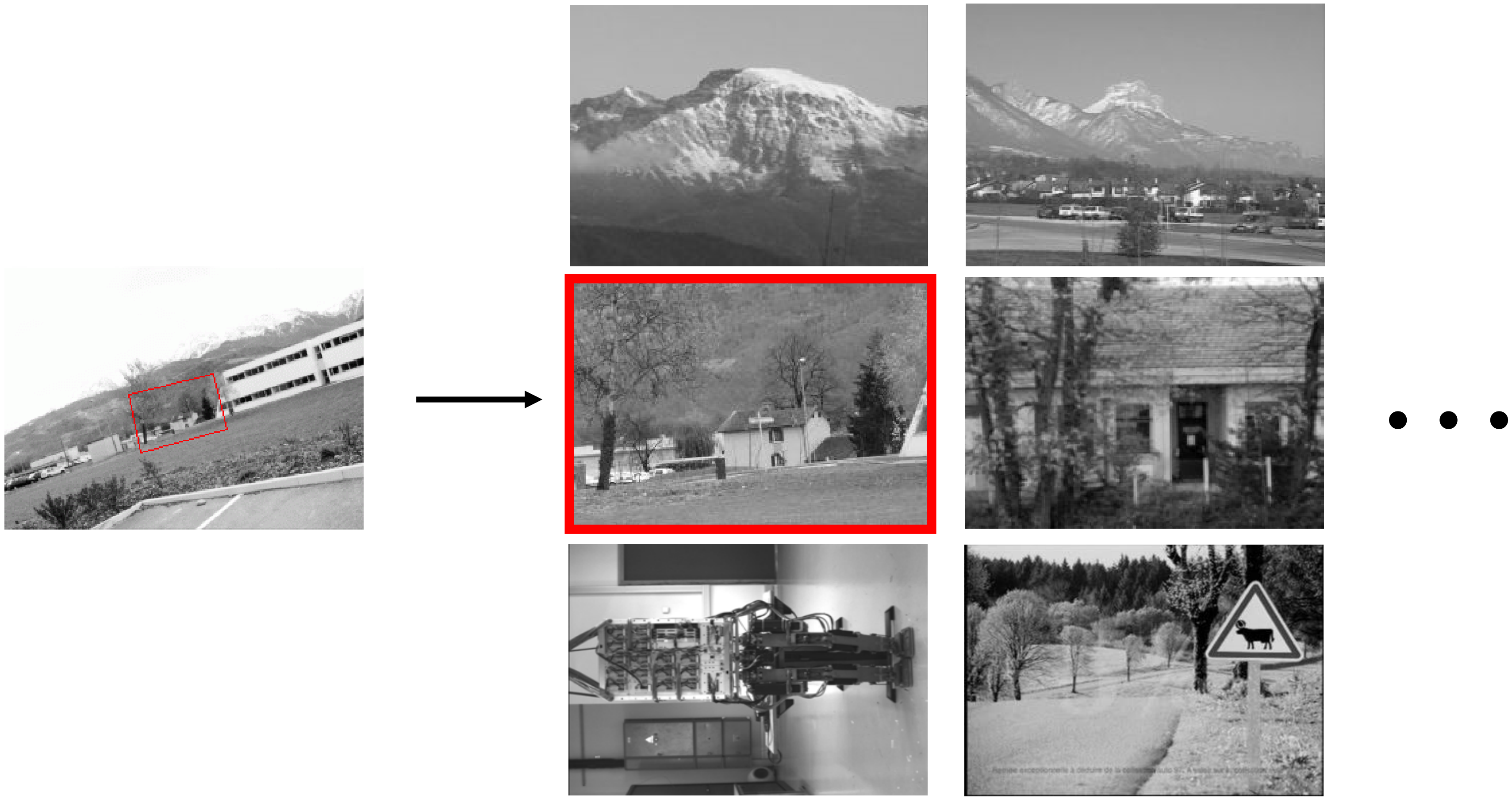
Instance-level recognition

Cordelia Schmid
INRIA, Grenoble

Instance-level recognition

Search for particular objects and scenes in large databases

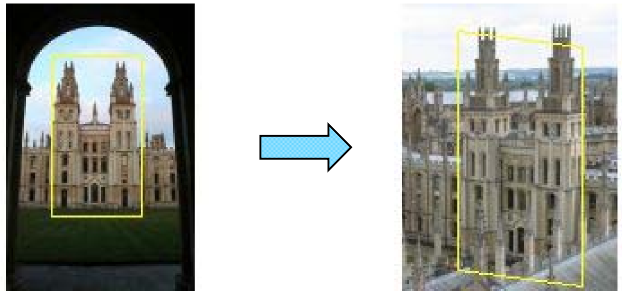


Difficulties

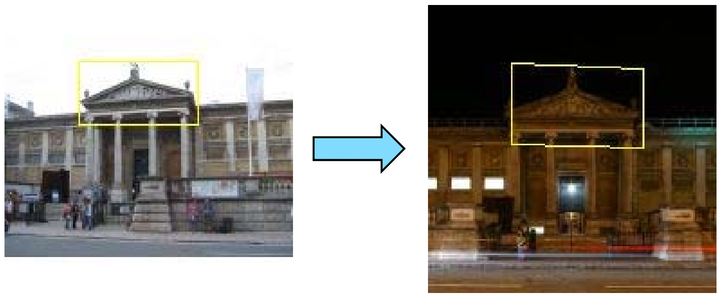
Finding the object despite possibly large changes in scale, viewpoint, lighting and partial occlusion

→ **requires invariant description**

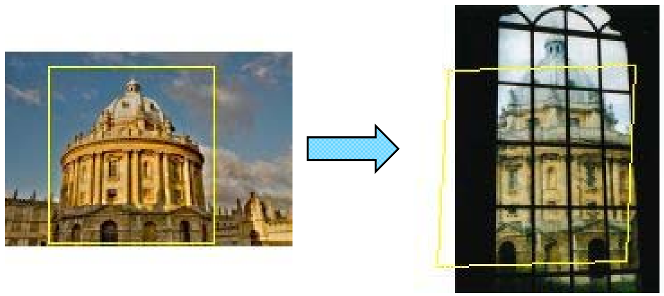
Scale



Viewpoint



Lighting



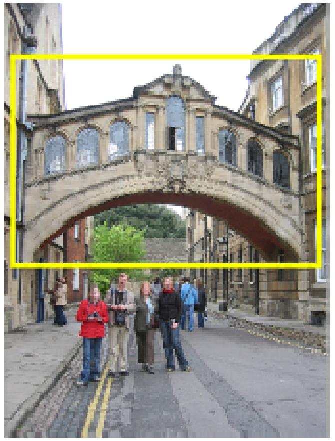
Occlusion

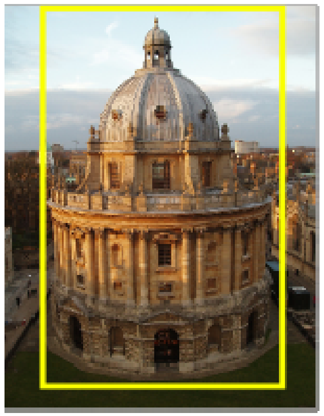
Difficulties

- Very large images collection → need for efficient indexing
  - Flickr has 2 billion photographs, more than 1 million added daily
  - Facebook has 15 billion images (~27 million added daily)
  - Large personal collections
  - Video collections, i.e., YouTube

Applications

Search photos on the web for particular places





Find these landmarks

...in these images and 1M more

Applications

- Take a picture of a product or advertisement
  → find relevant information on the web

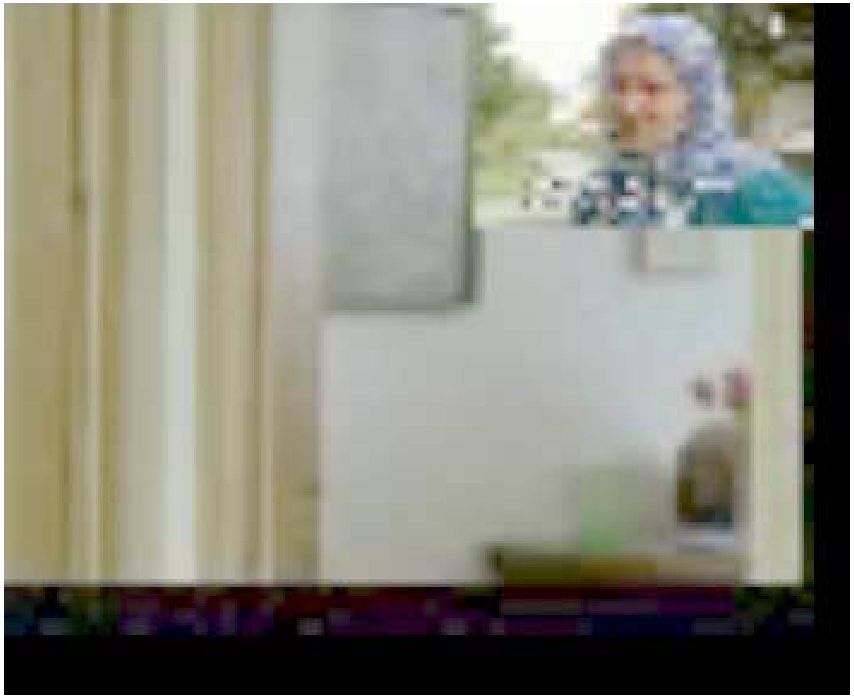
Applications

- Finding stolen/missing objects in a large collection

Applications

- Copy detection for images and videos

Query video



Search in 200h of video

Applications

- Sony Aibo – Robotics
  - Recognize docking station
  - Communicate with visual cards
  - Place recognition
  - Loop closure in SLAM

