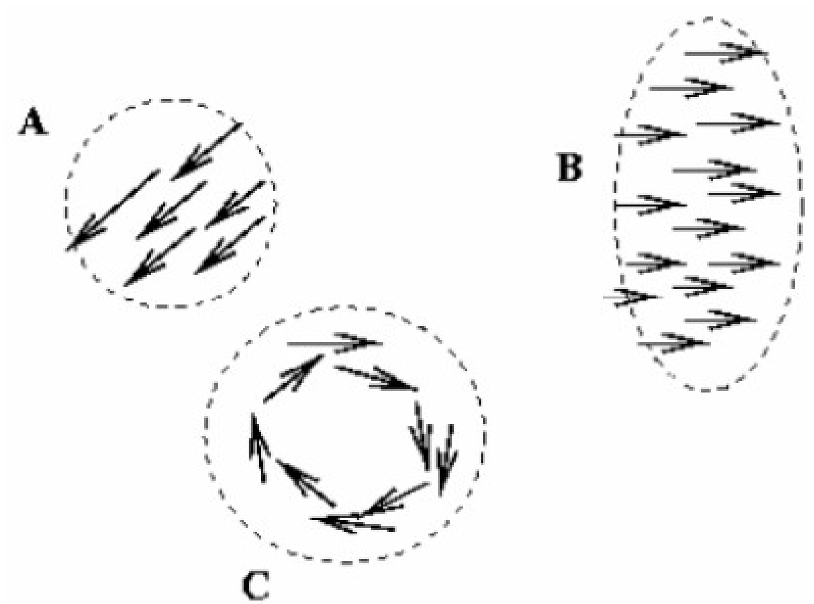
Motion and Tracking

Andrea Torsello
DAIS
Università Ca' Foscari
via Torino 155,
30172 Mestre (VE)

