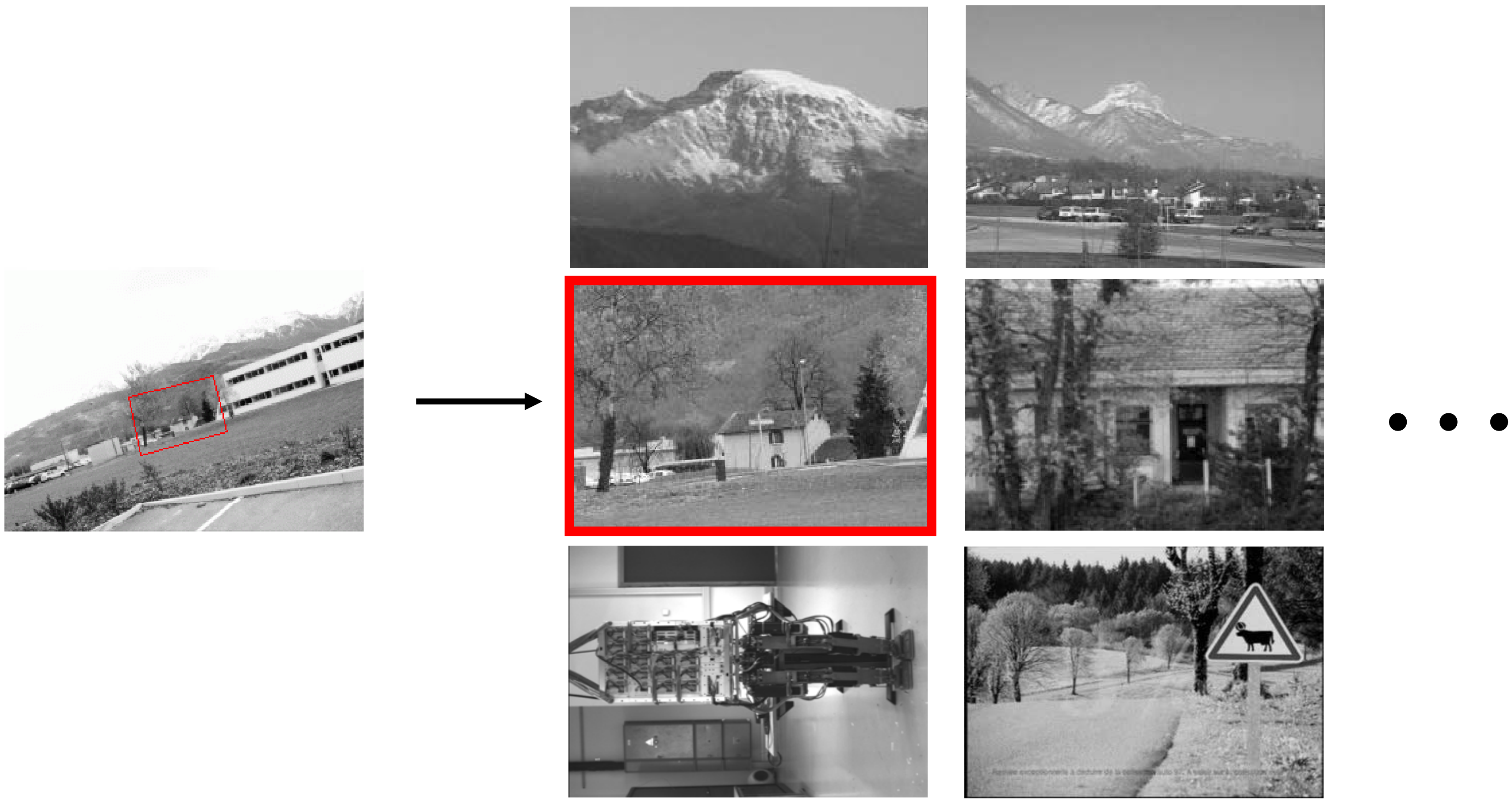
Instance-level recognition

Cordelia Schmid & Josef Sivic
INRIA

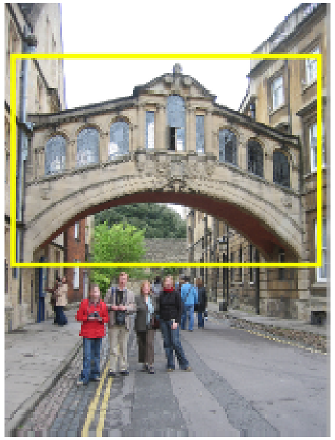
Instance-level recognition

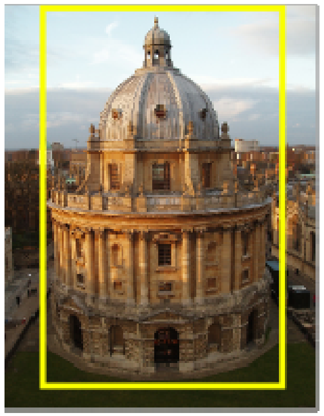
Particular objects and scenes, large databases



Application

Search photos on the web for particular places





Find these landmarks

...in these images and 1M more

Applications

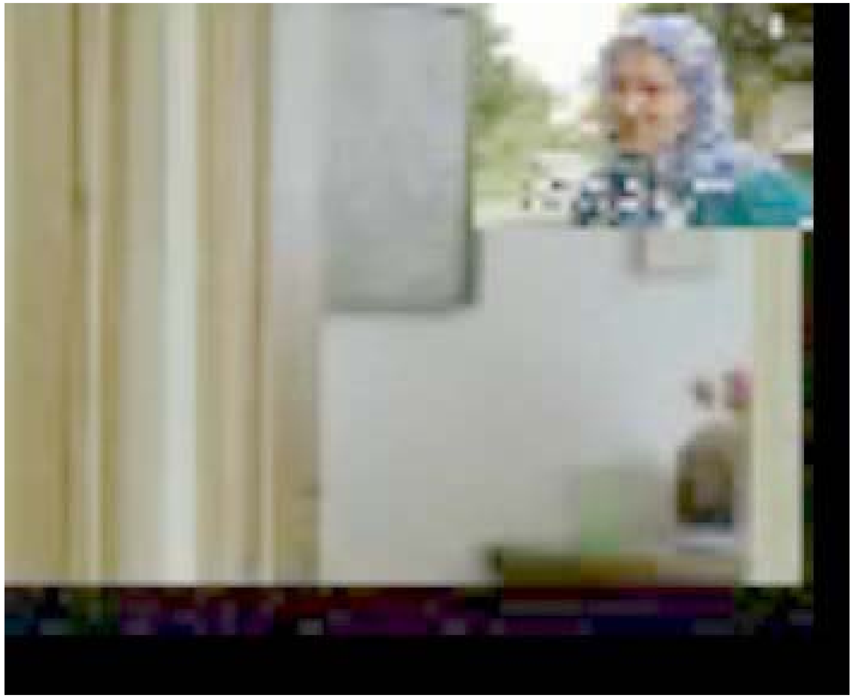
- Take a picture of a product or advertisement
  → find relevant information on the web