Instance-level recognition

1) Local invariant features

2) Matching and recognition with local features

3) Efficient visual search

4) Very large scale indexing

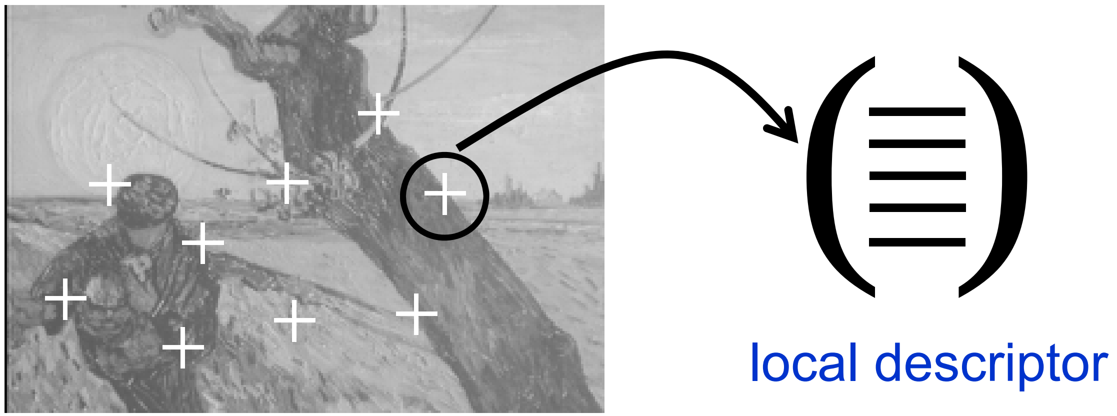
Local invariant features

- **Introduction to local features**

- Harris interest points + SSD, ZNCC, SIFT

- Scale & affine invariant interest point detectors

Local features



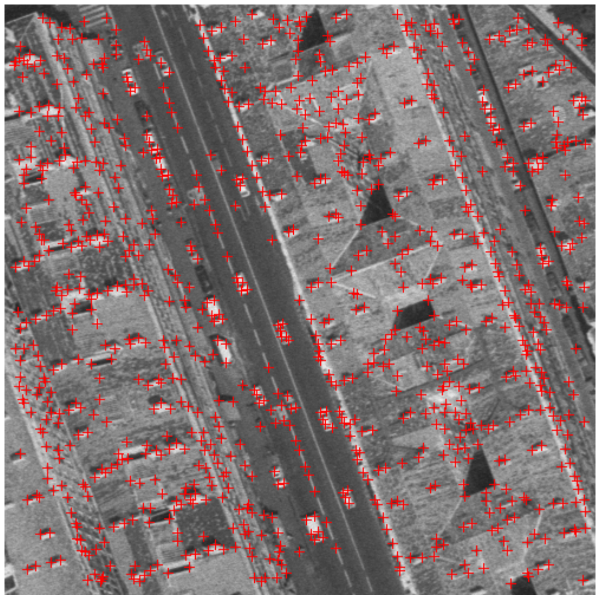
Many local descriptors per image

Robust to occlusion/clutter + no object segmentation required

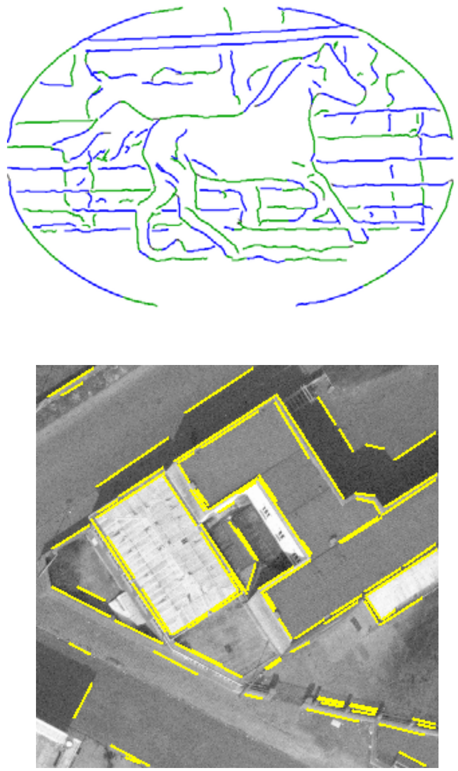
Photometric : distinctive

Invariant : to image transformations + illumination changes

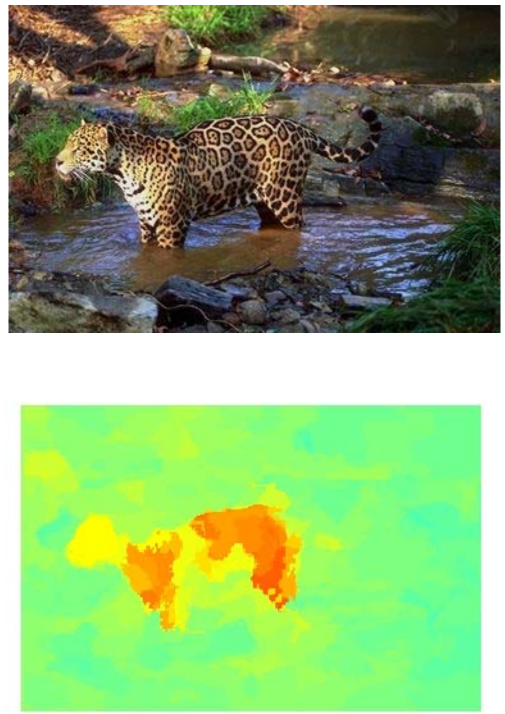
Local features



Interest Points

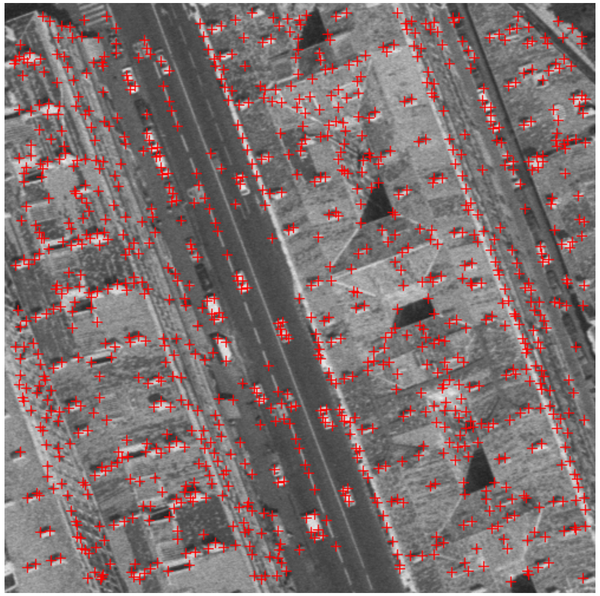


Contours/lines



Region segments

Local features

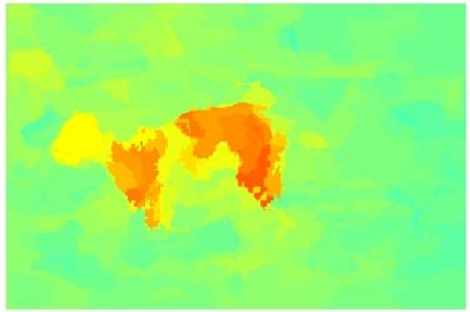


Interest Points

Patch descriptors, i.e. SIFT

Contours/lines

Mi-points, angles



Region segments

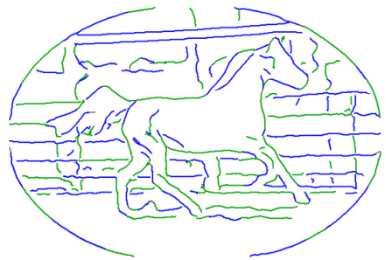
Color/texture histogram

Interest points / invariant regions

Harris detector

Scale/affine inv. detector

Contours / lines

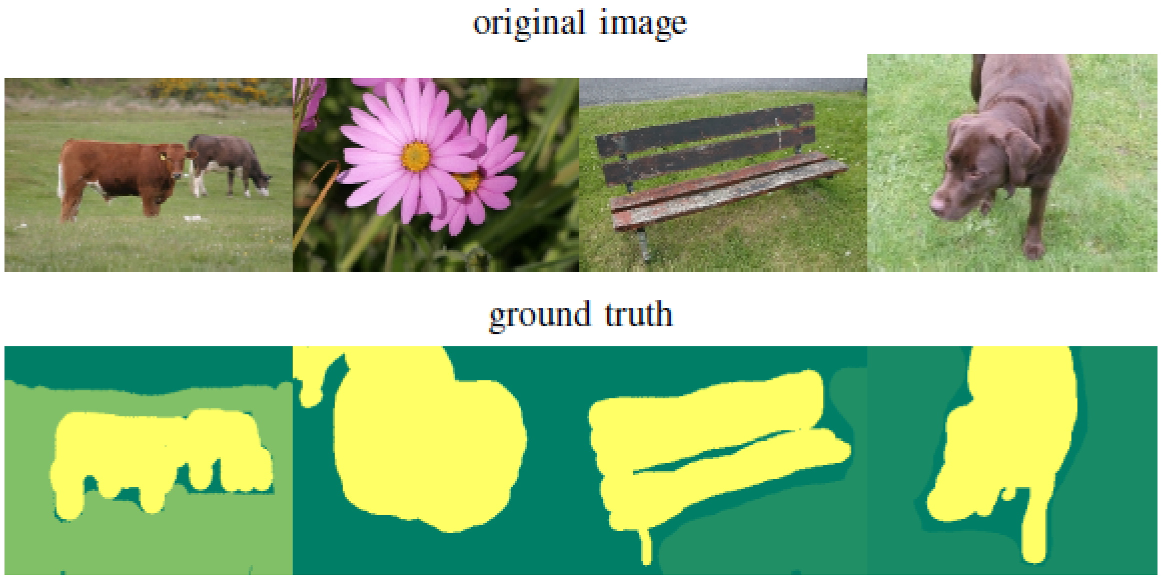


- Extraction de contours
  - Zero crossing of Laplacian
  - Local maxima of gradients

- Chain contour points (hysteresis) , Canny detector

- Recent contour detectors
  - global probability of boundary (**gPb**) detector [Malik et al., UC Berkeley, CVPR'08]
  - Structured forests for fast edge detection **(SED)** [Dollar and Zitnick, ICCV'13]

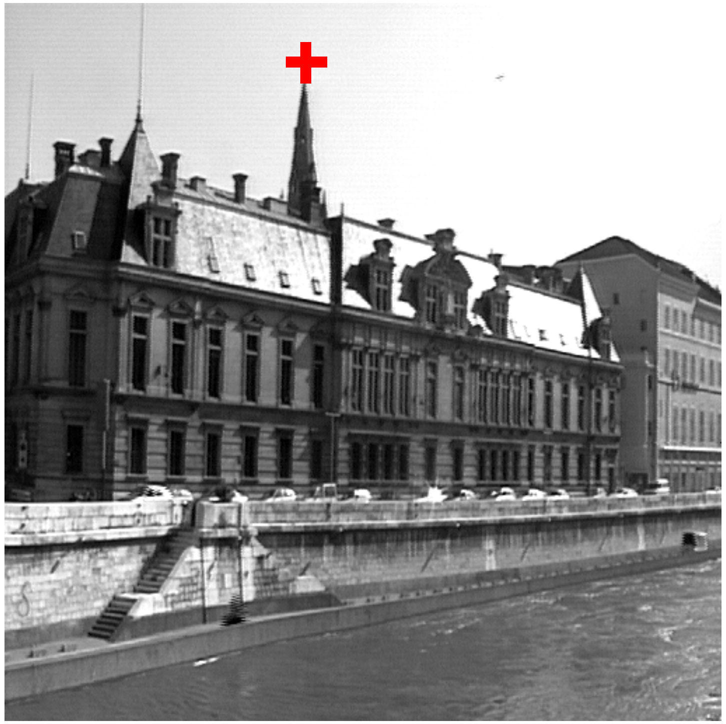
Regions segments / superpixels



Simple linear iterative clustering (SLIC)

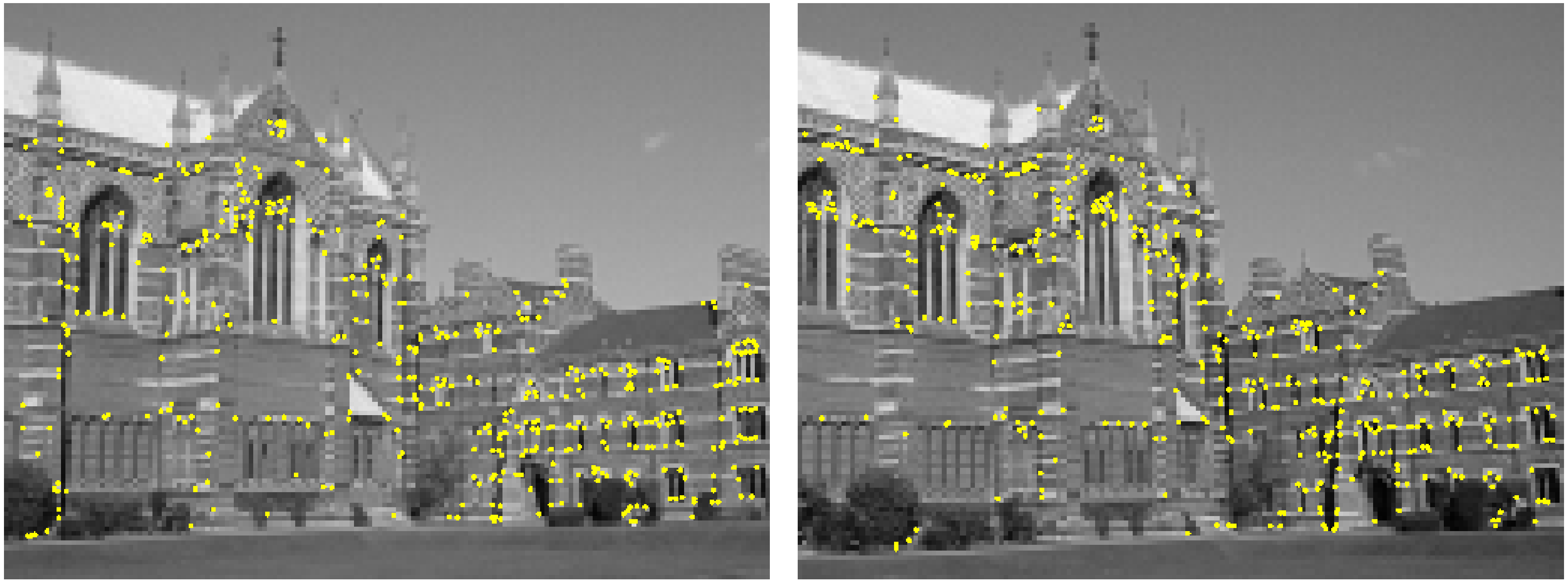
Normalized cut [Shi & Malik], Mean Shift [Comaniciu & Meer], ….

Matching of local descriptors



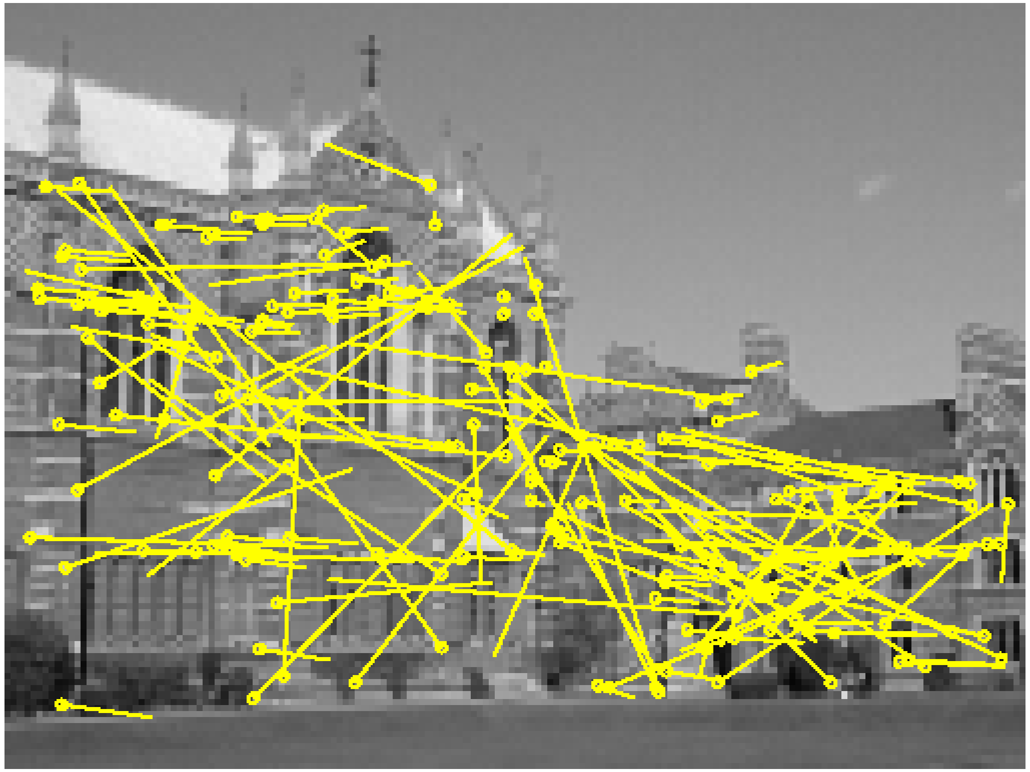
Find corresponding locations in the image