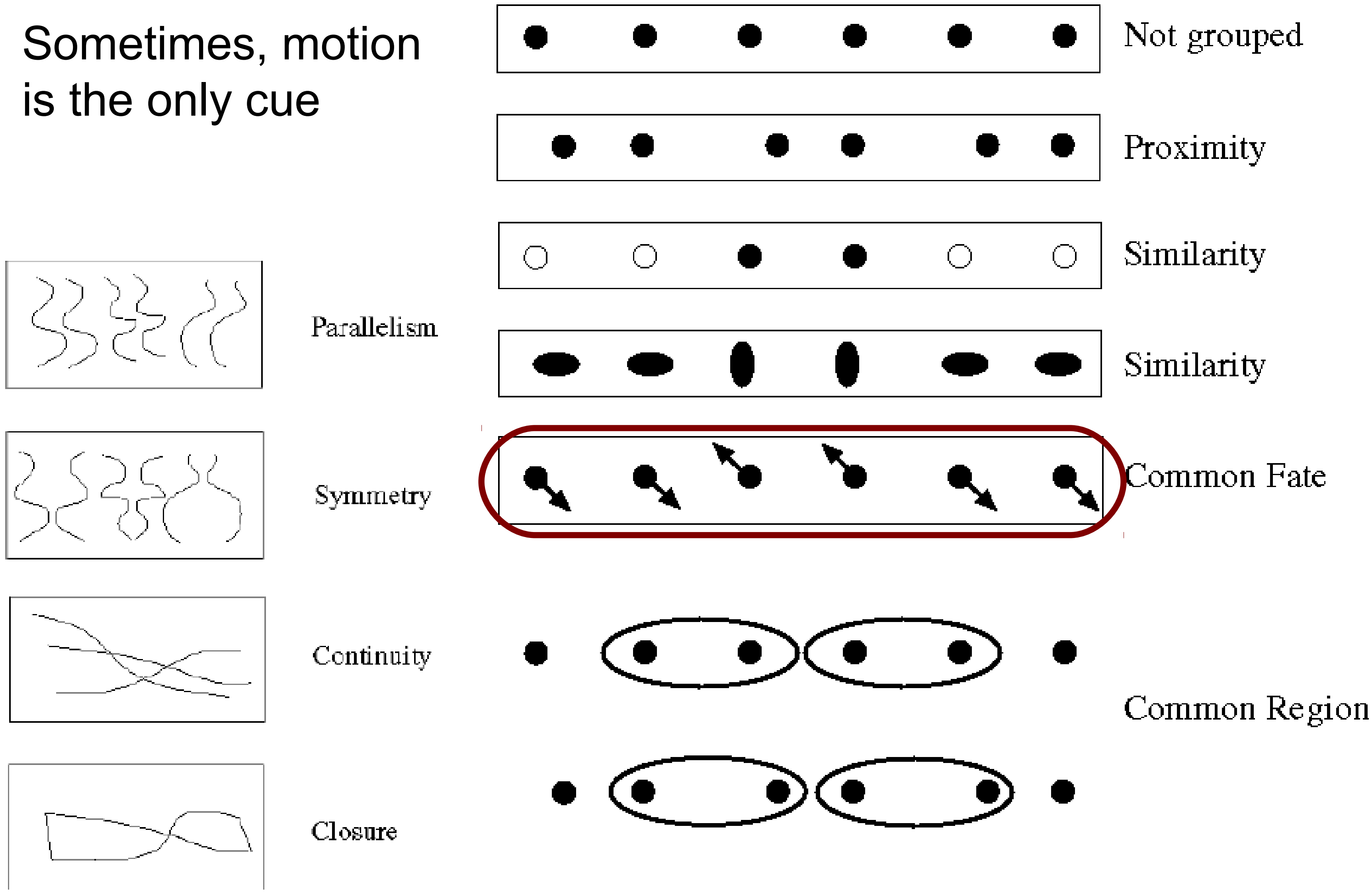
Motion Segmentation

- Segment the video into multiple coherently moving objects



Motion and Perceptual Organization

- Sometimes, motion is the only cue

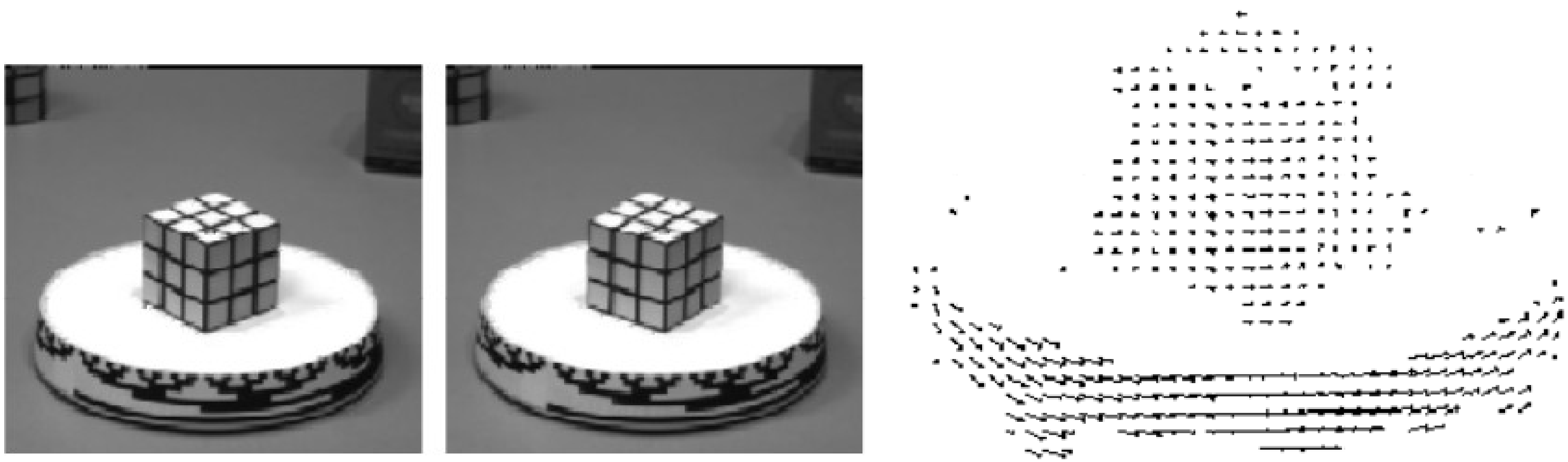


Motion and Perceptual Organization

- Sometimes, motion is the foremost cue

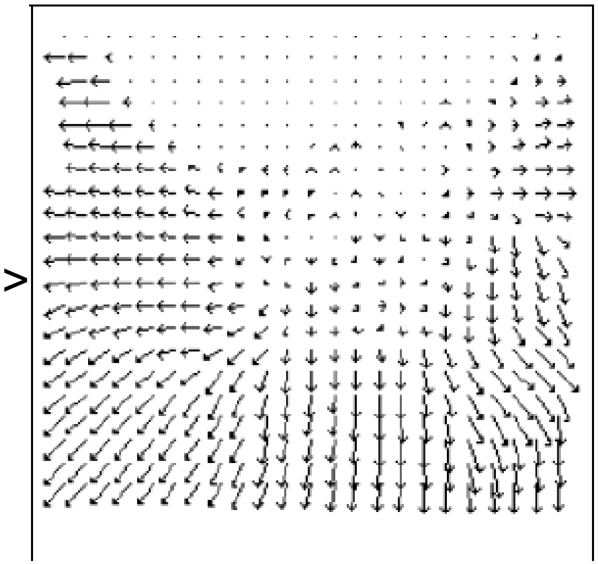
Motion Field

- The motion field is the projection of the 3D scene motion into the image



Motion field + camera motion

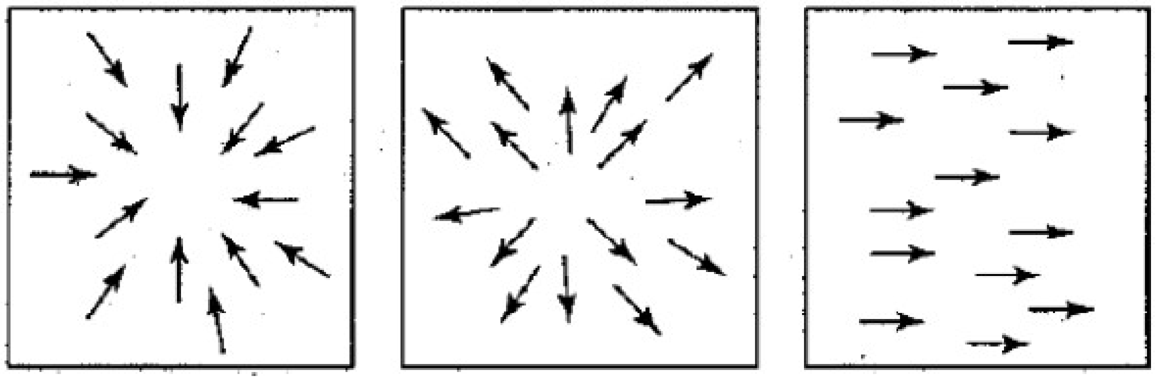
=>



Length of flow vectors inversely proportional to depth Z of 3d point

points closer to the camera move more quickly across the image plane

Motion field + camera motion



Zoom out Zoom in Pan right to left

Motion Estimation Techniques

