Target tracking in the image sequence

Václav Hlaváč

Czech Technical University, Faculty of Electrical Engineering
Department of Cybernetics, Center for Machine Perception
121 35 Praha 2, Karlovo nám. 13, Czech Republic

hlavac@fel.cvut.cz, http://cmp.felk.cvut.cz

Lecture plan

1. Tracking, task formulation.
2. Motion capture.
3. Mean shift tracking.

Tracking, informally

- The aim is to continuously locate a target region in a videosequence.
- It is usually assumed that the target region to be tracked is given in the initial frame.
- The target trajectory is built incrementally by repeatedly estimating relative displacement of a the target between successive frames.
- The target or the camera or both are moving.

Courtesy: Pavel Krsek

- *Tracking a car (manually selected target) from a helicopter.*
- *The trajectory serves for image stabilization.*

Tracking, alternative problem formulation

- Following a moving target through the image sequence.
- Estimating the target location in a new frame from the information in previous frames.
- Refining the target position using the estimate.

Tracking involves two basic problems:

Motion predicts the limited search region in which the tract target is likely to be in the next frame.

Matching (also detection or localization) identifies the target within the designed search region.

Difficulties in tracking

Tracking of a general object in a dynamically changing environment is difficult because of:

- The object appearance changes due to illumination variation, object motion, and object scaling.
- The object 3D pose variations and information loss due to the perspective projection.
- Partial and full object occlusions.
- Background clutters.
- Similar objects from the same class as the tracked object (ambiguity, e.g., when tracking landmarks).

Generative vs. discriminative tracking

Discriminative tracking methods

- Formulate the tracking as a classification problem.
- The trained classifier is used to discriminate the target from background.
- The classifier can be online updated during the tracking procedure.

Generative tracking methods

- Represent the target observations as an appearance model.
- The tracking problem is formulated as searching for the region within the highest probability generated from the appearance model.
- Some works update the target appearance model incrementally for adapting to dynamic environmental changes and target appearance variations.

Motion capture

- Simplifies the situation by putting markers on a tracked object / subject.
- Used in movies (virtual characters), computer games, medical rehabilitation, car crash tests, . . .