Illustration – Matching



Interest points extracted with Harris detector (~ 500 points)

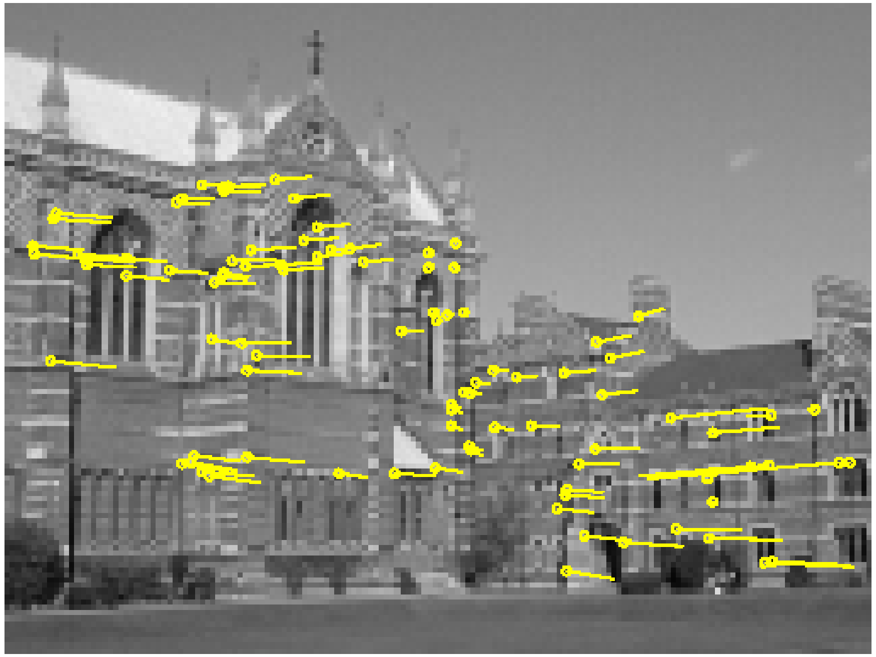
Illustration – Matching



Interest points matched based on cross-correlation (188 pairs)

Illustration – Matching

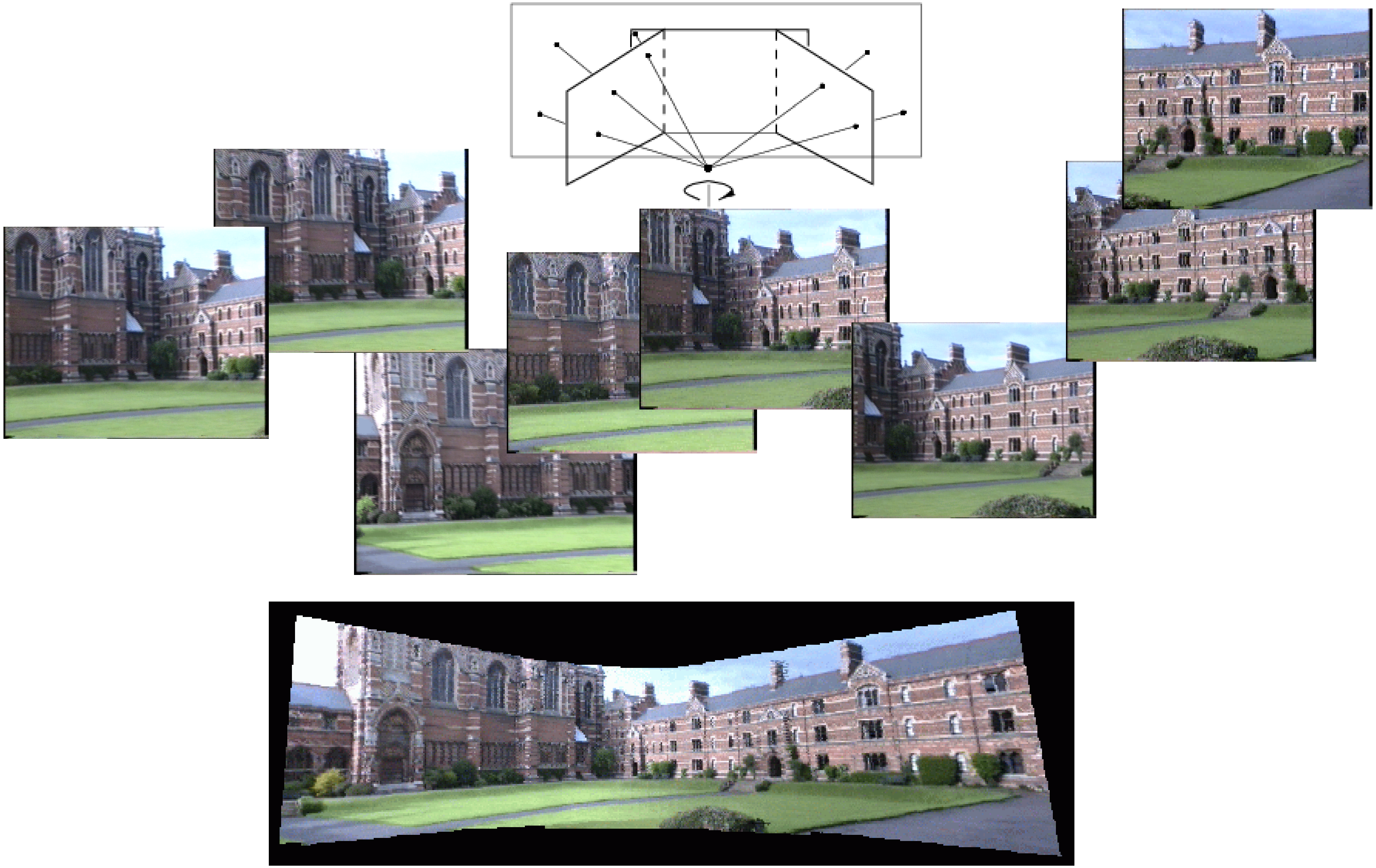
Global constraint - Robust estimation of the fundamental matrix



99 inliers

89 outliers

Application: Panorama stitching



Overview

- Introduction to local features
- **Harris interest points + SSD, ZNCC, SIFT**
- Scale & affine invariant interest point detectors

Harris detector [Harris & Stephens'88]

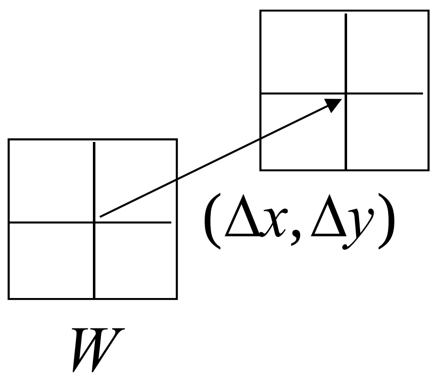
Based on the idea of auto-correlation

Important difference in all directions => interest point

Harris detector

Auto-correlation function for a point (x, y) and a shift $(\Delta x, \Delta y)$

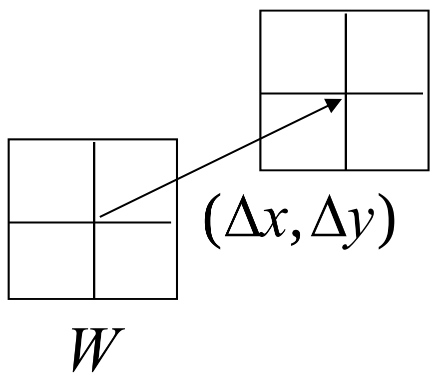
$$A(x, y) = \sum_{(x_k, y_k) \in W(x,y)} (I(x_k, y_k) - I(x_k + \Delta x, y_k + \Delta y))^2$$



Harris detector

Auto-correlation function for a point (x, y) and a shift $(\Delta x, \Delta y)$

$$A(x, y) = \sum_{(x_k, y_k) \in W(x,y)} \left(I(x_k, y_k) - I(x_k + \Delta x, y_k + \Delta y) \right)^2$$



$$A(x, y) \begin{cases} \text{small in all directions} & \rightarrow \text{uniform region} \\ \text{large in one directions} & \rightarrow \text{contour} \\ \text{large in all directions} & \rightarrow \text{interest point} \end{cases}$$

Harris detector

Discret shifts are avoided based on the auto-correlation matrix

with first order approximation

$$I(x_k + \Delta x, y_k + \Delta y) = I(x_k, y_k) + \begin{pmatrix} I_x(x_k, y_k) & I_y(x_k, y_k) \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

$$A(x, y) = \sum_{(x_k, y_k) \in W(x,y)} (I(x_k, y_k) - I(x_k + \Delta x, y_k + \Delta y))^2$$

$$= \sum_{(x_k, y_k) \in W} \left(\begin{pmatrix} I_x(x_k, y_k) & I_y(x_k, y_k) \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \right)^2$$

Harris detector

$$= \begin{pmatrix} \Delta x & \Delta y \end{pmatrix} \begin{bmatrix} \sum_{(x_k, y_k) \in W} (I_x(x_k, y_k))^2 & \sum_{(x_k, y_k) \in W} I_x(x_k, y_k) I_y(x_k, y_k) \\ \sum_{(x_k, y_k) \in W} I_x(x_k, y_k) I_y(x_k, y_k) & \sum_{(x_k, y_k) \in W} (I_y(x_k, y_k))^2 \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

Auto-correlation matrix

the sum can be smoothed with a Gaussian

$$= \begin{pmatrix} \Delta x & \Delta y \end{pmatrix} G \otimes \begin{bmatrix} I_x^{\ 2} & I_x I_y \\ I_x I_y & I_y^{\ 2} \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

Harris detector

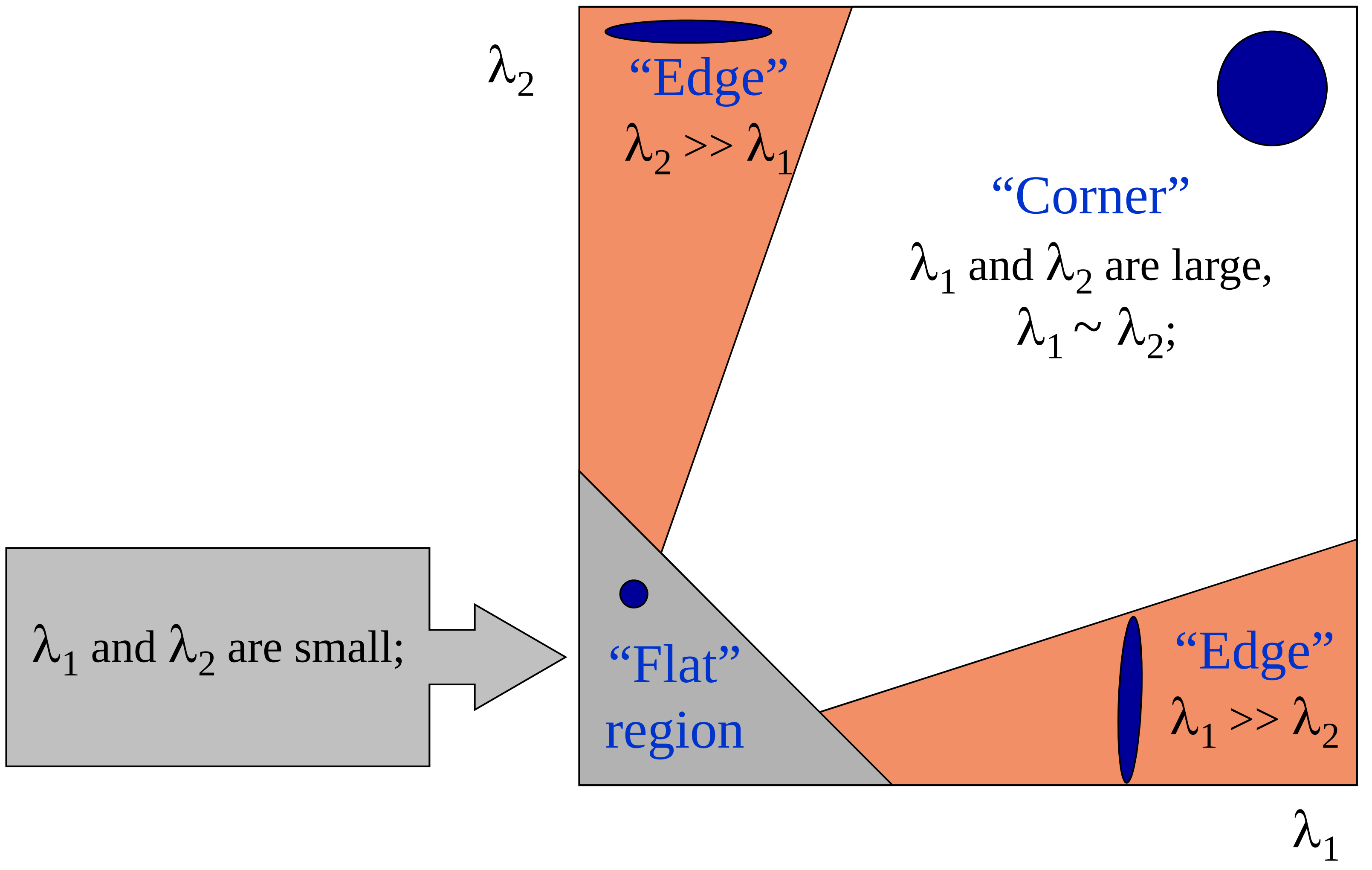
- Auto-correlation matrix

$$A(x, y) = G \otimes \begin{bmatrix} I_x^{\ 2} & I_x I_y \\ I_x I_y & I_y^{\ 2} \end{bmatrix}$$