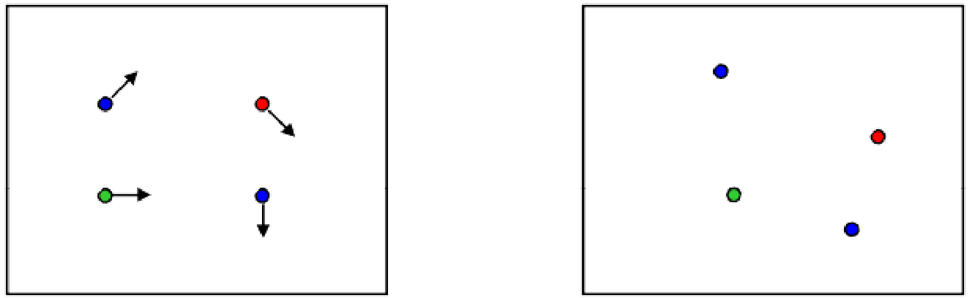
- Direct methods
  - Directly recover image motion at each pixel from spatio-temporal image brightness variations
  - Dense motion fields, fields but sensitive to appearance variations
  - Suitable for video and when image motion is small

- Feature-based methods
  - Extract visual features (corners, textured areas) and track them over multiple frames
  - Sparse motion fields, but more robust tracking
  - Suitable when image motion is large (10s of pixels)

Optical Flow

- **Optical flow** is the apparent motion of brightness patterns in the image
- Ideally, optical flow would be the same as the motion field
- Have to be careful: apparent motion can be caused by lighting changes without any actual motion