Motion capture – principles

optical

magnetic

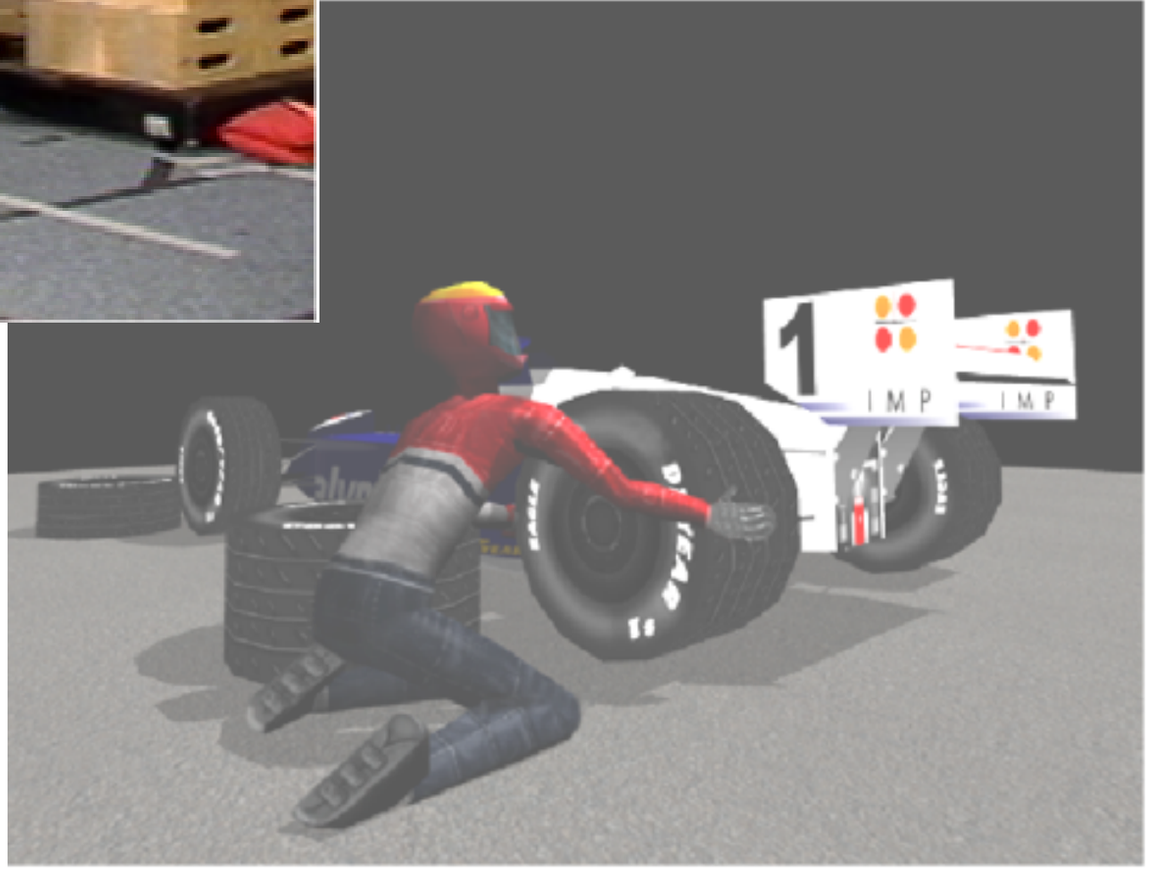
mechanical

Courtesy: Mark Mayer, Caltech

Motion capture – in movies, games

- Collect sensor data (triangulate optical data or read magnetic fields or read mechanical joints).
- Fill in missing data.
- Use inverse kinematics (if necessary) to determine joint angles.
- Apply the motion to the computer graphics model of a character.

Motion capture – F1 result



Example – walk

Example – volleyball spike

Example – exercises

Example – reactions to pushes

Example – application in humanoid robotics

Tracking, applications

Applications:

- radar / planes
- cars
- pedestrians
- face features / expressions
- small particles

There are many ad-hoc approaches to tracking.

General probabilistic formulation: track model density over time.

Why is tracking difficult?

- Targets can change appearance while being tracked.
  - Targets may get occluded.
  - Illumination condition change over time which influences target/scene appearance. Occurred specularities cause qualitative changes in appearance which are difficult to handle.
  - Moving camera or fast moving target can cause blurred images.
  - Tracked region (target) is often a projection from a 3D object. The loss of information occurs.
  - Target of interest can vary drastically from texture-less targets to highly textured ones. Consequently, used cues (features) for tracking have to be chosen automatically.
- Due much data, the efficiency is an issue.

Four cues categories for distinguishing the target from the background

1. The target object and it appearance (e.g., edges, corners, texture).
2. Scene in which the target appears (e.g., various background models).
3. Discontinuities in the image function on the border between the tracked object and background.
4. Specific object motion (e.g., ballistic trajectory, jumping table tennis thrown obliquely).

Used models in tracking

- Simple dynamic models (second order dynamics).
- Kinematic models, e.g., human body is represented by a 'stick figure' with kinematic constraints in joints.
- Template (image, 2D or 3D geometric model), multiple templates.
- 2D spline curves, active contours.
- etc.

Tracking – theory

The theory is established in the control theory, i.e., analysis of stochastic systems.

Wiener filtration. ≈1940. Off-line. Input/output description. Gaussian distribution. Linear dynamics. Linear estimate optimal in the least square sense.

Kalman filtration. 1960. On-line. State description. The rest as Wiener filtration . . .

Extended Kalman filtration. ≈1970. Nonlinear dynamics. Linearization. The rest as Kalman filtration.

Condensation. 1996. (conditional density propagation). Copes with multimodal distributions.

Linear stochastic system

Deterministic case

$$\begin{aligned} x(t+1) &= A\,x(t) + B\,u(t) \\ y(t) &= A\,x(t) + D\,u(t) \end{aligned}$$

Linear stochastic system

$$\begin{aligned} x(t+1) &= A\,x(t) + B\,u(t) + v(t) \\ y(t) &= A\,x(t) + D\,u(t) + e(t) \end{aligned}$$

$v(t)$ – process noise, $e(t)$ – measurement noise. Random stationary sequences.

White noise assumption

Mean value

$$\mathcal{E}\left\{\begin{bmatrix} v(t) \\ e(t) \end{bmatrix}\right\} = 0$$

Covariance

$$\mathcal{E}\left\{\begin{bmatrix} v(t) \\ e(t) \end{bmatrix} \cdot \begin{bmatrix} v(t) \\ e(t) \end{bmatrix}^{\top}\right\} = \begin{bmatrix} Q & S \\ S^{\top} & R \end{bmatrix} \delta(t_1 - t_2)$$