[Google Goggles, Milpix Pixee]

Applications

- Copy detection for images and videos

Query video



Search in 200h of video

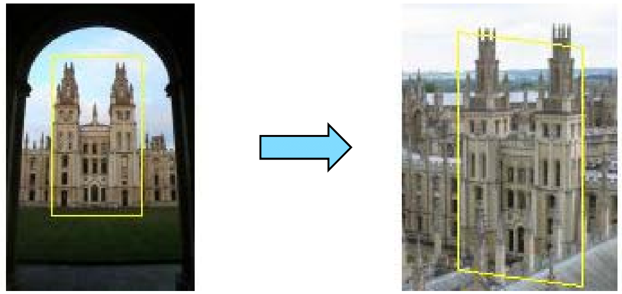
Difficulties

Find the object despite

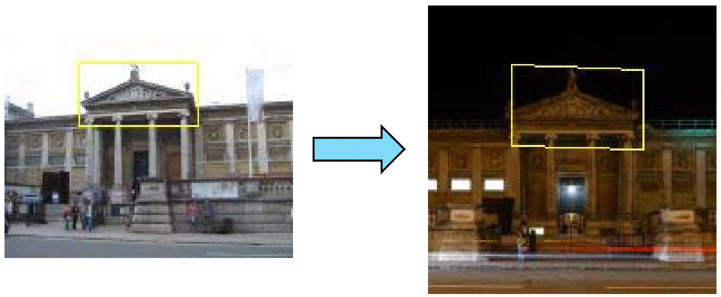
- large changes in scale, viewpoint, lighting
- crop and occlusion
- not much texture/structure

➢ requires local invariant descriptors

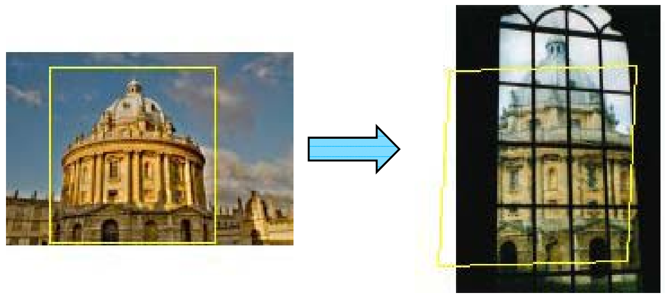
Scale



Viewpoint



Lighting



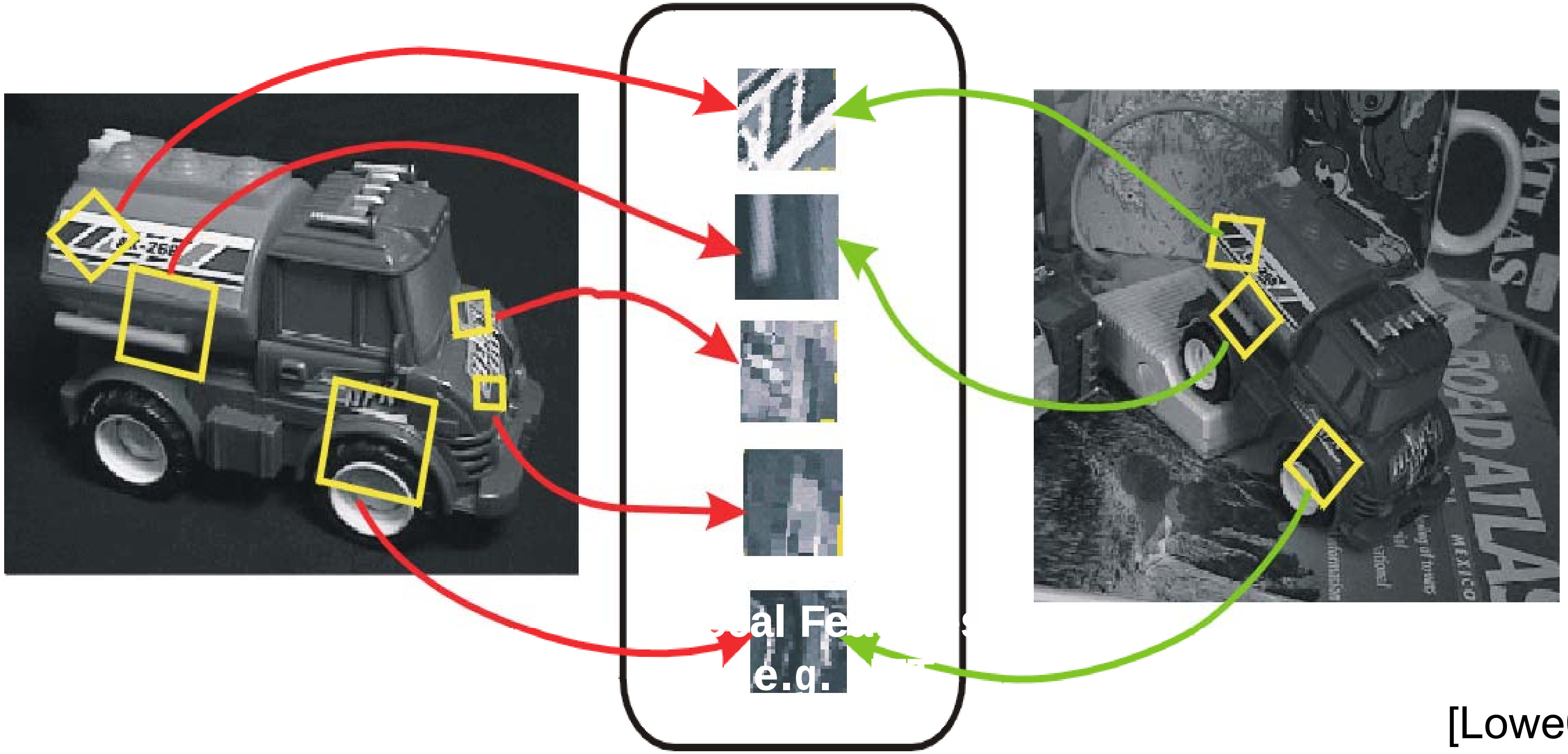
Occlusion

Difficulties

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

Approach: matching local invariant descriptors

Image content is transformed into local features that are invariant to geometric and photometric transformations