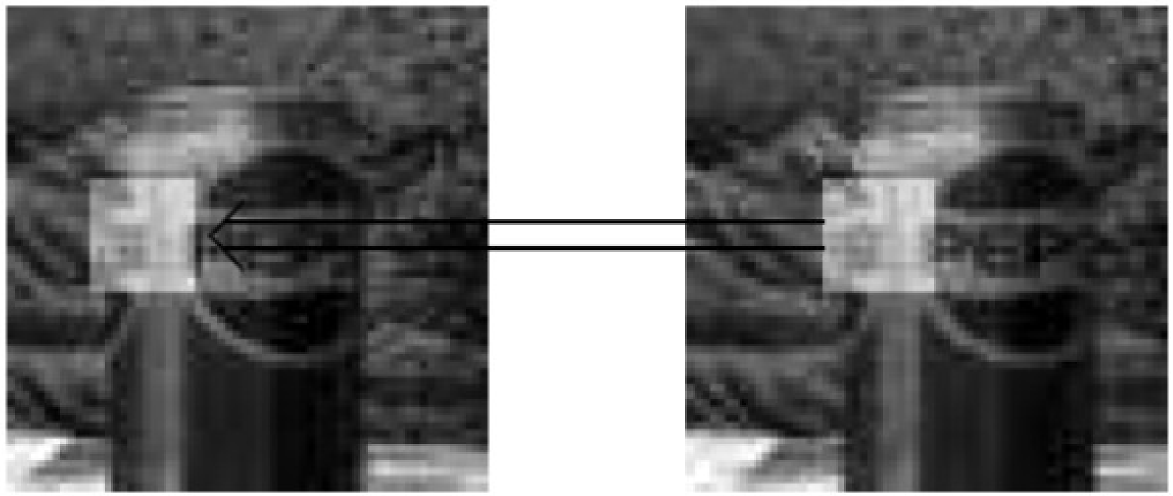
Optical Flow



$H(x, y)$ $I(x, y)$

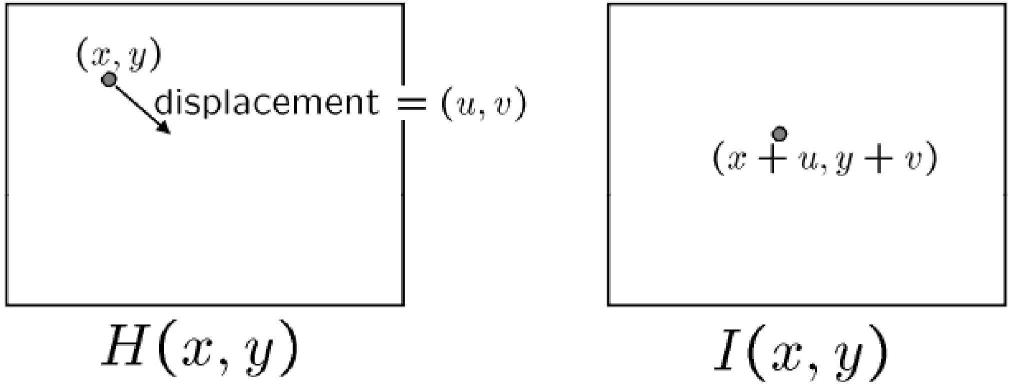
- How to estimate pixel motion from image H to image I?
  - Solve pixel correspondence problem: given a pixel in H, H look for nearby pixels of the same color in I
- Key assumptions
  - color constancy: a point in H looks the same in I
    - For grayscale images, this is brightness constancy
  - small motion: points do not move very far
- This is called the optical flow problem

Brightness Constancy



The highlighted region in the right image looks roughly the same as the region in the left image

Optical Flow Constraints



- Brightness constancy:

  H(x,y)=I(x+u,y+v)

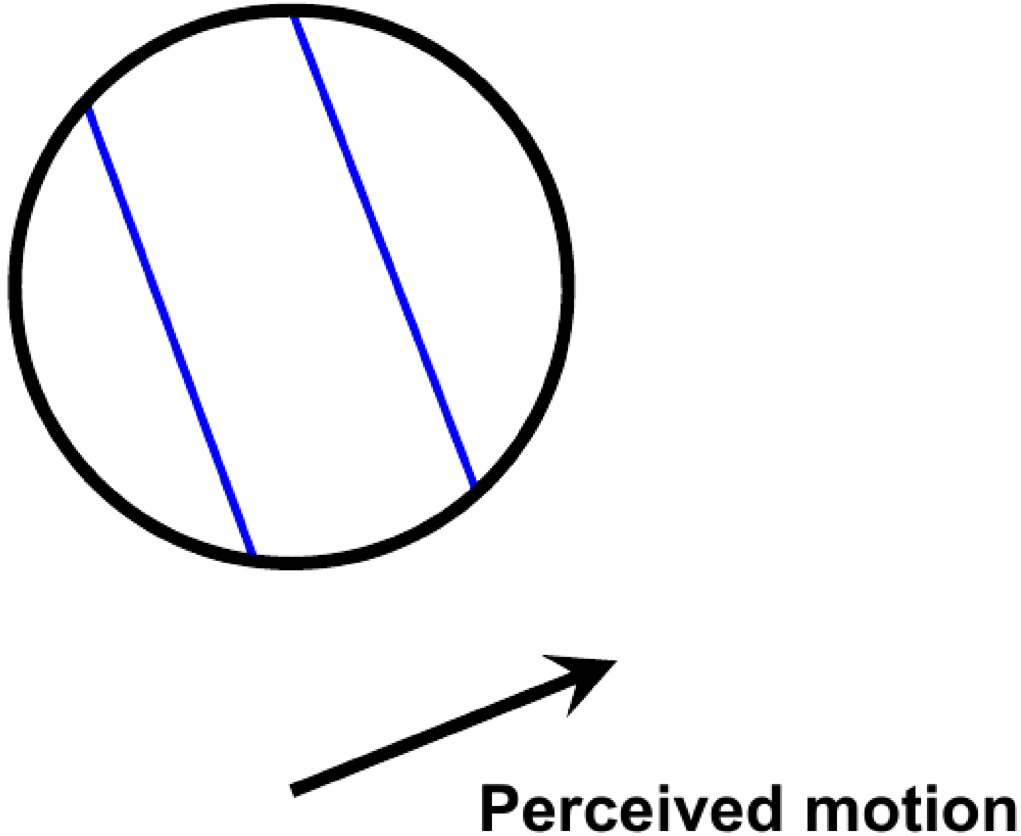
- Small motion:

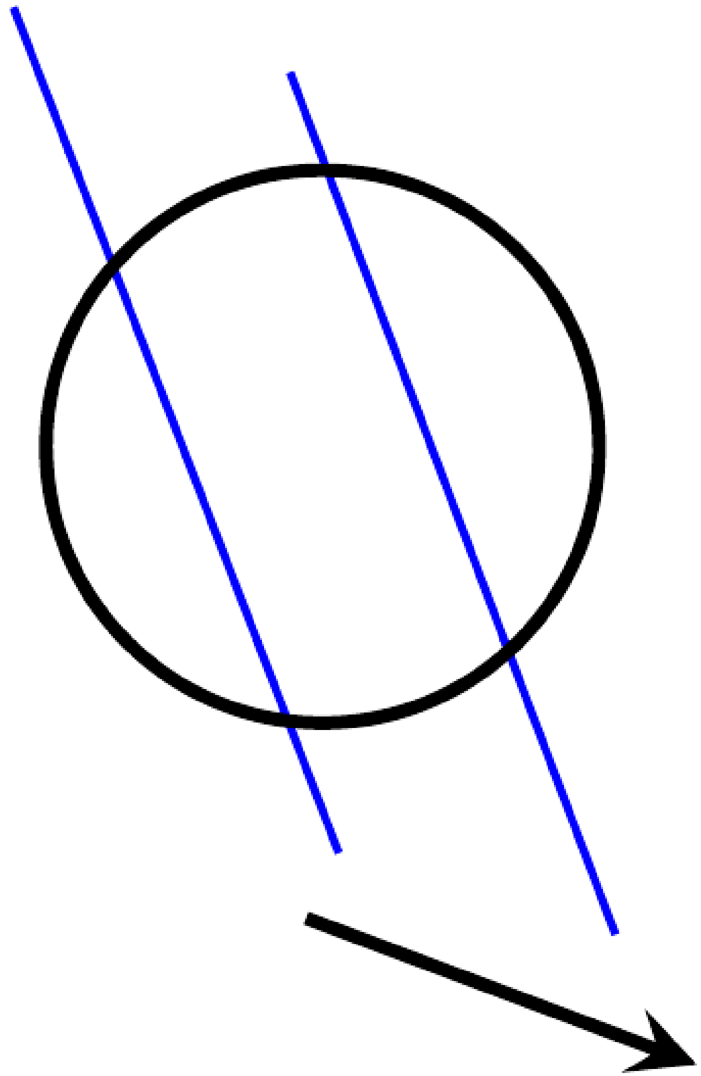
$$I(x+u, y+v) \approx I(x, y) + \frac{\partial I}{\partial x} u + \frac{\partial I}{\partial y} v$$

- Combining these equations

$$\begin{aligned} 0 &= I(x+u, y+v) - H(x, y) \\ &\approx (I(x, y) - H(x, y)) + I_x u + I_y v \\ &\approx I_t + \nabla I (u, v)^T \end{aligned}$$