Mean shift, principle

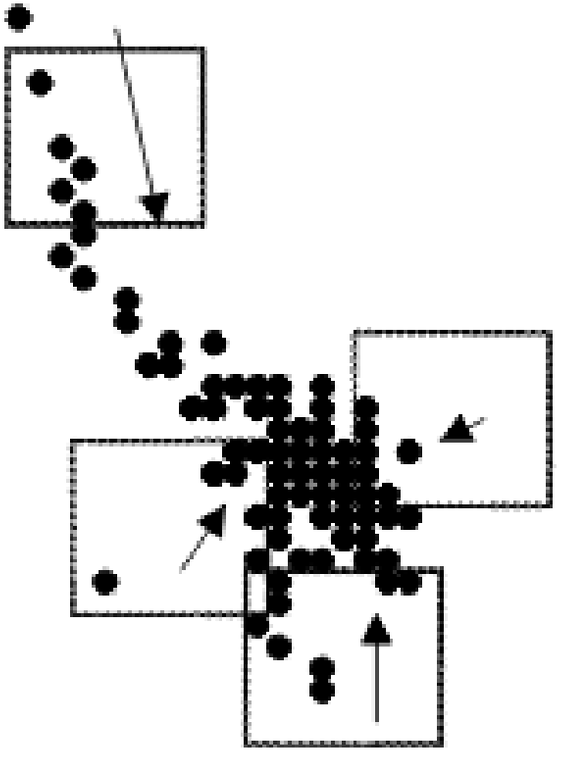
- We already introduced mean shift in the lecture about image segmentation.
- Estimation of the density gradient - Fukunaga K.: Introduction to Statistical Pattern Recognition, Academic Press, New York, 1972.
- Sample mean of local samples points in the direction of higher density. It provides the estimate of the gradient.
- Mean shift vector m of each point p

$$m = \sum_{i \in \text{window}} w_i \left(p_i - p\right), \quad w_i = \text{dist}(p, p_i)$$

- Based on the assumption that points are more and more dense as we are getting near the cluster 'central mass'.

Mean shift algorithm

- Input: points in the Euclidean (feature) space.
- Determine a search window size (usually small).
- Choose the initial location of the search window.
- Compute the mean location (centroid of the data) in the search window.
- Center the search window at the mean location computed in the previous step.
- Repeat until convergence.



Mean shift & Tracking

- Comaniciu D., Ramesh V., and P. Meer. Real-time tracking of non rigid objects using mean shift. In IEEE Conf. Computer Vision and Pattern Recognition, 2000.
- Comaniciu D., and Meer P. Mean shift: A robust approach toward feature space analysis. In IEEE Trans. Pattern Analysis and Machine Inteligence, 2002.
- Collins R. T. Mean-shift blob tracking through scale space. In IEEE Conf. Computer Vision and Pattern Recognition, 2003.

Target localization

- Target localization means finding the discrete location $\mathbf{y}$ whose associated density function p is the most similar to the target density q.
- Similarity measure is expressed by a metric derived from the Bayesian framework (statistical hypothesis testing).
- Minimizing distance

$$d(\mathbf{y}) = \sqrt{1 - \rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}]} \tag{1}$$

between two distributions is equivalent to maximizing the Bayes error, here expressed by the Bhattacharyya coefficient.

Object representation

- Visual features are represented by its discrete densities estimated from the m-bin color histogram.
- Target model distribution.
- Target candidate distribution.

Distance minimization

- Finding the most probable location $\mathbf{y}$ of the target in the current frame means maximizing the Bhattacharyya coefficient.

$$\hat{\rho}(\mathbf{y}) \equiv \rho[\hat{p}(\mathbf{y}), \hat{q}] = \sum_{u=1}^{m} \sqrt{\hat{p}_u(\mathbf{y})\hat{q}}, \tag{2}$$

- The mean shift algorithm is employed to find the maximum.

Mean shift tracking algorithm

Given: target model distribution $\{\hat{q}_u\}_{u=1\ldots m}$ and the estimation of the target location $\hat{\mathbf{y}}_0$ in the previous frame.

1. Compute the target candidate distribution $\{\hat{p}_u(\hat{\mathbf{y}}_0)\}_{u=1\ldots m}$ at the location $\hat{\mathbf{y}}_0$ in the current frame and calculate the Bhattacharyya coefficient.
2. Compute weights $\{w_i\}_{i=1\ldots n_h}$.
3. Derive the new location of the target using the mean shift vector computed over the area of the employed kernel. Recalculate the distribution $\{\hat{p}_u(\hat{\mathbf{y}}_1)\}_{u=1\ldots m}$, and compute the Bhattacharyya coefficient at the new location.
4. If $\rho[\hat{p}(\hat{\mathbf{y}}_1), \hat{q}] < \rho[\hat{p}(\hat{\mathbf{y}}_0), \hat{q}]$ then set $\hat{\mathbf{y}}_1$ equal $\frac{1}{2}(\hat{\mathbf{y}}_0 + \hat{\mathbf{y}}_1)$.
5. Repeat Steps 3 through 5 until convergence, i.e., $\|\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_0\| < \epsilon$, yielding the position of the feature in the current frame. Set $\hat{\mathbf{y}}_0$ equal to $\hat{\mathbf{y}}_1$ and proceed with Step 1.