  - captures the structure of the local neighborhood
  - measure based on eigenvalues of this matrix
    - 2 strong eigenvalues => interest point
    - 1 strong eigenvalue => contour
    - 0 eigenvalue => uniform region

Interpreting the eigenvalues

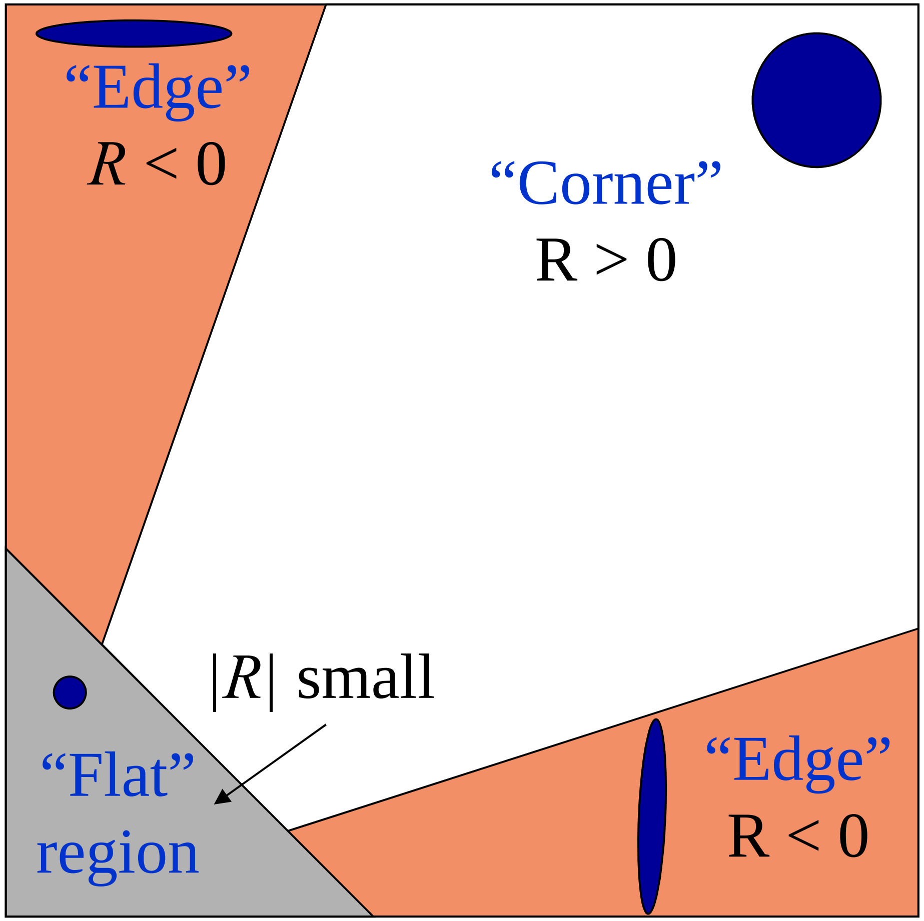
Classification of image points using eigenvalues of autocorrelation matrix



Corner response function

$$R = \det(A) - \alpha\, \mathrm{trace}(A)^2 = \lambda_1\lambda_2 - \alpha(\lambda_1 + \lambda_2)^2$$

α: constant (0.04 to 0.06)



Harris detector

- Cornerness function

$$R = \det(A) - k(trace(A))^2 = \lambda_1\lambda_2 - k(\lambda_1 + \lambda_2)^2$$

Reduces the effect of a strong contour

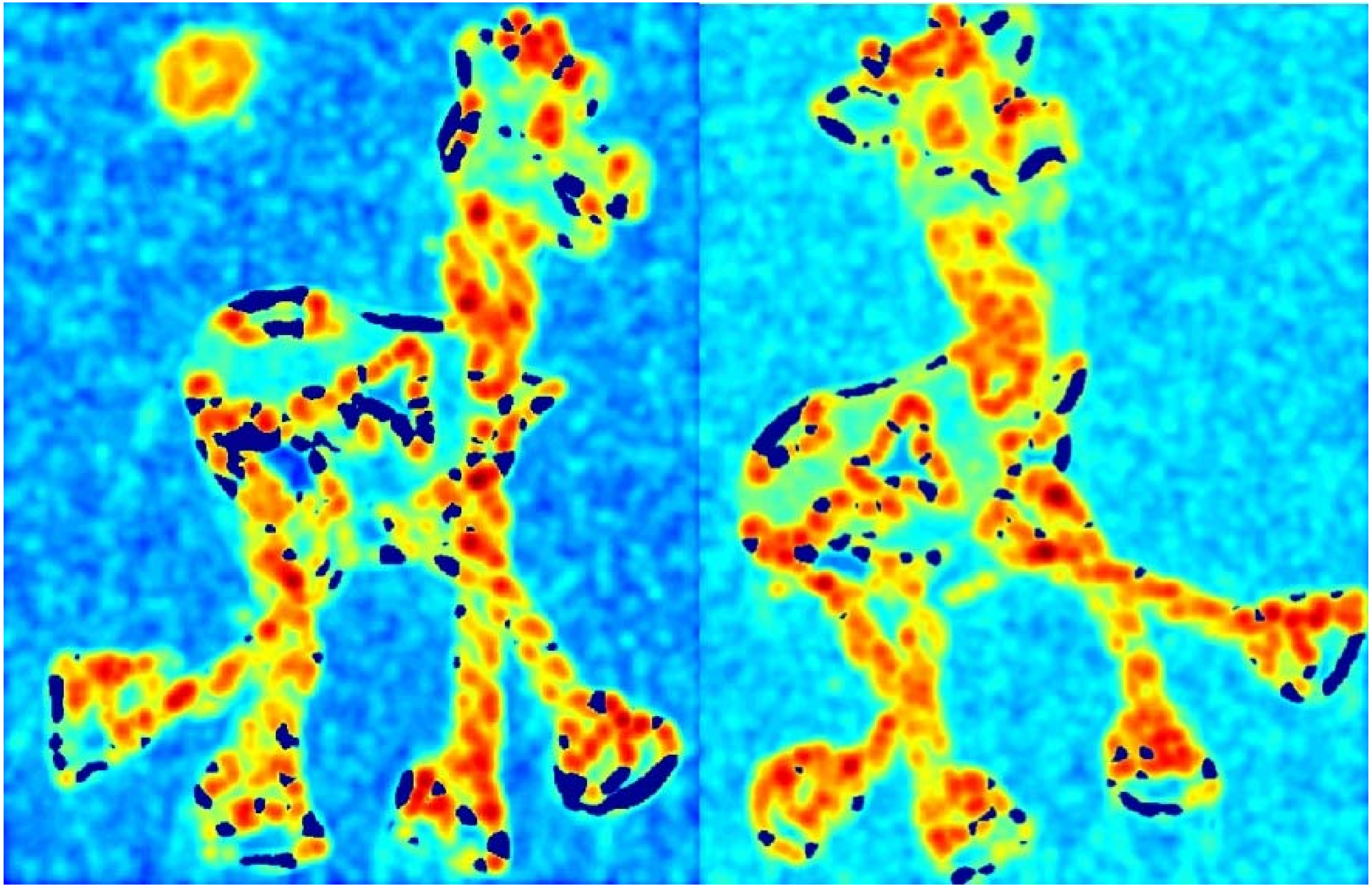
- Interest point detection
  - Treshold (absolut, relatif, number of corners)
  - Local maxima

$$f > thresh \;\wedge\; \forall x, y \in 8-neighbourhood \;\; f(x,y) \geq f(x',y')$$