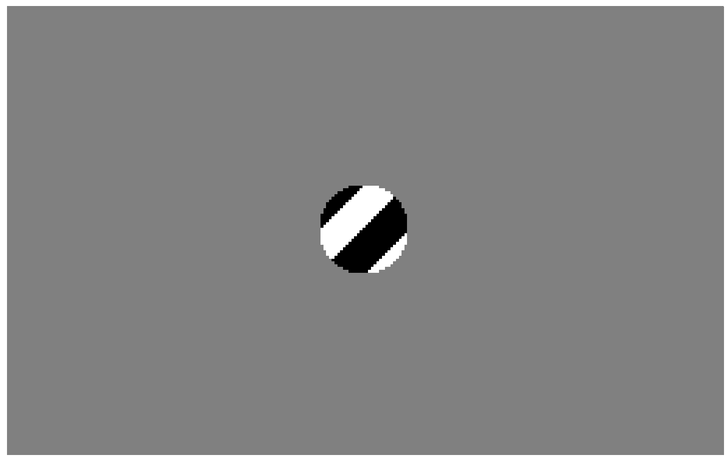
The Aperture Problem



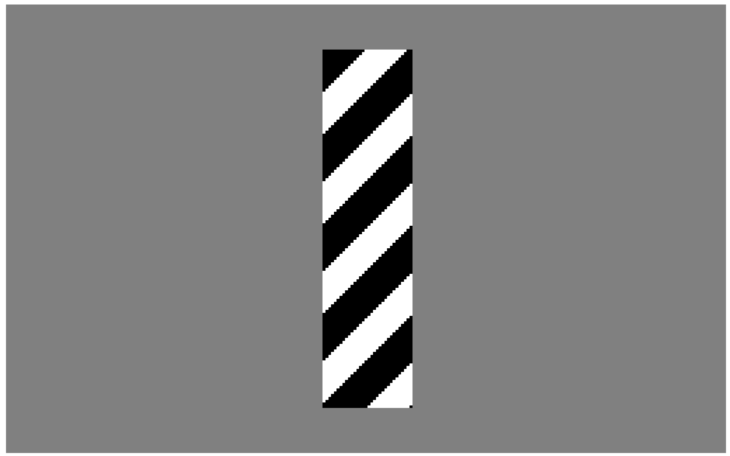


Actual motion

The barber pole illusion



The barber pole illusion



Solving the Aperture Problem

- How to get more equations for a pixel?
- Spatial coherence constraint: pretend the pixel's neighbors have the same (u,v)
  - If we use a 5x5 window, that gives us 25 equations per pixel

$$\begin{bmatrix} I_x(\mathbf{p}_1) & I_y(\mathbf{p}_1) \\ I_x(\mathbf{p}_2) & I_y(\mathbf{p}_2) \\ \vdots & \vdots \\ I_x(\mathbf{p}_{25}) & I_y(\mathbf{p}_{25}) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_t(\mathbf{p}_1) \\ I_t(\mathbf{p}_2) \\ \vdots \\ I_t(\mathbf{p}_{25}) \end{bmatrix}$$

Solving the Aperture Problem

- Prob: we have more equations than unknowns
- Solution: solve least squares problem

$$\underset{25\text{x}2}{A}\ \underset{2\text{x}1}{d} = \underset{25\text{x}1}{b} \longrightarrow \text{minimize } ||Ad - b||^2$$

  - Solved by $\underset{2\text{x}2}{(A^T A)}\ \underset{2\text{x}1}{d} = \underset{2\text{x}1}{A^T b}$

$$\underbrace{\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix}}_{A^T A} \begin{bmatrix} u \\ v \end{bmatrix} = - \underbrace{\begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}}_{A^T b}$$

UNIVERSITÀ CA' FOSCARI VENEZIA

$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}$$

$$A^T A \qquad\qquad\qquad A^T b$$

- When is this solvable?
  - $A^T A$ should be invertible
  - $A^T A$ should not be too small
    - eigenvalues λ1 and λ2 of A T A should not be too small
  - $A^T A$ should be well-conditioned
    - λ1/ λ2 should not be too large (λ1 = larger eigenvalue)