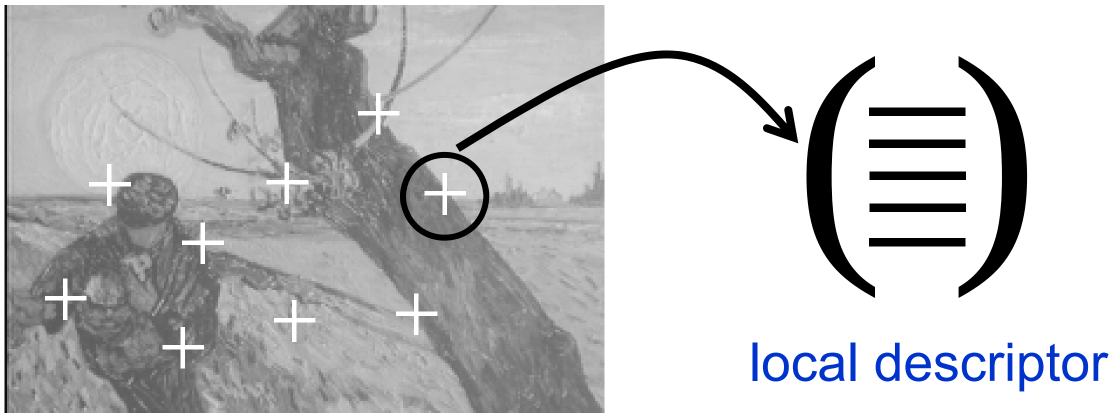


[Lowe04]

Overview

- **Local invariant features** (C. Schmid)
- Matching and recognition with local features (J. Sivic)
- Efficient visual search (J. Sivic)
- Very large scale search (C. Schmid)
- Practical session

Local features



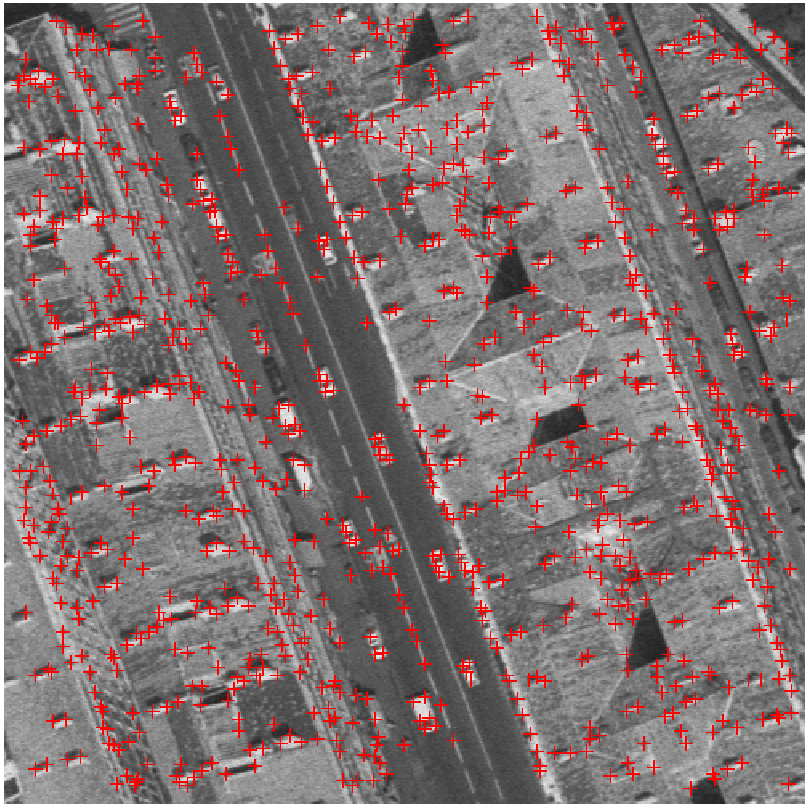
Several / many local descriptors per image