Illustration – Original sequence

Illustration – Result of tracking

Illustration – Weights w_i

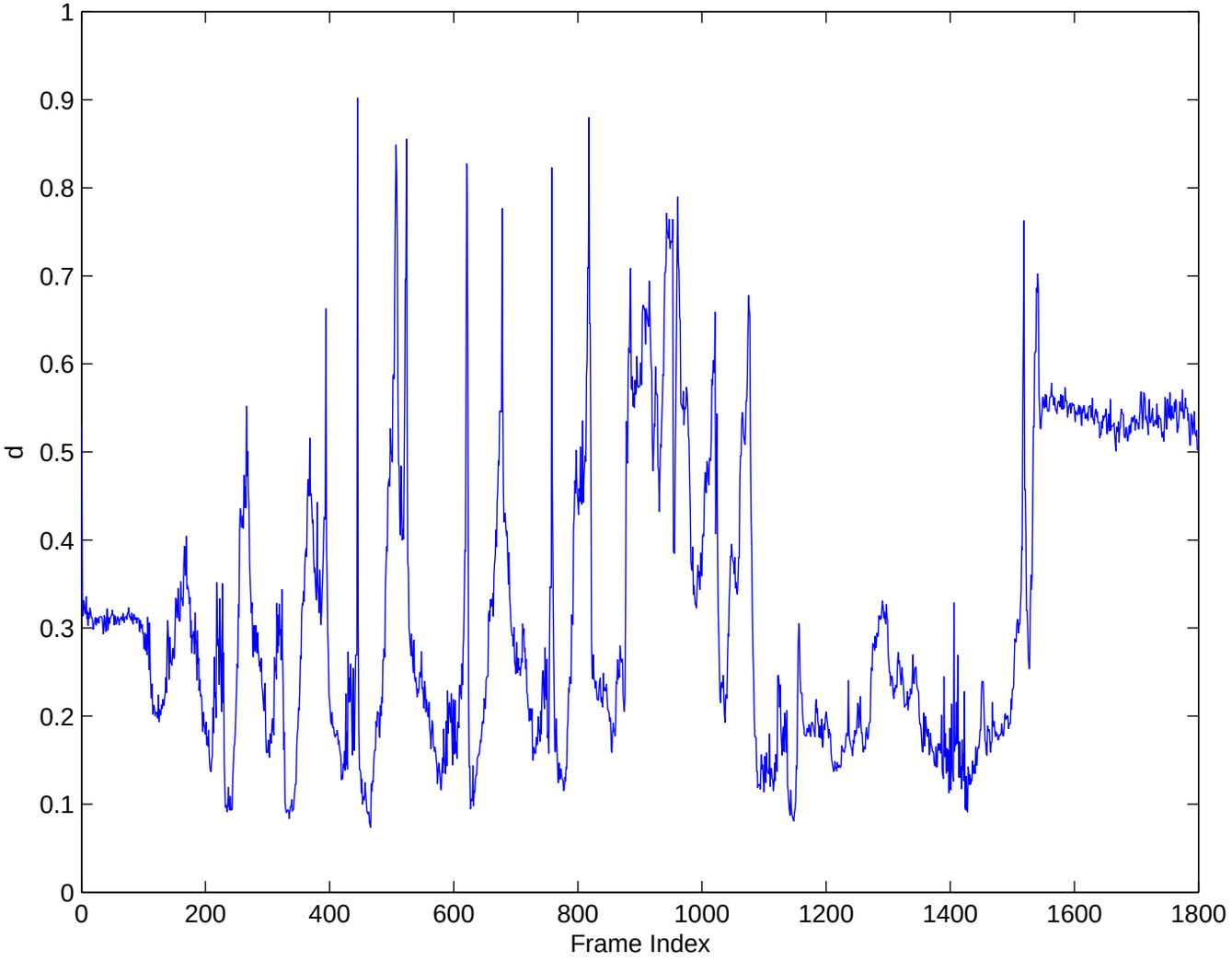
Data are weighted only within the mean shift window.