Harris Detector: Steps

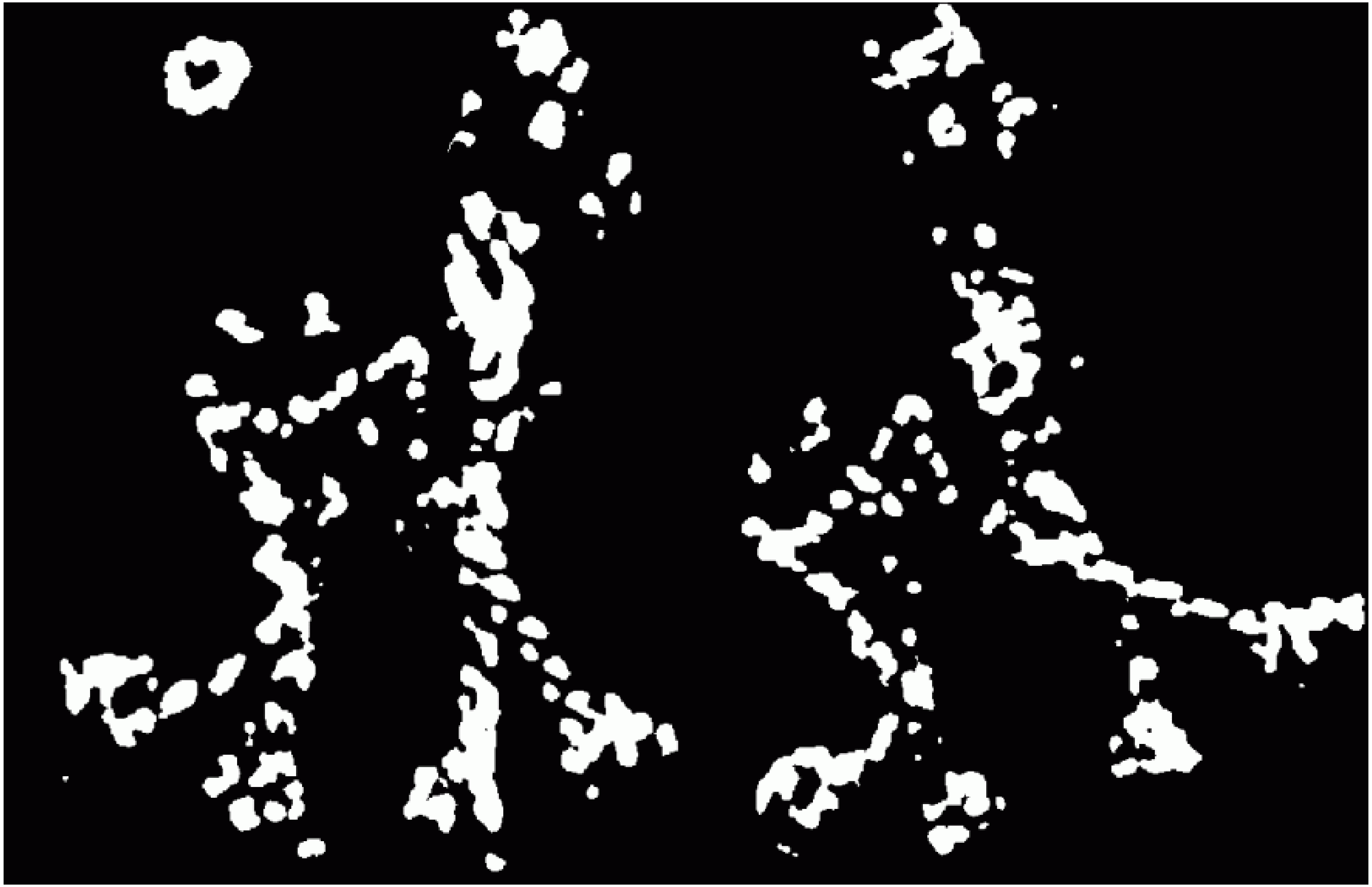
Harris Detector: Steps

Compute corner response R



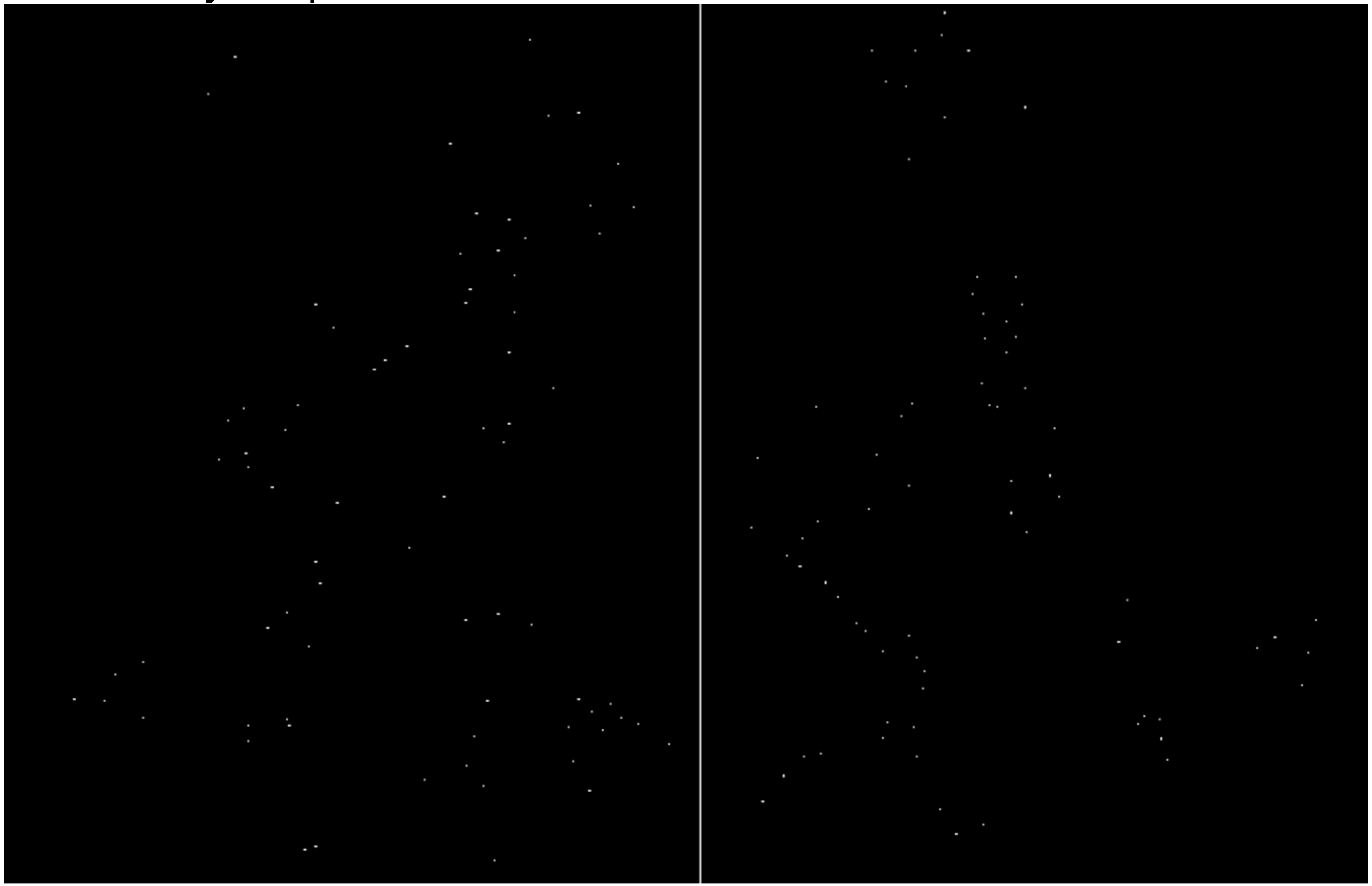
Harris Detector: Steps

Find points with large corner response: *R*>threshold



Harris Detector: Steps

Take only the points of local maxima of *R*



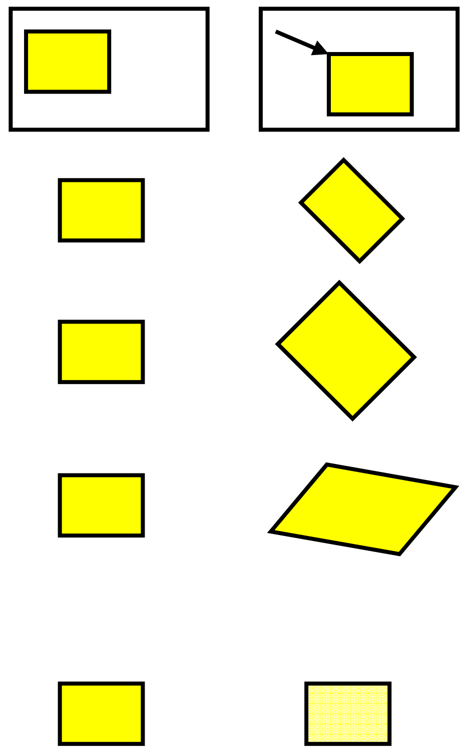
Harris Detector: Steps

Harris detector: Summary of steps

1. Compute Gaussian derivatives at each pixel
2. Compute second moment matrix *A* in a Gaussian window around each pixel
3. Compute corner response function *R*
4. Threshold *R*
5. Find local maxima of response function (non-maximum suppression)

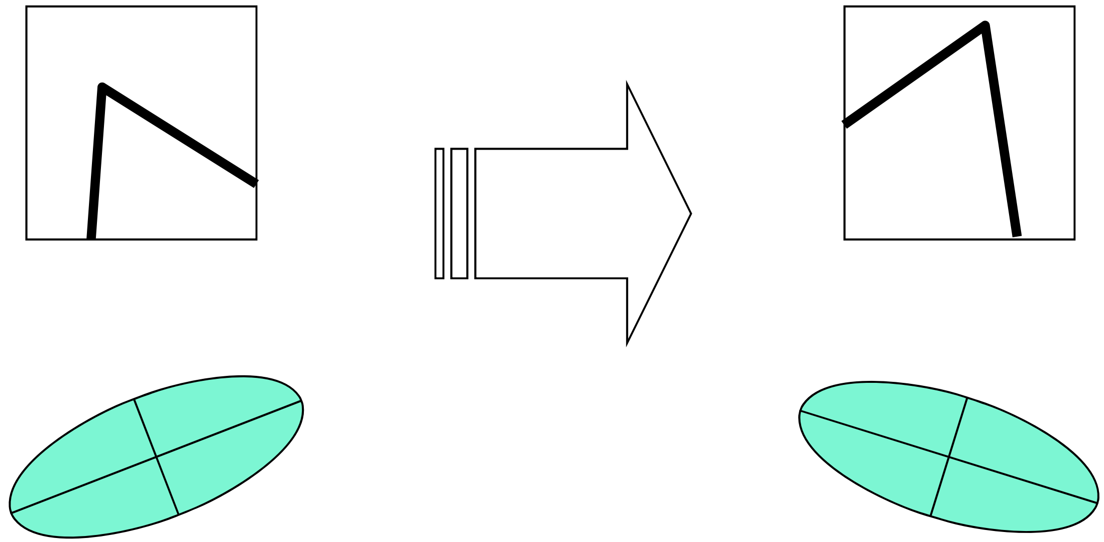
Harris - invariance to transformations

- Geometric transformations
  - translation
  - rotation
  - similitude (rotation + scale change)
  - affine (valide for local planar objects)
- Photometric transformations
  - Affine intensity changes ($I \rightarrow a\,I + b$)



Harris Detector: Invariance Properties

- Rotation

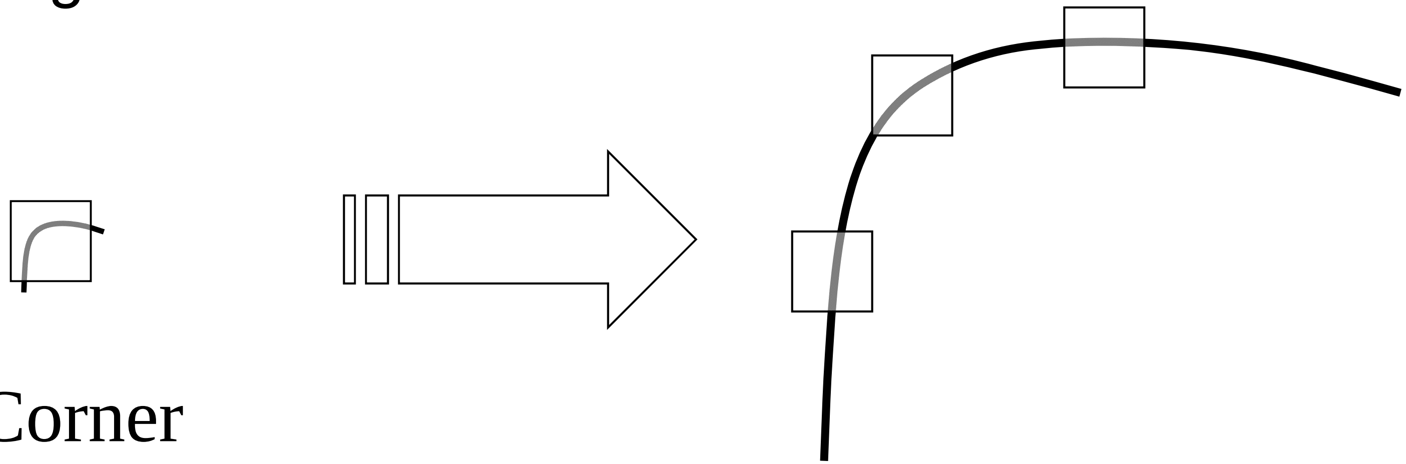


Ellipse rotates but its shape (i.e. eigenvalues) remains the same

Corner response R is invariant to image rotation

Harris Detector: Invariance Properties

- Scaling



All points will be classified as edges

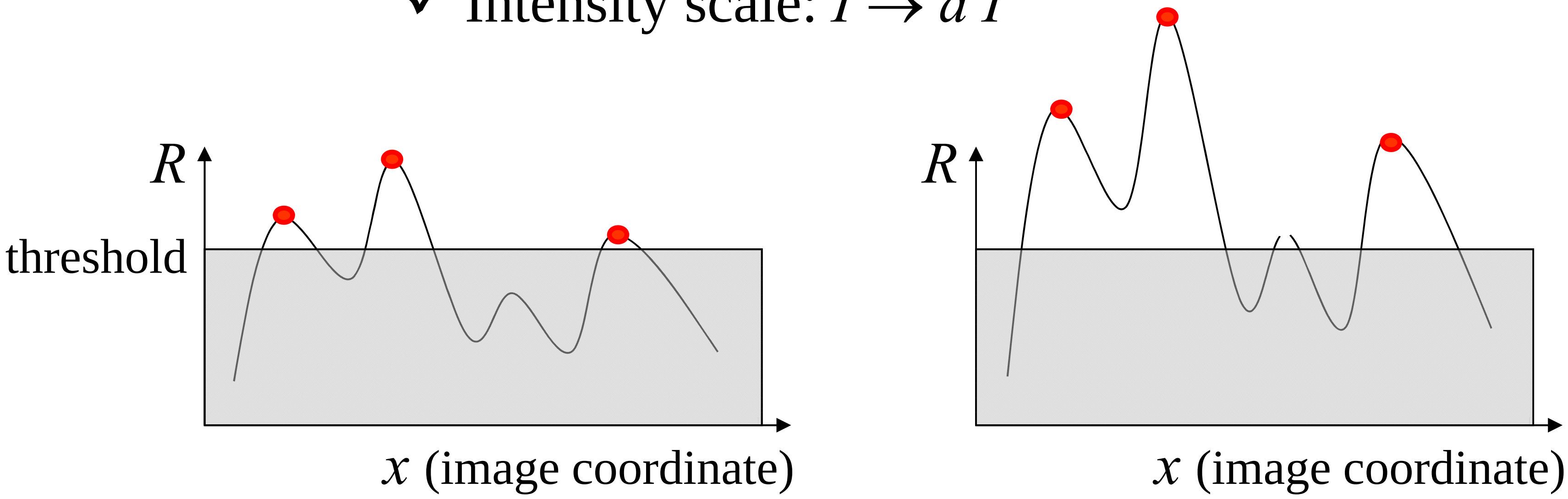
Not invariant to scaling

Harris Detector: Invariance Properties

- Affine intensity change

✓ Only derivatives are used => invariance to intensity shift $I \rightarrow I + b$

✓ Intensity scale: $I \rightarrow a\, I$



Partially invariant to affine intensity change, dependent on type of threshold

Comparison of patches - SSD

Comparison of the intensities in the neighborhood of two interest points

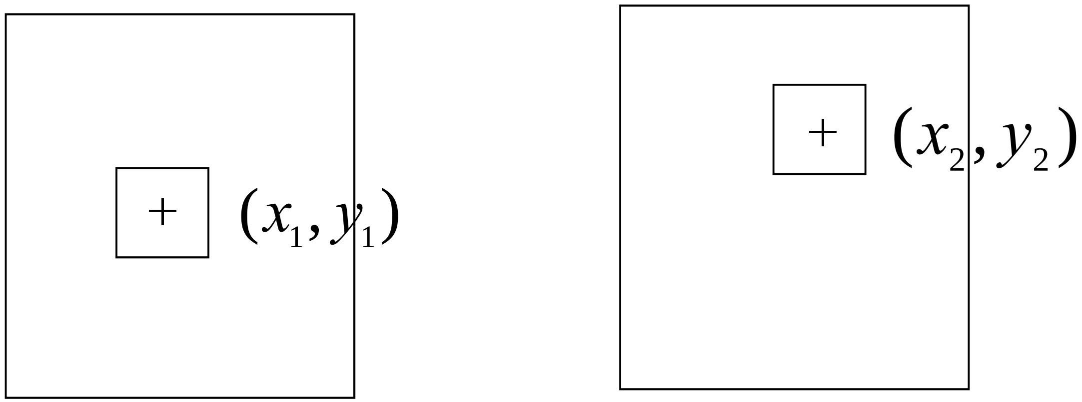


image 1 image 2

SSD : sum of square difference

$$\frac{1}{(2N+1)^2}\sum_{i=-N}^{N}\sum_{j=-N}^{N}(I_1(x_1+i,y_1+j)-I_2(x_2+i,y_2+j))^2$$

Small difference values → similar patches

Comparison of patches

SSD : $\frac{1}{(2N+1)^2}\sum_{i=-N}^{N}\sum_{j=-N}^{N}(I_1(x_1+i,y_1+j)-I_2(x_2+i,y_2+j))^2$

Invariance to photometric transformations?