Robust to occlusion/clutter, no object segmentation required

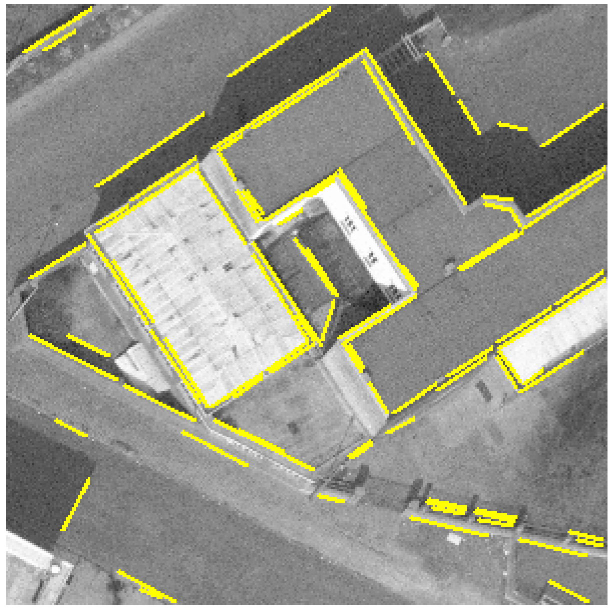
Photometric : distinctive

Invariant : to image transformations + illumination changes

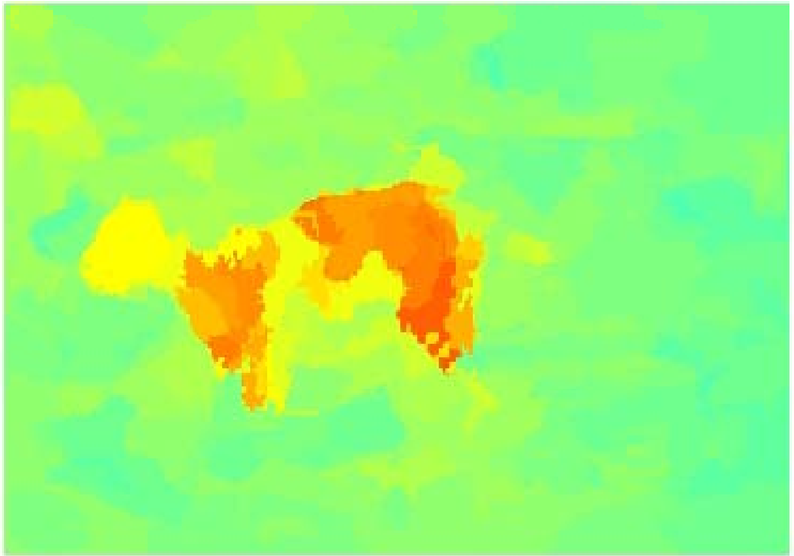
Local features: interest points



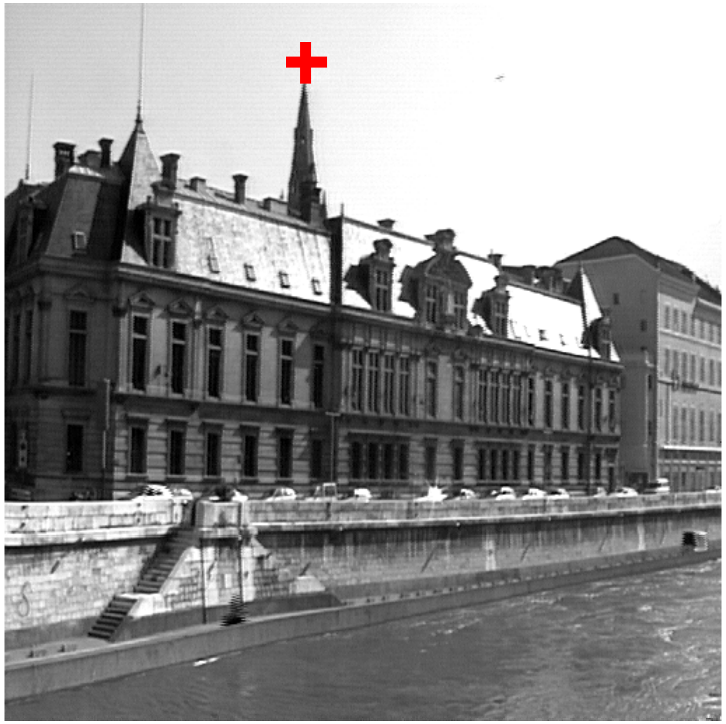
Local features: Contours/lines



Local features: regions

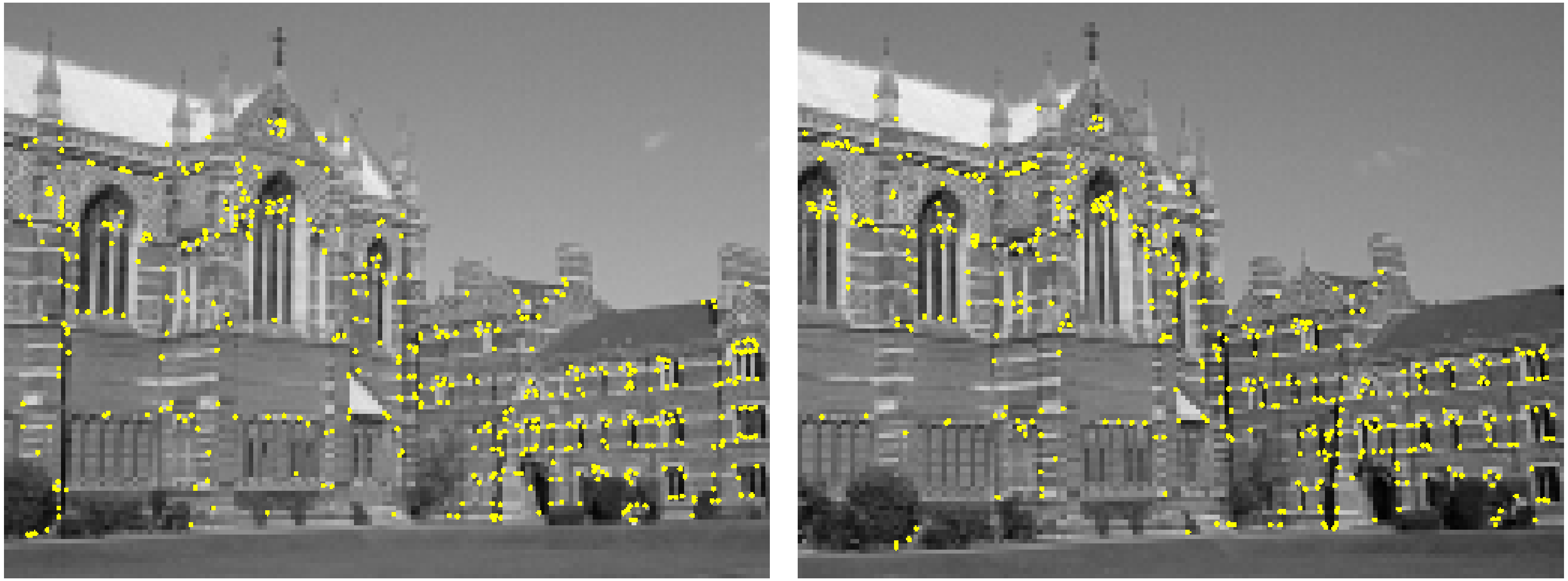


Matching & instance-level recognition → Interest points



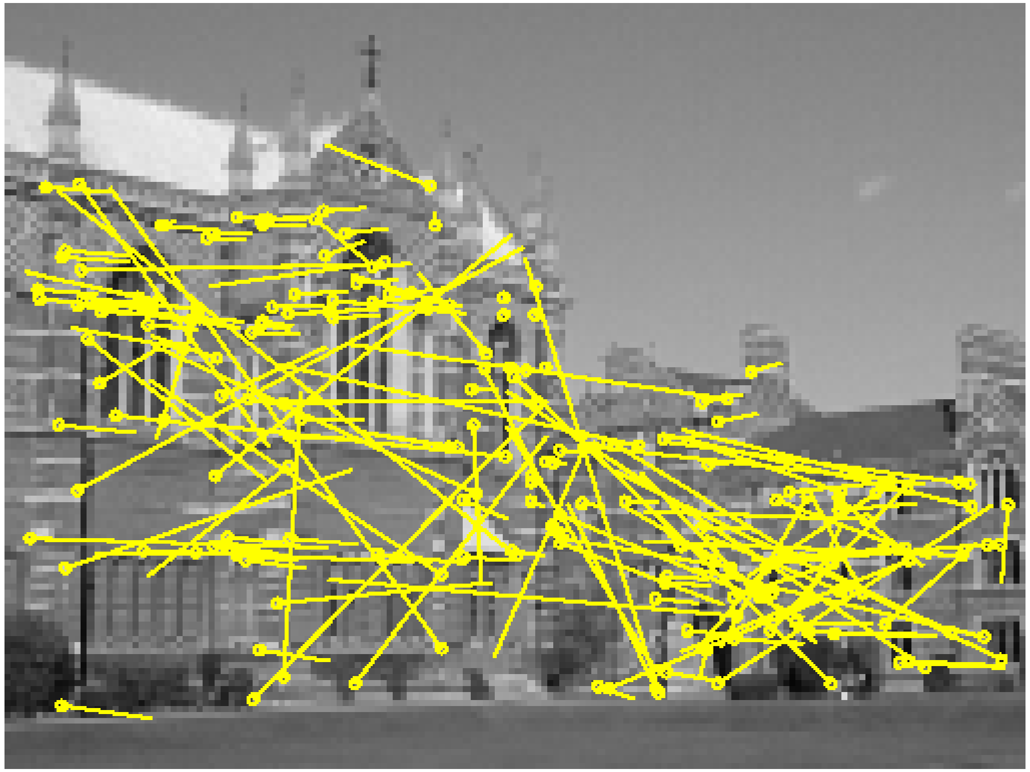
Find corresponding locations in two images