Edge

UNIVERSITÀ CA' FOSCARI VENEZIA

- gradients very large or very small
- large λ1, small λ2

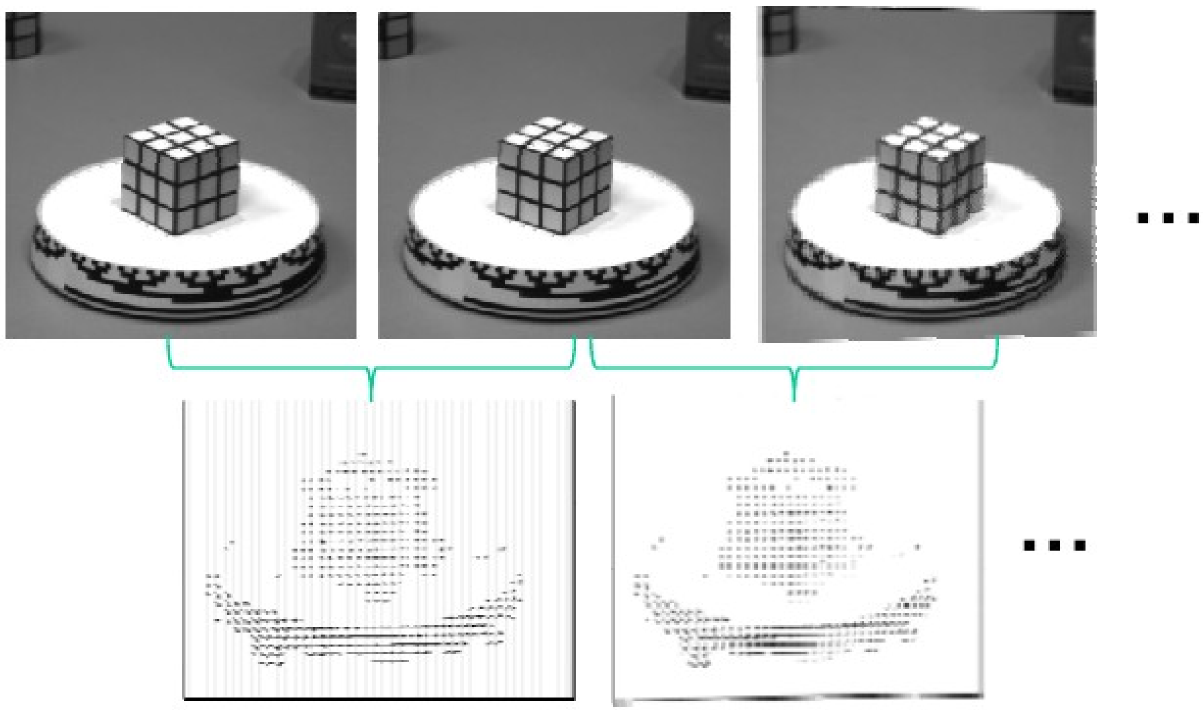
Low-texture region

- gradients have small magnitude
- small $\lambda 1$, small $\lambda 2$

- gradients are different, large magnitudes
- large $\lambda1$, large $\lambda2$

Optical flow for tracking

- If we have more than just a pair of frames, we could compute flow from one to the next:



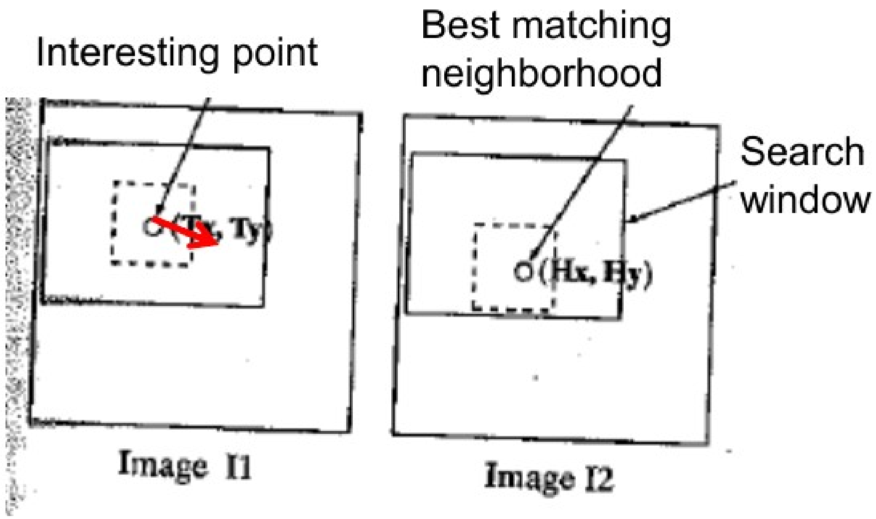
- But flow only reliable for small motions, and we may have occlusions, textureless regions that yield bad estimates anyway...

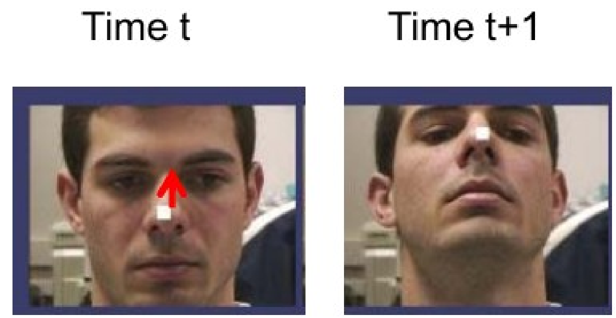
Motion Estimation Techniques

- Direct methods
  - Directly recover image motion at each pixel from spatio-temporal image brightness variations
  - Dense motion fields, fields but sensitive to appearance variations
  - Suitable for video and when image motion is small

- Feature-based methods
  - Extract visual features (corners, textured areas) and track them over multiple frames
  - Sparse motion fields, but more robust tracking
  - Suitable when image motion is large (10s of pixels)