Intensity changes ($I \rightarrow I + b$)

=> Normalizing with the mean of each patch

$$\frac{1}{(2N+1)^2}\sum_{i=-N}^{N}\sum_{j=-N}^{N}((I_1(x_1+i,y_1+j)-m_1)-(I_2(x_2+i,y_2+j)-m_2))^2$$

Intensity changes ($I \rightarrow aI + b$)

=> Normalizing with the mean and standard deviation of each patch

$$\frac{1}{(2N+1)^2}\sum_{i=-N}^{N}\sum_{j=-N}^{N}\left(\frac{I_1(x_1+i,y_1+j)-m_1}{\sigma_1}-\frac{I_2(x_2+i,y_2+j)-m_2}{\sigma_2}\right)^2$$

Cross-correlation ZNCC

zero normalized SSD

$$\frac{1}{(2N+1)^2}\sum_{i=-N}^{N}\sum_{j=-N}^{N}\left(\frac{I_1(x_1+i,y_1+j)-m_1}{\sigma_1}-\frac{I_2(x_2+i,y_2+j)-m_2}{\sigma_2}\right)^2$$

⬇

ZNCC: zero normalized cross correlation

$$\frac{1}{(2N+1)^2}\sum_{i=-N}^{N}\sum_{j=-N}^{N}\left(\frac{I_1(x_1+i,y_1+j)-m_1}{\sigma_1}\right)\cdot\left(\frac{I_2(x_2+i,y_2+j)-m_2}{\sigma_2}\right)$$

ZNCC values between -1 and 1, 1 when identical patches
in practice threshold around 0.5

Local descriptors

- Pixel values
- Greyvalue derivatives, differential invariants [Koenderink'87]
- SIFT descriptor [Lowe'99]
- SURF descriptor [Bay et al.'08]
- DAISY descriptor [Tola et al.'08, Windler et al'09]
- LIOP descriptor [Wang et al.'11]
- Recent patch descriptors based on CNN features [Brox et al.'15, Paulin et al.'15,…]

Local descriptors

- Greyvalue derivatives
  - Convolution with Gaussian derivatives

$$\mathbf{v}(x,y) = \begin{pmatrix} I(x,y) * G(\sigma) \\ I(x,y) * G_x(\sigma) \\ I(x,y) * G_y(\sigma) \\ I(x,y) * G_{xx}(\sigma) \\ I(x,y) * G_{xy}(\sigma) \\ I(x,y) * G_{yy}(\sigma) \\ \vdots \end{pmatrix}$$

$$I(x,y) * G(\sigma) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x',y',\sigma) I(x-x', y-y') dx'dy'$$

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2})$$

Local descriptors

Notation for greyvalue derivatives [Koenderink'87]

$$\mathbf{v}(x,y) = \begin{pmatrix} I(x,y) * G(\sigma) \\ I(x,y) * G_x(\sigma) \\ I(x,y) * G_y(\sigma) \\ I(x,y) * G_{xx}(\sigma) \\ I(x,y) * G_{xy}(\sigma) \\ I(x,y) * G_{yy}(\sigma) \\ \vdots \end{pmatrix} = \begin{pmatrix} L(x,y) \\ L_x(x,y) \\ L_y(x,y) \\ L_{xx}(x,y) \\ L_{xy}(x,y) \\ L_{yy}(x,y) \\ \vdots \end{pmatrix}$$

Invariance?

Local descriptors – rotation invariance

Invariance to image rotation : differential invariants [Koen87]

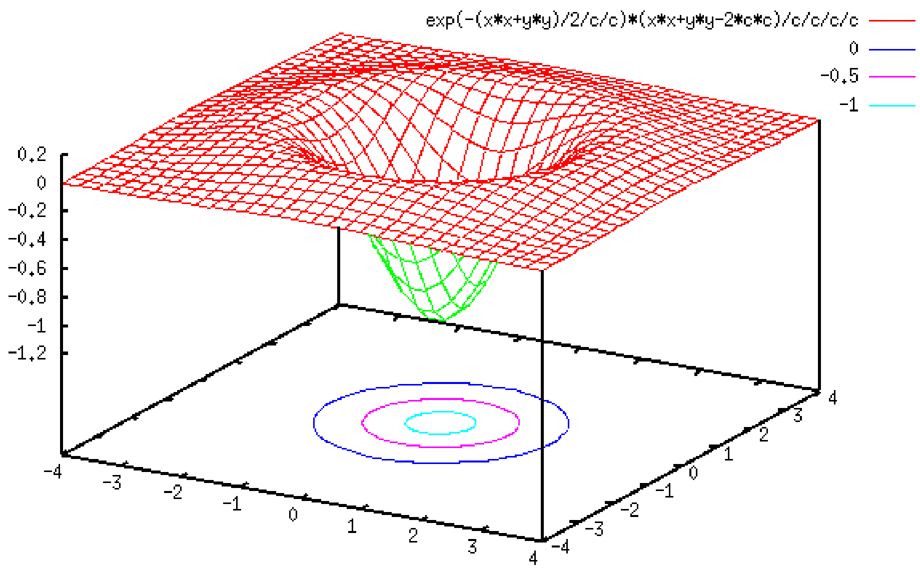
gradient magnitude → $L_x L_x + L_y L_y$

Laplacian → $L_{xx} + L_{yy}$

$$\begin{bmatrix} L \\ L_x L_x + L_y L_y \\ L_{xx} L_x L_x + 2 L_{xy} L_x L_y + L_{yy} L_{yy} \\ L_{xx} + L_{yy} \\ L_{xx} L_{xx} + 2 L_{xy} L_{xy} + L_{yy} L_{yy} \\ \cdots \\ \cdots \\ \cdots \\ \cdots \end{bmatrix}$$

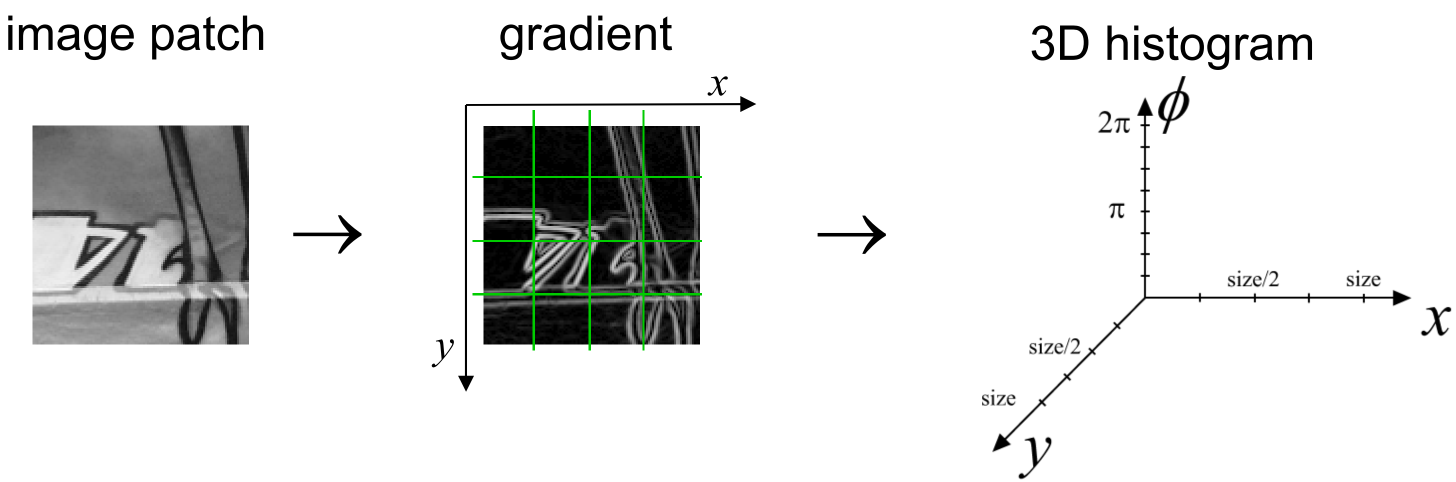
Laplacian of Gaussian (LOG)

$$LOG = G_{xx}(\sigma) + G_{yy}(\sigma)$$



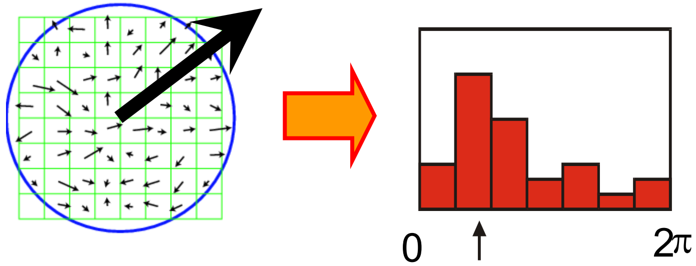
SIFT descriptor [Lowe'99]

- Approach
  - 8 orientations of the gradient
  - 4x4 spatial grid
  - Dimension 128
  - soft-assignment to spatial bins
  - normalization of the descriptor to norm one
  - comparison with Euclidean distance

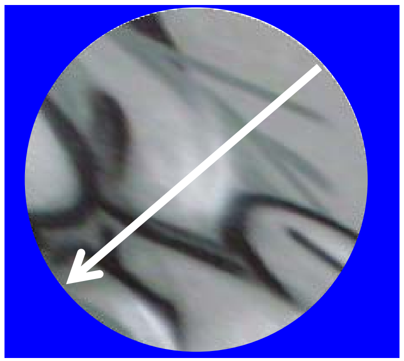


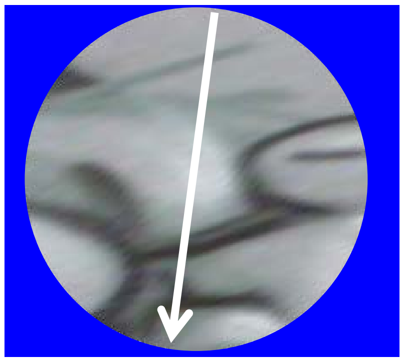
Local descriptors - rotation invariance

- Estimation of the dominant orientation
  - extract gradient orientation
  - histogram over gradient orientation
  - peak in this histogram



- Rotate patch in dominant direction





Local descriptors – illumination change

- Robustness to illumination changes

  in case of an affine transformation $I_1(\mathbf{x}) = aI_2(\mathbf{x}) + b$

- Normalization of the image patch with mean and variance

Invariance to scale changes

- Scale change between two images

- Scale factor s can be eliminated

- Support region for calculation!!
  - In case of a convolution with Gaussian derivatives defined by σ

$$I(x,y) * G(\sigma) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x',y',\sigma) I(x-x',y-y') dx'dy'$$

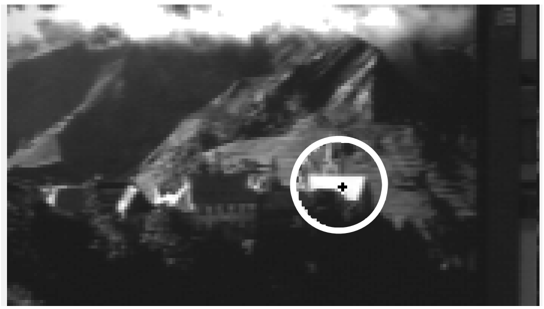
$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)$$

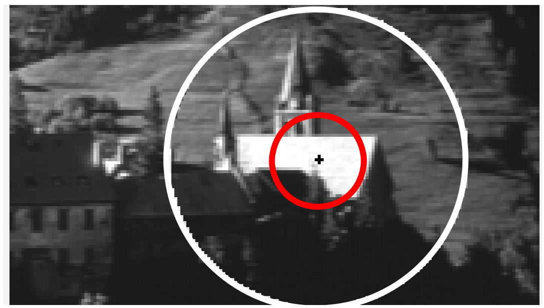
Overview

- Introduction to local features
- Harris interest points + SSD, ZNCC, SIFT
- **Scale invariant interest point detectors**

Scale invariance - motivation

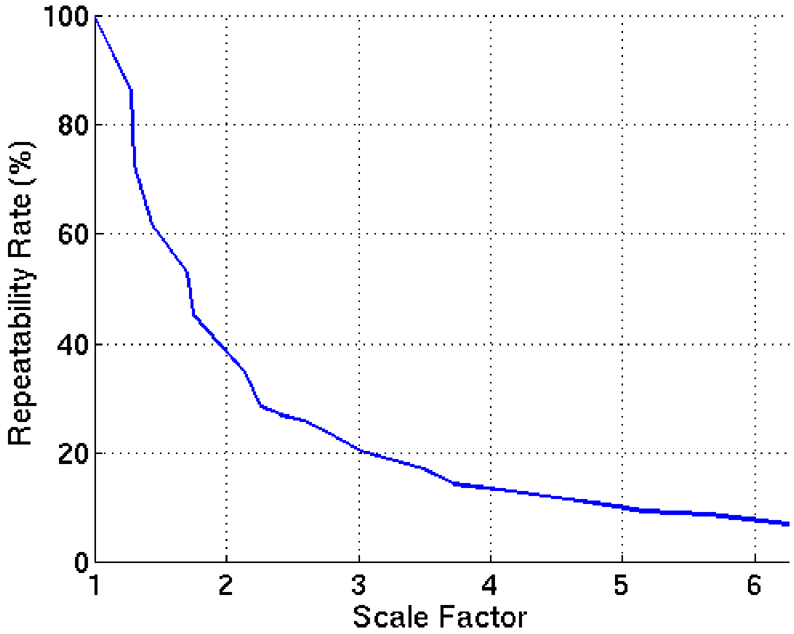
- Description regions have to be adapted to scale changes