Illustration – Matching



Interest points extracted with Harris detector (~ 500 points)

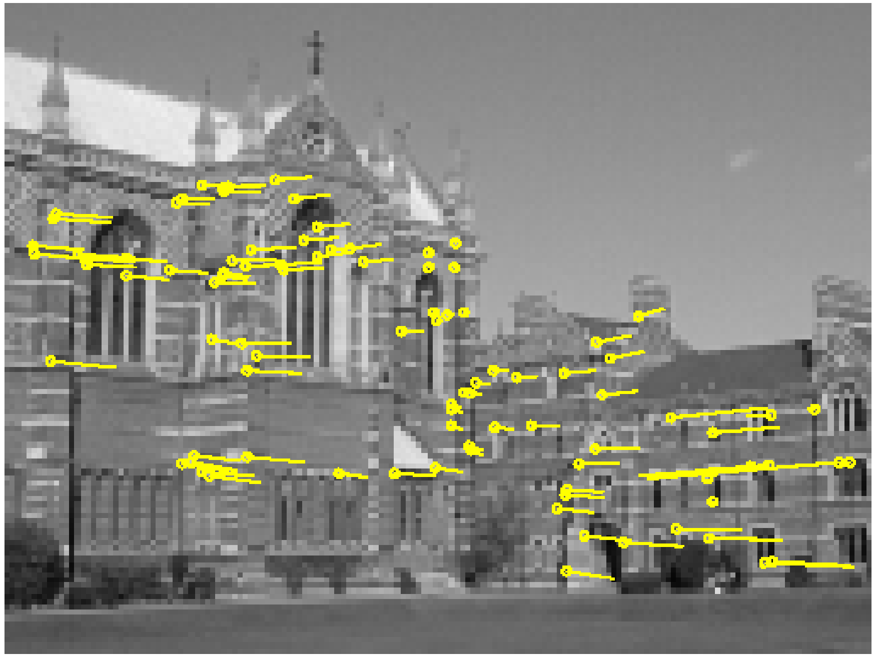
Illustration – Matching



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

Illustration – Matching

Global constraint - Robust estimation of the fundamental matrix



99 inliers

89 outliers

Harris detector [Harris & Stephens'88]

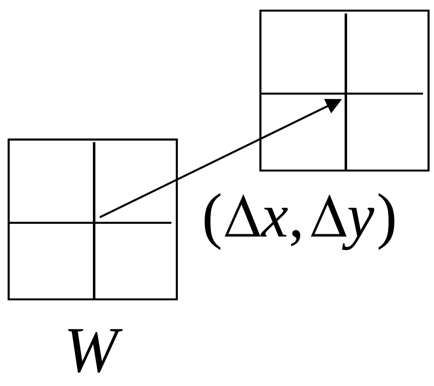
Based on auto-correlation

Important difference in all directions => interest point

Harris detector

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

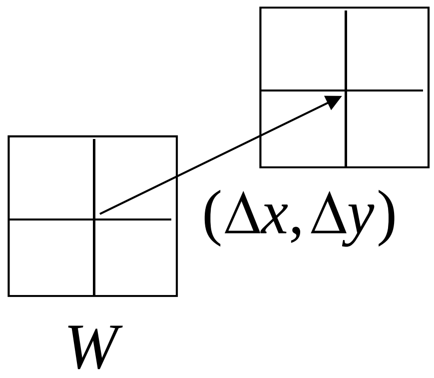
$$a(x,y) = \sum_{(x_k, y_k) \in W(x,y)} \left(I(x_k, y_k) - I(x_k + \Delta x, y_k + \Delta y) \right)^2$$



Harris detector

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

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



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

Harris detector

Discret shifts are avoided based on the auto-correlation matrix

with first order approximation

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

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

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

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

$$G \otimes \begin{bmatrix} I_x^{\ 2} & I_x I_y \\ I_x I_y & I_y^{\ 2} \end{bmatrix}$$

  – 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

Harris detector

- Cornerness function

$$f = \det(a) - k(trace(a))^2 = \lambda_1\lambda_2 - k(\lambda_1 + \lambda_2)^2$$