Feature-based matching for motion



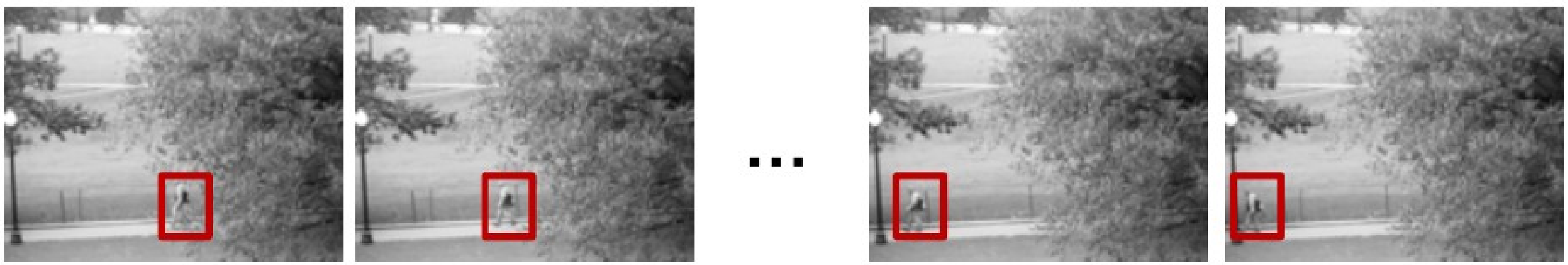


- Search window is centered at the point where we last saw the feature, in image I1
- Best match = position where we have the highest normalized cross-correlation value
- Where should the search window be placed?
  - Near match at previous frame
  - More generally, taking into account the expected dynamics of the object

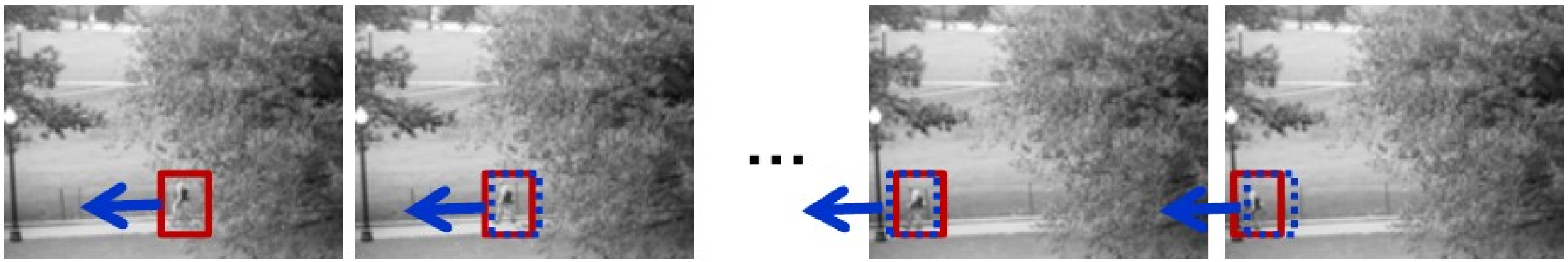
Detection vs. tracking

t=1 t=2 … t=20 t=21

Detection vs. tracking



Detection: We detect the object independently in each frame and can record its position over time, e.g., based on blob's centroid or detection window coordinates



Tracking with dynamics: We use image measurements to estimate position of object, but also incorporate position predicted by dynamics, i.e., our expectation of object's motion pattern.

Tracking with dynamics

- Use model of expected motion to predict where objects will occur in next frame, even before seeing the image.

- **Intent**:
  - Do less work looking for the object, restrict the search.
  - Get improved estimates since measurement noise is tempered by smoothness, dynamics priors.

- **Assumption**: continuous motion patterns:
  - Camera is not moving instantly to new viewpoint
  - Objects do not disappear and reappear in different places in the scene
  - Gradual change in pose between camera and scene

Tracking as inference

- The hidden state consists of the true parameters we care about, denoted X.
- The measurement is our noisy observation that results from the underlying state, denoted Y.
- At each time step, state changes (from X_{t-1} to X_t) and we get a new observation Y_t
- Our goal: recover most likely state X_t given
  - All observations seen so far.
  - Knowledge about dynamics of state transitions.

Independence Assumptions

- Only immediate past state influences current state

$$P(\boldsymbol{X}_i|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{i-1}) = P(\boldsymbol{X}_i|\boldsymbol{X}_{i-1})$$

- Measurements at time i only depend on the current state

$$P(\boldsymbol{Y}_i,\boldsymbol{Y}_j,\ldots\boldsymbol{Y}_k|\boldsymbol{X}_i) = P(\boldsymbol{Y}_i|\boldsymbol{X}_i)P(\boldsymbol{Y}_j,\ldots,\boldsymbol{Y}_k|\boldsymbol{X}_i)$$

Tracking via deformable contours

- Use final contour/model extracted at frame t as an initial solution for frame t+1
- Evolve initial contour to fit exact object boundary at frame t+1
- Repeat, initializing with most recent frame.