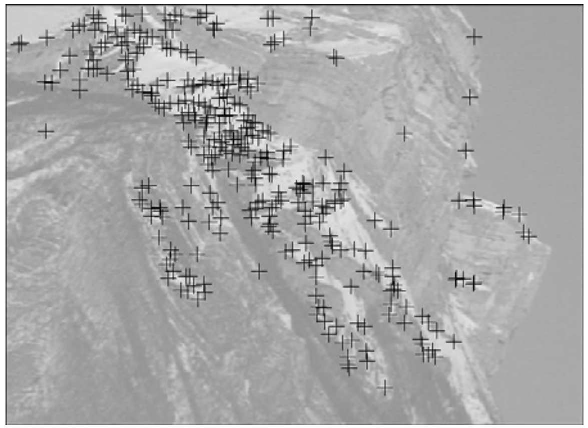


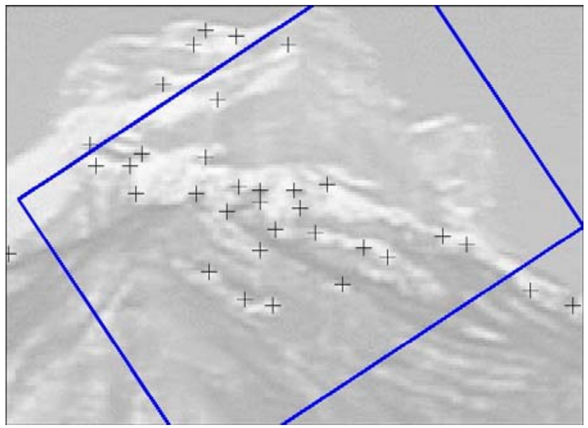


- Interest points have to be repeatable for scale changes

Harris detector + scale changes







Repeatability rate

$$R(\varepsilon) = \frac{|\{(\mathbf{a}_i, \mathbf{b}_i) \mid dist(H(\mathbf{a}_i), \mathbf{b}_i) < \varepsilon\}|}{\max(|\mathbf{a}_i|, |\mathbf{b}_i|)}$$

Scale adaptation

Scale change between two images

$$I_1\begin{pmatrix}x_1\\y_1\end{pmatrix}=I_2\begin{pmatrix}x_2\\y_2\end{pmatrix}=I_2\begin{pmatrix}sx_1\\sy_1\end{pmatrix}$$

Scale adapted derivative calculation

Scale adaptation

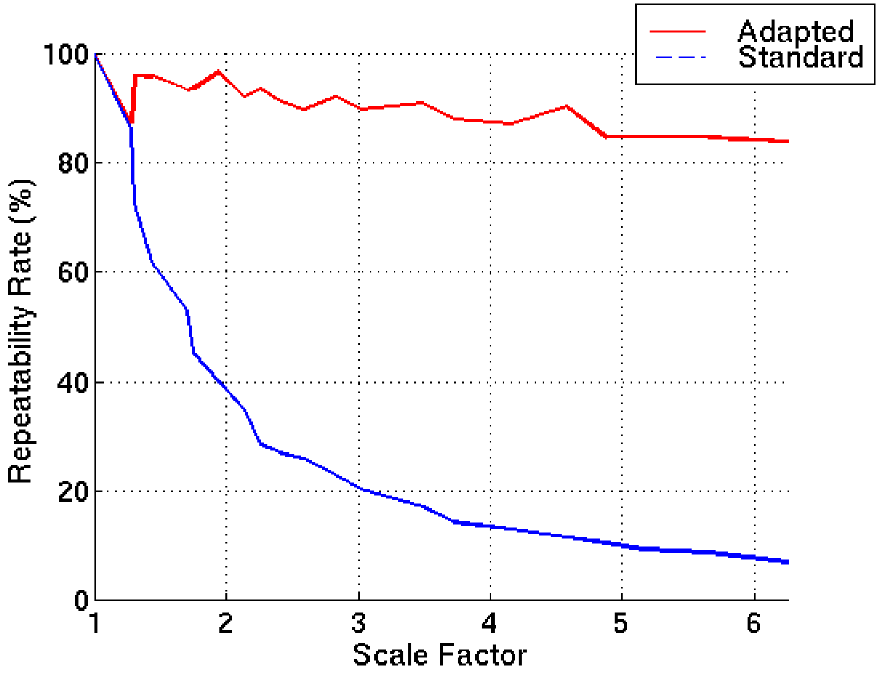
Scale change between two images

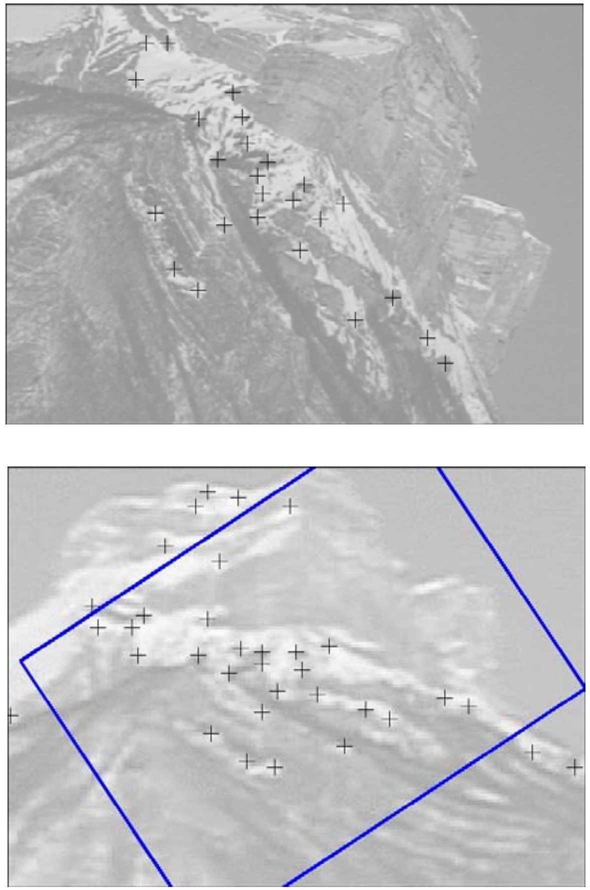
$$I_1\begin{pmatrix}x_1\\y_1\end{pmatrix} = I_2\begin{pmatrix}x_2\\y_2\end{pmatrix} = I_2\begin{pmatrix}sx_1\\sy_1\end{pmatrix}$$

Scale adapted derivative calculation

$$I_1\begin{pmatrix}x_1\\y_1\end{pmatrix} \otimes G_{i_1 \dots i_n}(\sigma) = s^n I_2\begin{pmatrix}x_2\\y_2\end{pmatrix} \otimes G_{i_1 \dots i_n}(s\sigma)$$

Harris detector – adaptation to scale





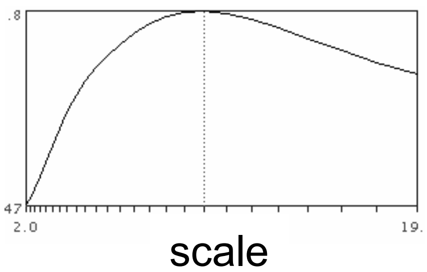
Scale selection

- For a point compute a value (gradient, Laplacian etc.) at several scales
- Normalization of the values with the scale factor

  e.g. Laplacian $| s^2 (L_{xx} + L_{yy}) |$

- Select scale s^* at the maximum → characteristic scale

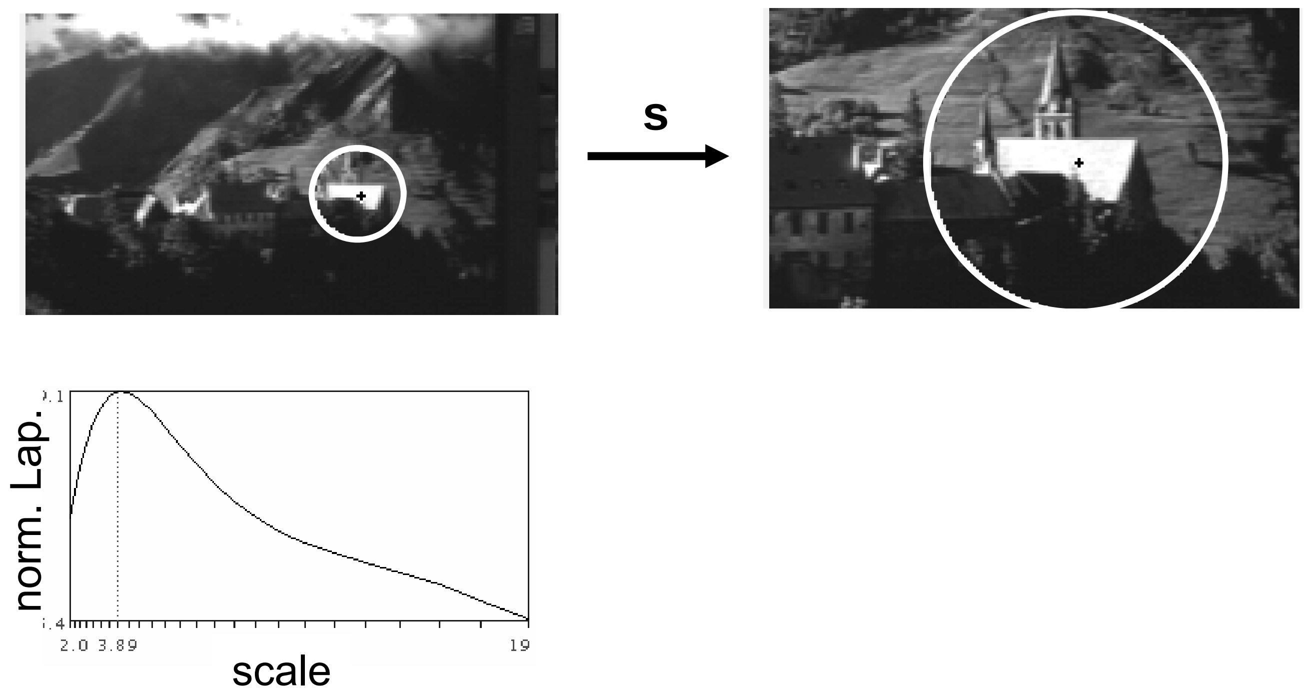
$| s^2 (L_{xx} + L_{yy}) |$



- Exp. results show that the Laplacian gives best results

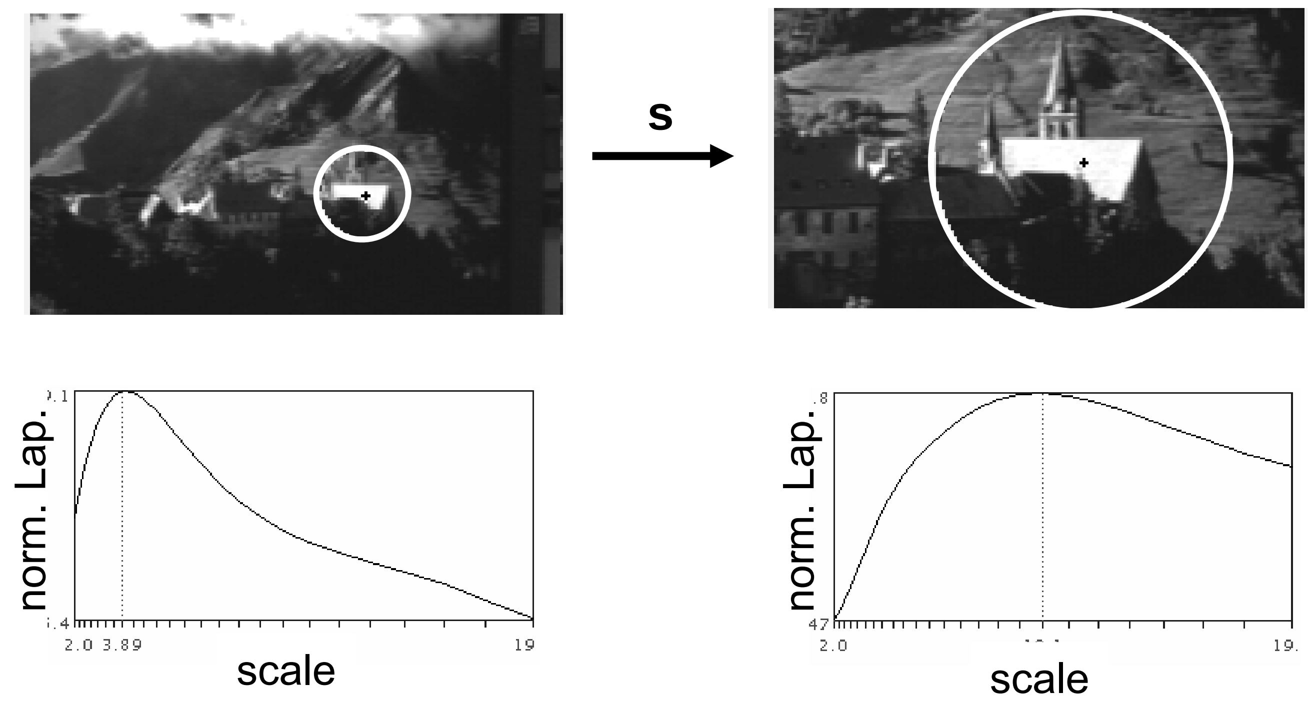
Scale selection

- Scale invariance of the characteristic scale



Scale selection

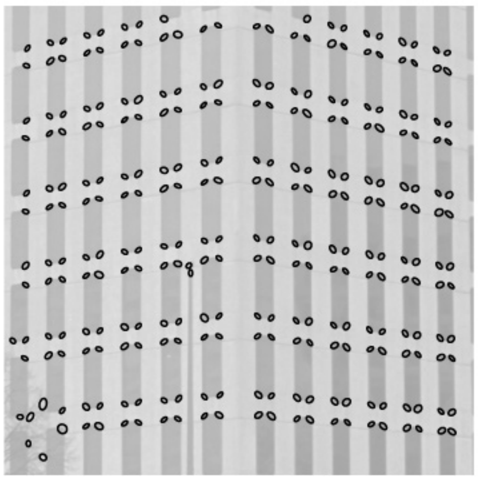
- Scale invariance of the characteristic scale