Reduces the effect of a strong contour

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

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

Comparison of patches - SSD

Comparison of the intensities in the neighborhood of two interest points

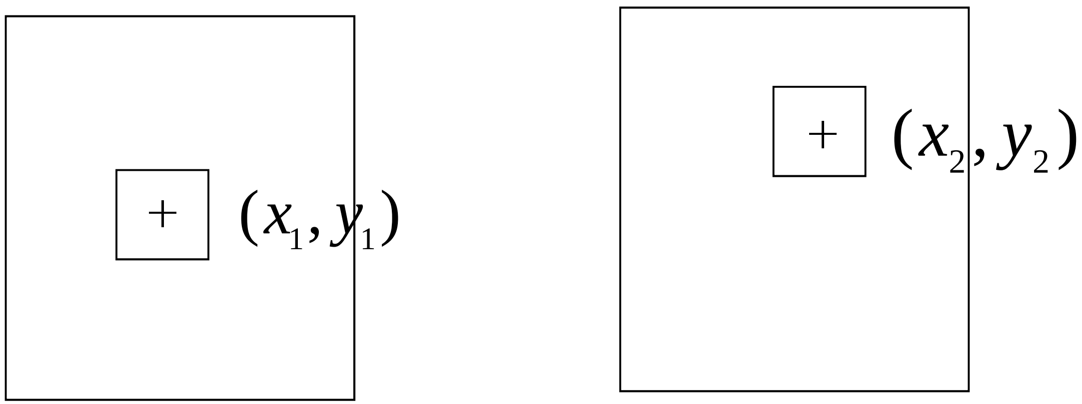


image 1 image 2

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

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

$$\Downarrow$$

ZNCC: zero normalized cross correlation

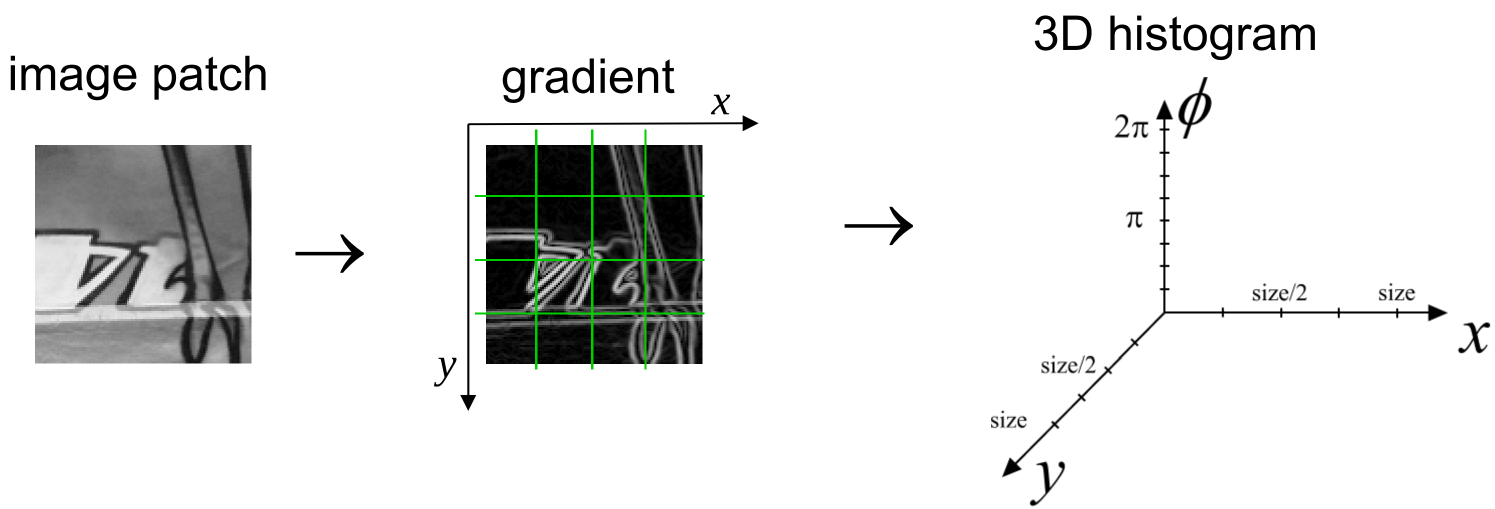
$$\frac{1}{(2N+1)^2}\sum_{i=-N}^{N}\sum_{j=-N}^{N}\left(\frac{I_1(x_1+i,y_1+j)-m_1}{\sigma_1}\right)\cdot\left(\frac{I_2(x_2+i,y_2+j)-m_2}{\sigma_2}\right)$$

ZNCC values between -1 and 1, 1 when identical patches

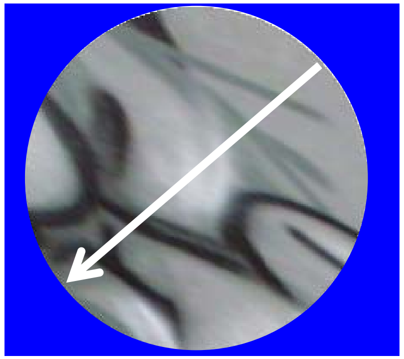
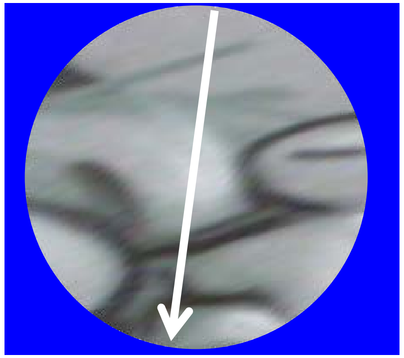
in practice threshold around 0.5

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



SIFT - rotation invariance

- Estimation of the dominant orientation
  - extract gradient orientation
  - histogram over gradient orientations
  - peak in this histogram
- Rotate patch in dominant direction

Other local descriptors

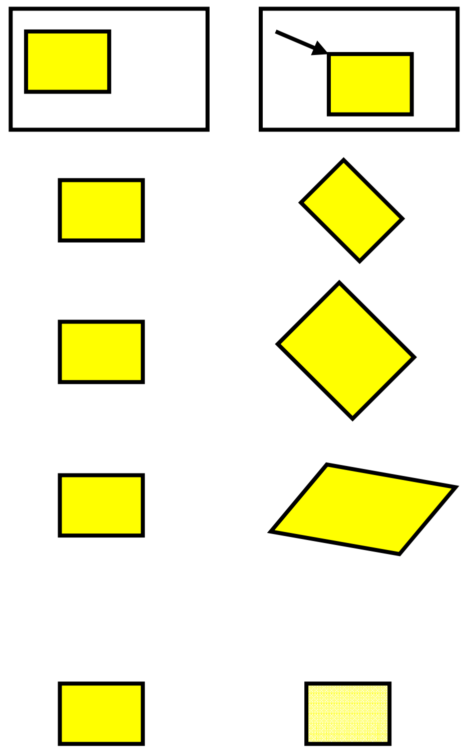
- Greyvalue derivatives, differential invariants [Koenderink'87]
- Shape context [Belongie et al.'02]
- SURF descriptor [Bay et al.'08]
- DAISY descriptor [Tola et al.'08, Windler et al'09]
- LIOP descriptor [Wang et al.'11]

Comparison – descriptors

- Robust region descriptors better than point-wise descriptors [Mikolajczyk & Schmid'05]
- Significant difference between SIFT and low dimensional descriptors as well as cross-correlation
- Performance of the descriptor is relatively independent of the detector
- Recently, faster descriptors