Illustration – Density in 2D

Mean shift is used to find modes.

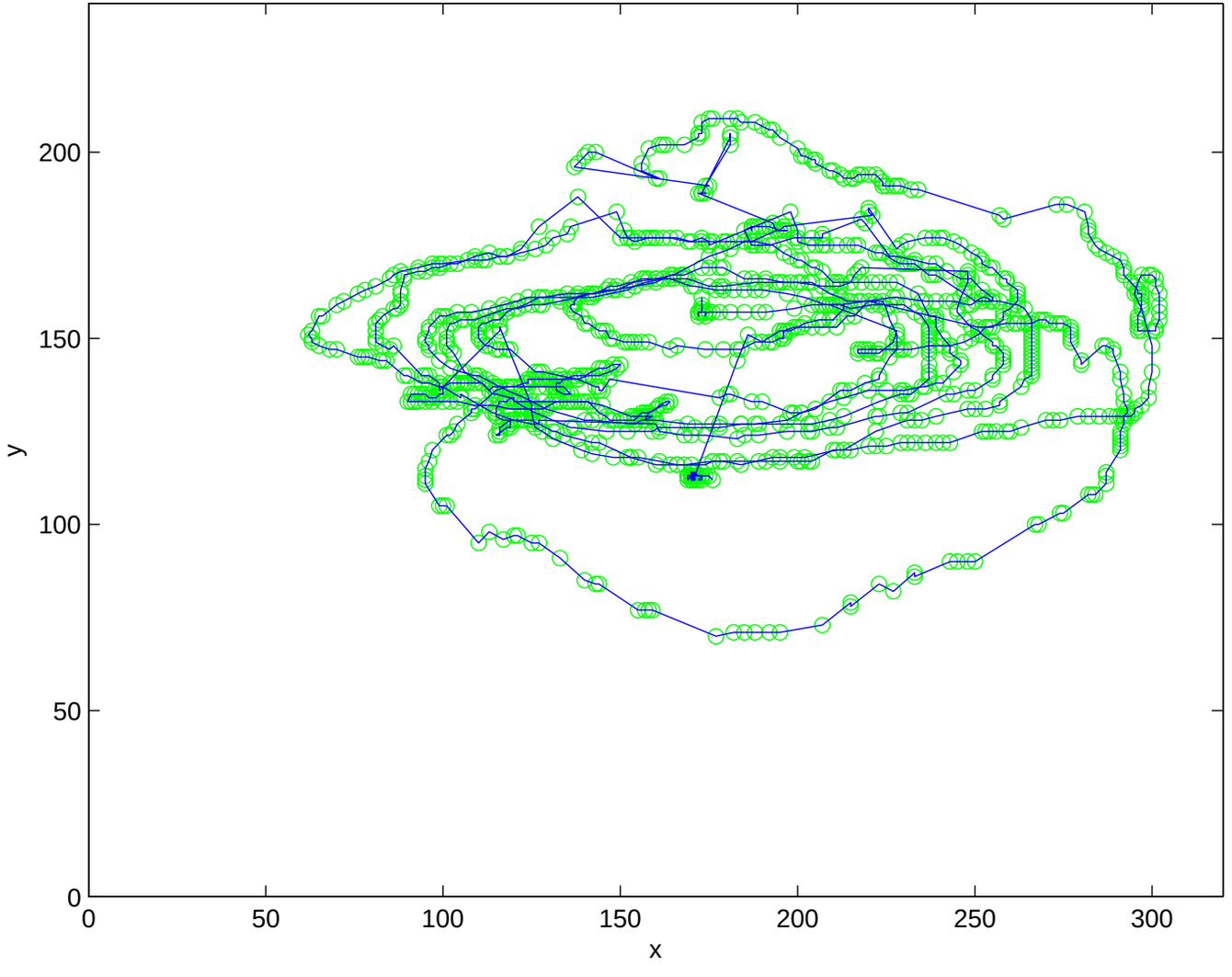
Illustration – Density in 3D

Illustration – Distance



Distance $d(\mathbf{y})$.

Illustration – Trajectory



Tracker locations in the video sequence.

Mean shift, Summary

- Tracking based on visual features (color, texture). Objects characterized by statistical distribution of these features.
- Pros - able to track variety of objects, capability to handle partial occlusions, significant clutter, rotation in depth, motion blur, target scale variations and changes in camera position.
- Cons - full occlusions cannot be handled, looses target when objects with similar probability distribution are present, unimodality.