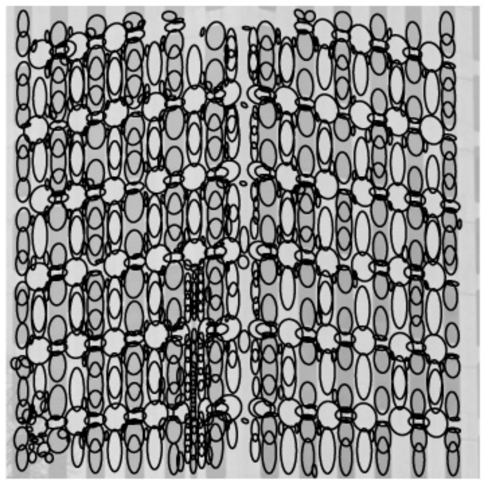
- Relation between characteristic scales $s \cdot s_1^* = s_2^*$

Scale-invariant detectors

- Harris-Laplace (Mikolajczyk & Schmid’01)
- Laplacian detector (Lindeberg’98)
- Difference of Gaussian (SIFT detector, Lowe’99)



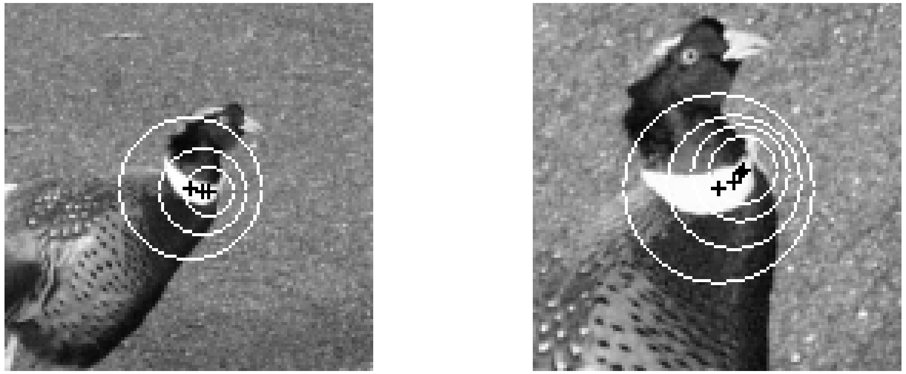
Harris-Laplace



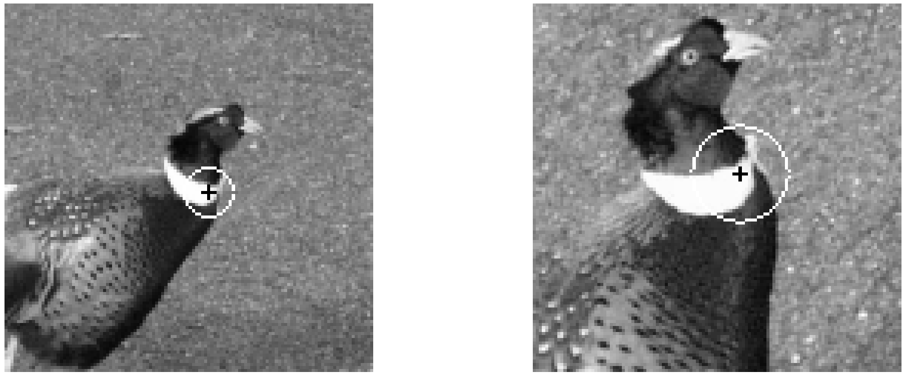
Laplacian

Harris-Laplace

multi-scale Harris points



selection of points at maximum of Laplacian



➡ invariant points + associated regions [Mikolajczyk & Schmid’01]

Matching results

213 / 190 detected interest points

Matching results

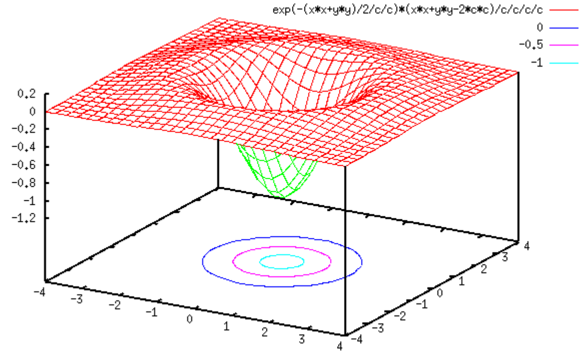
58 points are initially matched

Matching results

32 points are matched after verification – all correct

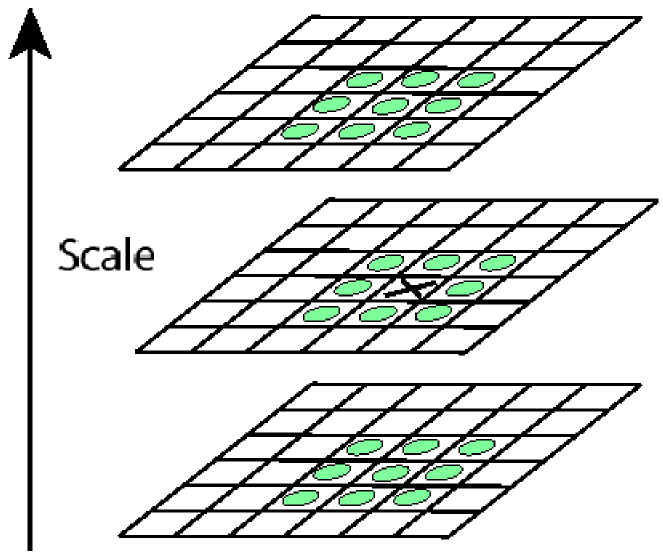
LOG detector

Convolve image with scale-normalized Laplacian at several scales



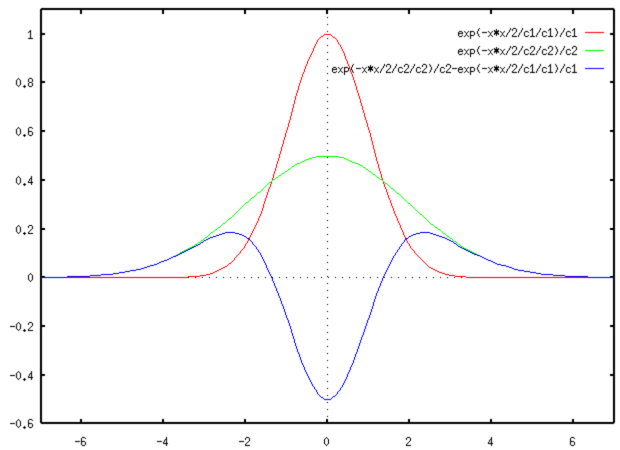
$$LOG = s^2 (G_{xx}(\sigma) + G_{yy}(\sigma))$$

Detection of maxima and minima of Laplacian in scale space

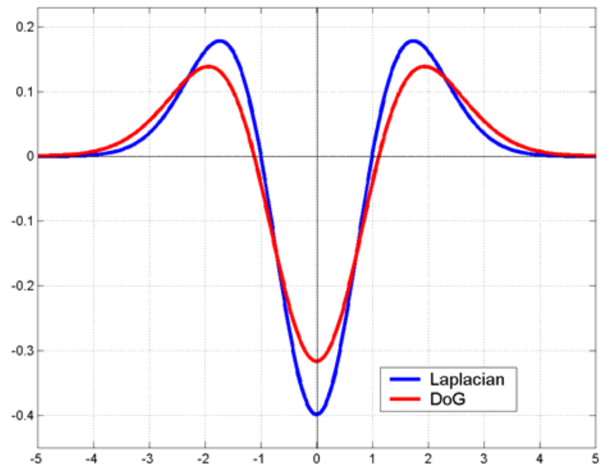


Efficient implementation

- Difference of Gaussian (DOG) approximates the Laplacian $DOG = G(k\sigma) - G(\sigma)$

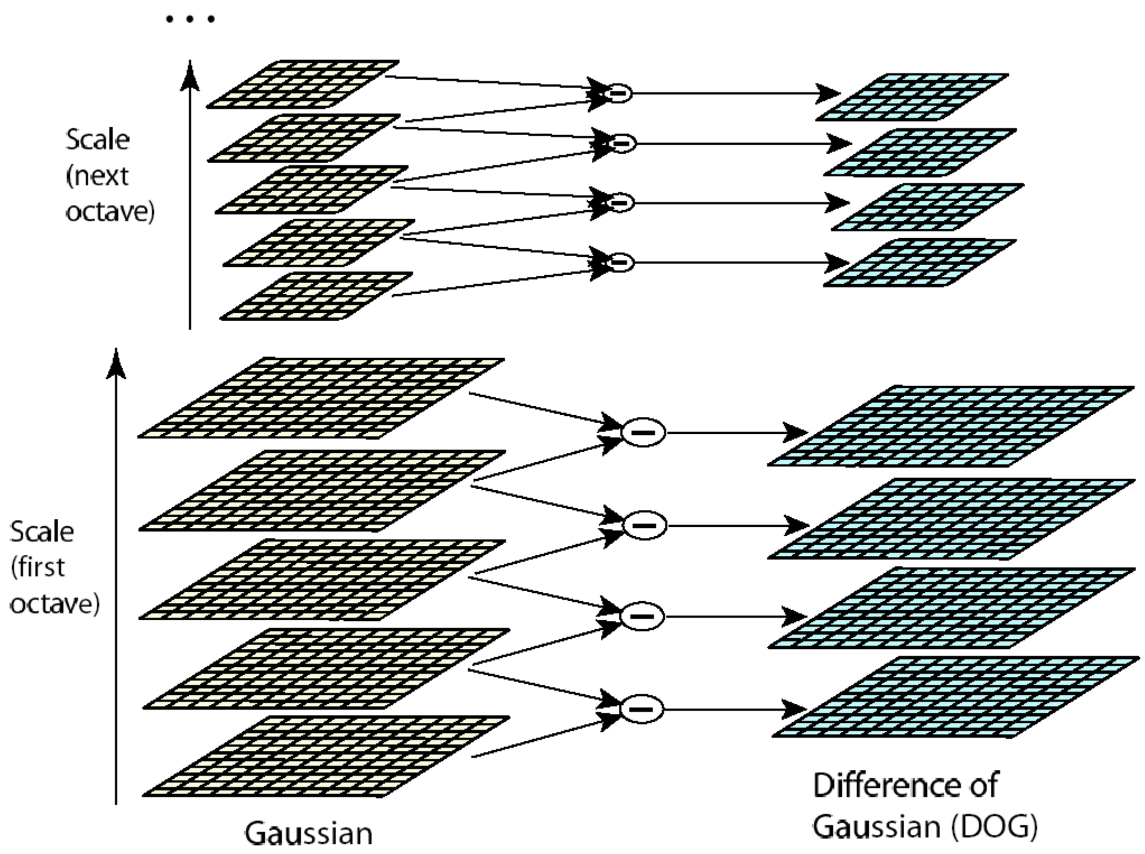


- Error due to the approximation



DOG detector

- Fast computation, scale space processed one octave at a time



David G. Lowe. "Distinctive image features from scale-invariant keypoints."I*JCV* 60 (2).

Maximally stable extremal regions (MSER) [Matas'02]

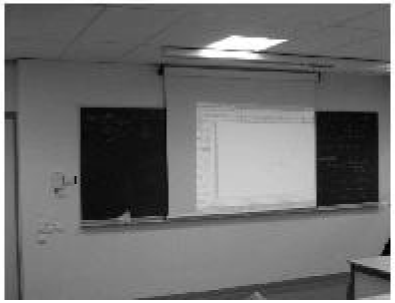
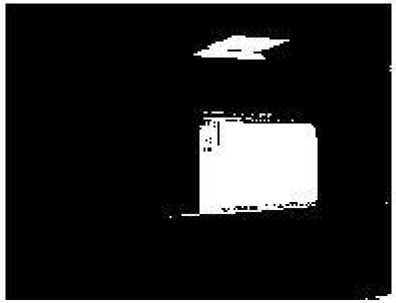
- Extremal regions: connected components in a thresholded image (all pixels above/below a threshold)

- Maximally stable: minimal change of the component (area) for a change of the threshold, i.e. region remains stable for a change of threshold

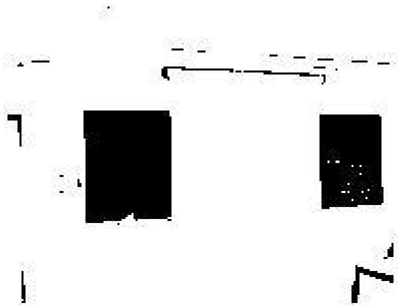
- Excellent results in a recent comparison

Maximally stable extremal regions (MSER)

Examples of thresholded images

high threshold

low threshold

MSER