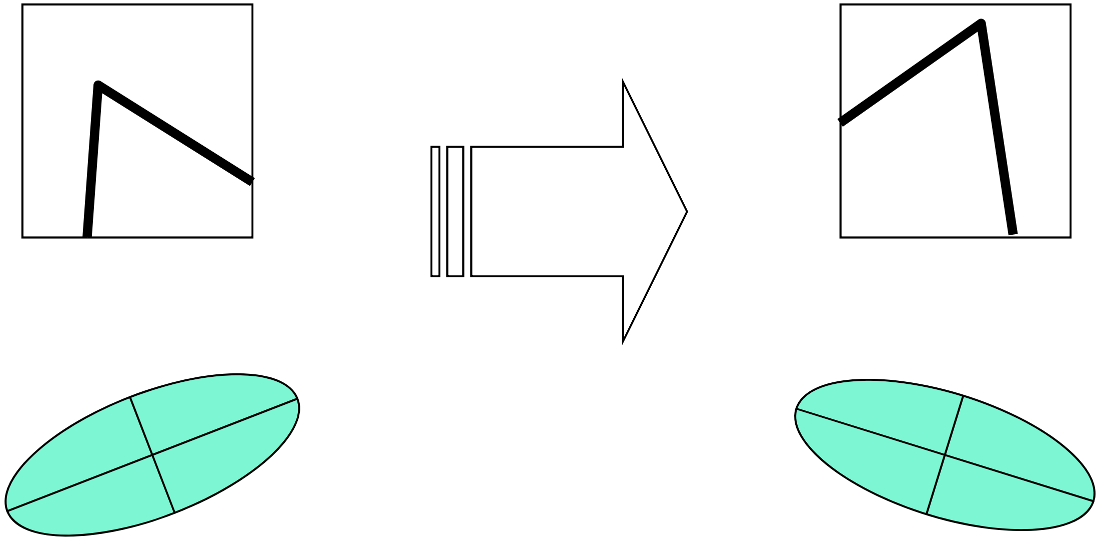
Invariance to transformations – Harris

- Geometric transformations
  - translation
  - rotation
  - similarity (rotation + scale change + translation)
  - affine (2x2 transformation matrix + translation)
- 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

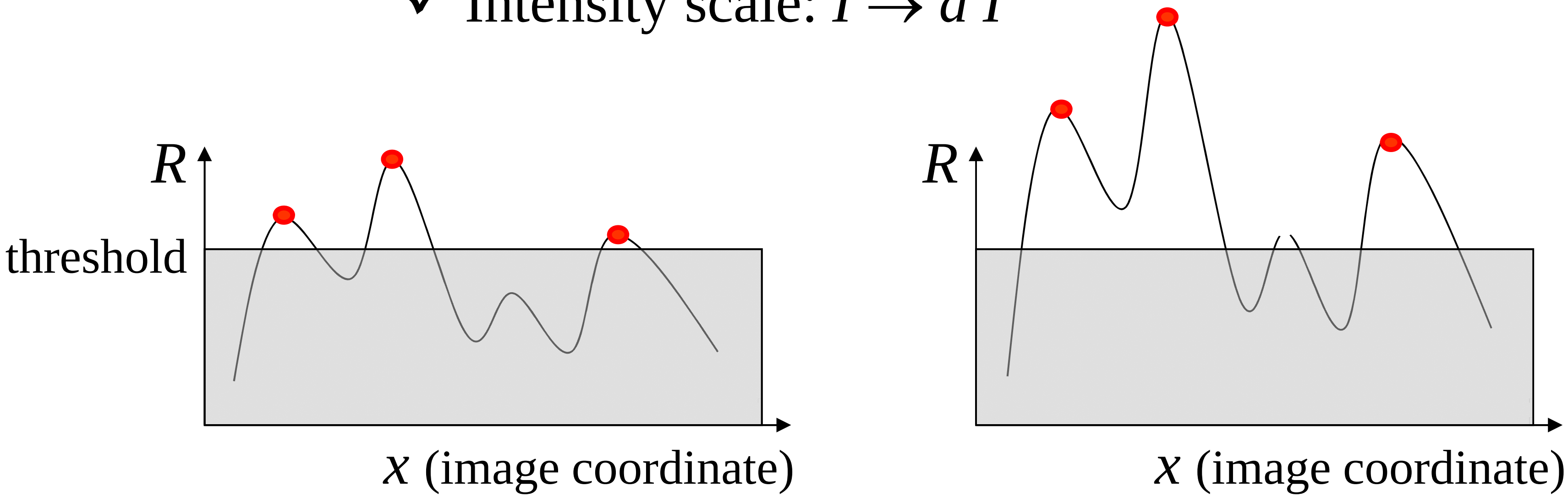
Corner response R is invariant to image rotation

Harris Detector: Invariance Properties

- Affine intensity change

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

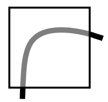
✓ Intensity scale: $I \rightarrow a\, I$



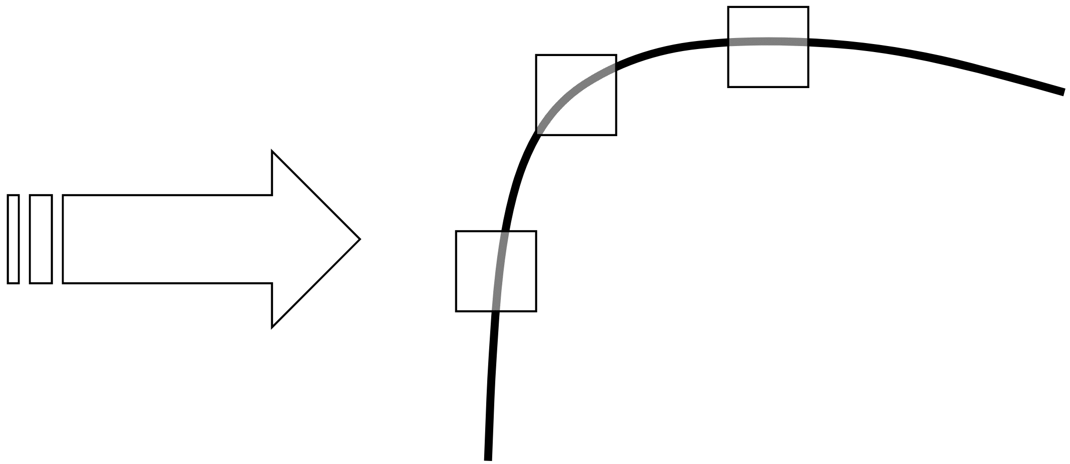
Partially invariant to affine intensity change, dependent on type of threshold

