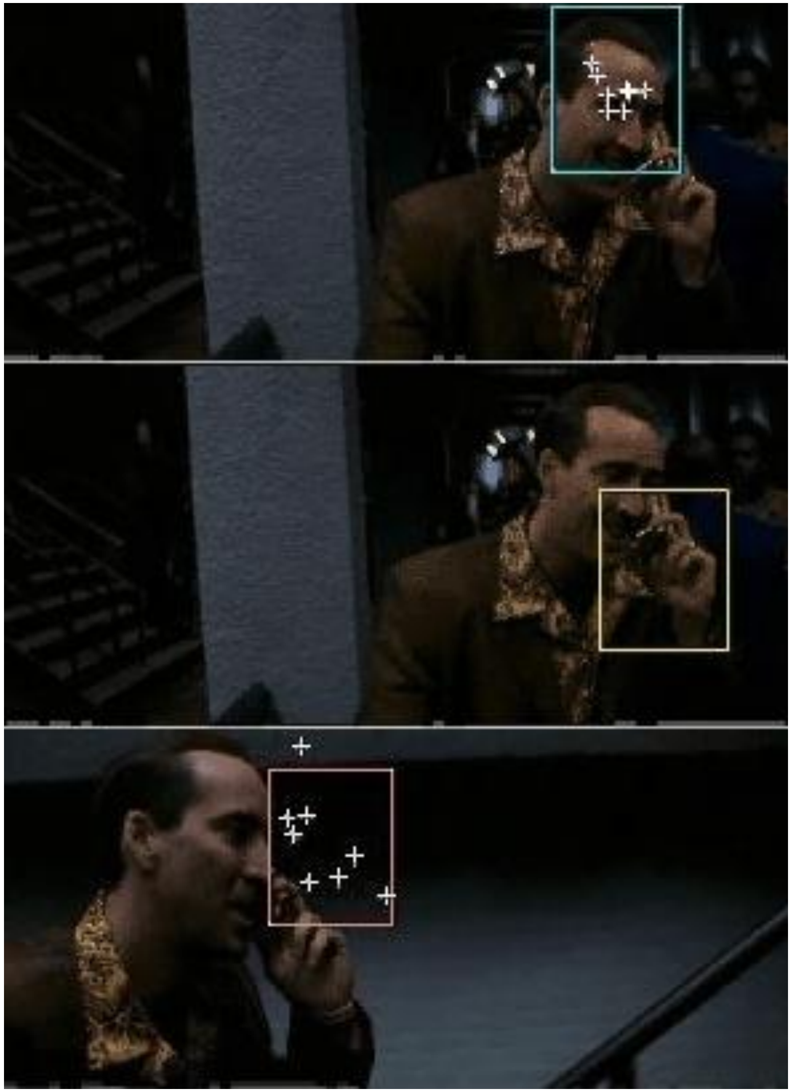



[Lehuger et al. ICIP'06]



technicolor

# Adaptation with multiple cues



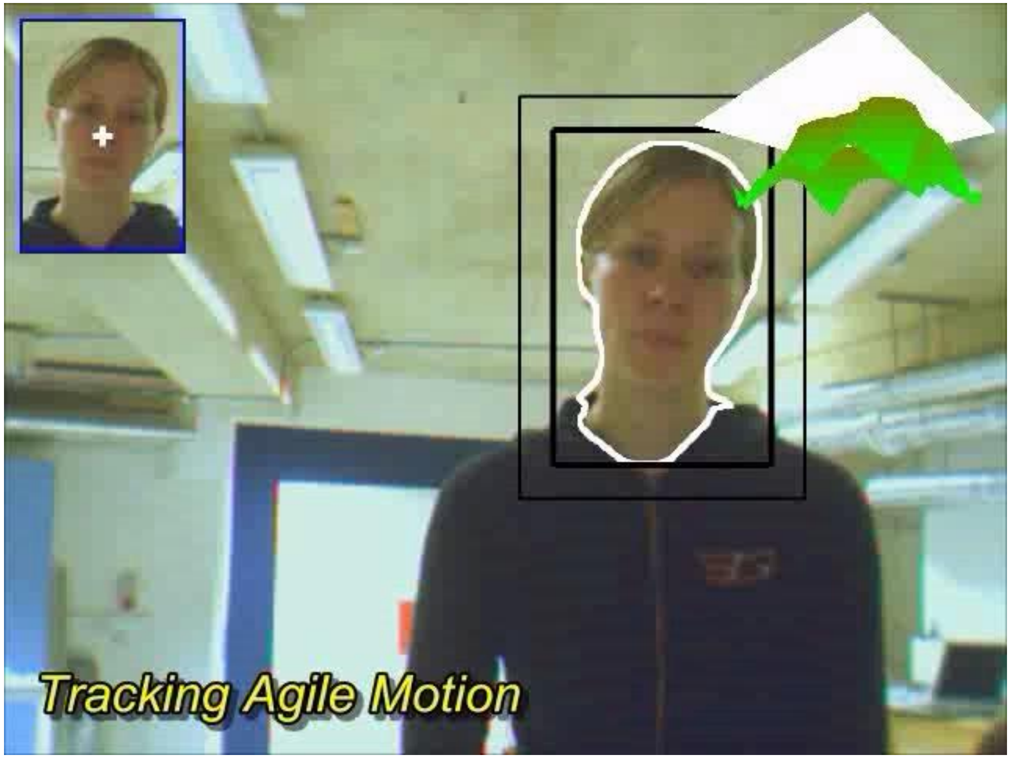
[Badrinarayanan et al. ICCV'07]



technicolor

# High dimensions…

- Interactions between parts of state space
  - Through evolution and/or observation
  - Permanently or intermittently
- Articulated objects
  - Admissible values (joints limits)
  - Kinematics
- MOT
  - Intermittent interaction through observation
- Segmentation
  - MRF on pixel labels




http://www.robots.ox.ac.uk/~cbibby/index.shtml