Harris Detector: Invariance Properties

- Scaling



Corner

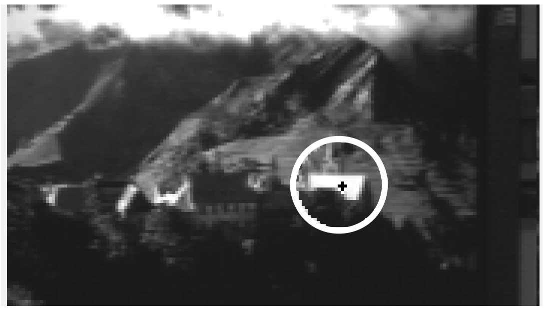


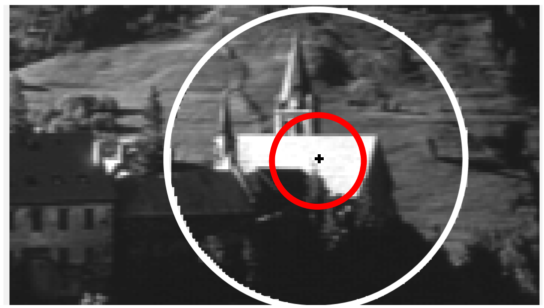
All points will be classified as edges

Not invariant to scaling

Scale invariance - motivation

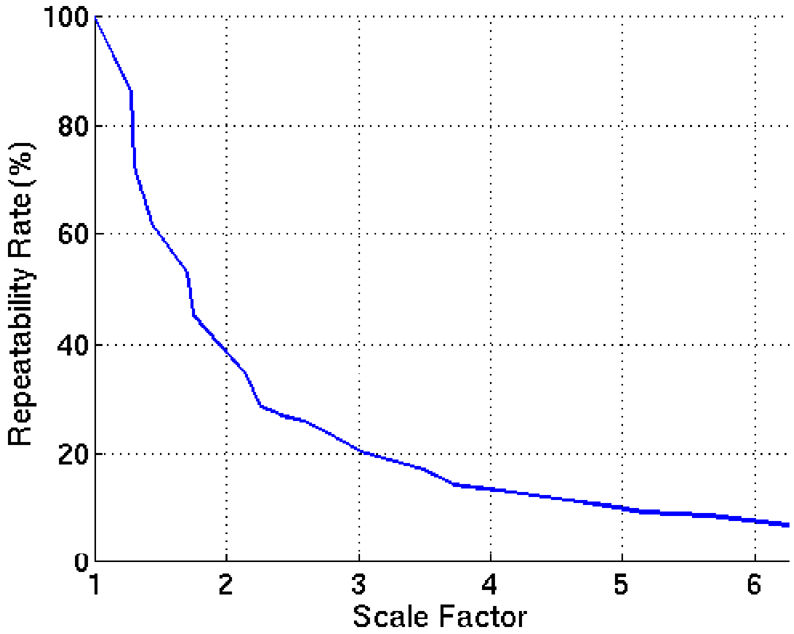
- Description regions have to be adapted to scale changes

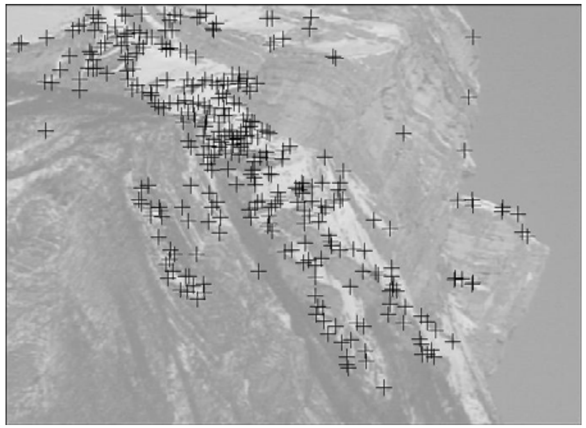


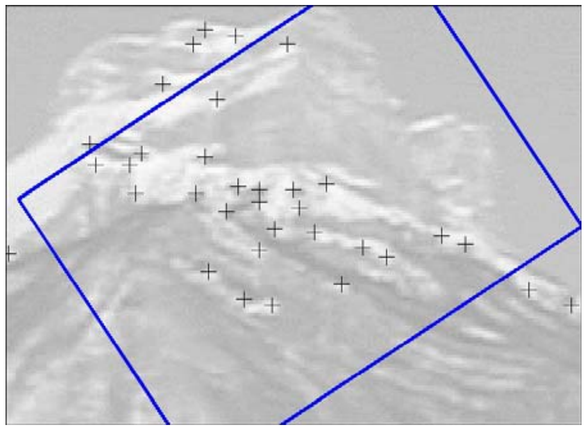


- Interest points have to be repeatable for scale changes

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

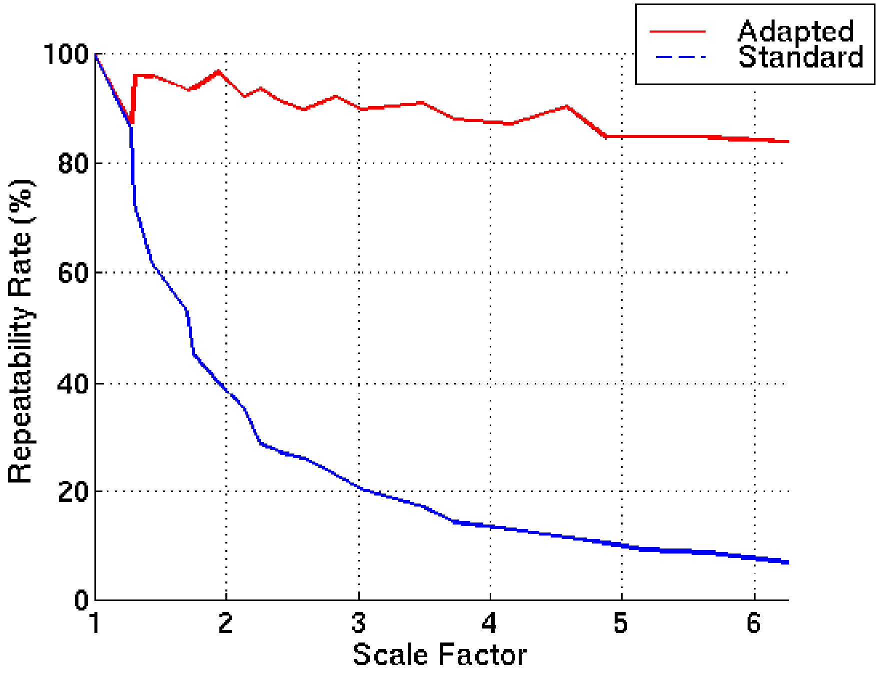
Scale adapted derivative calculation

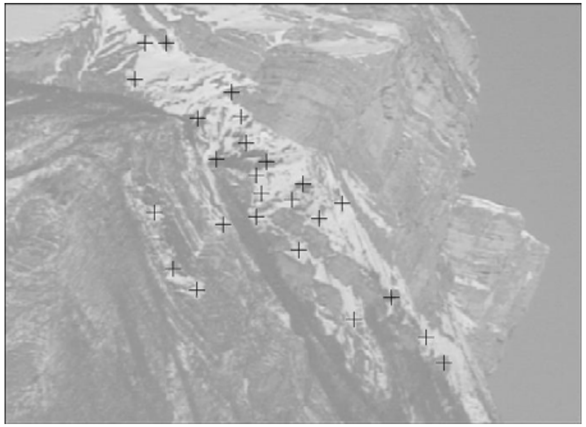
$$I_1\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \otimes G_{i_1 \ldots i_n}(\sigma) = s^n I_2\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \otimes G_{i_1 \ldots i_n}(s\sigma)$$

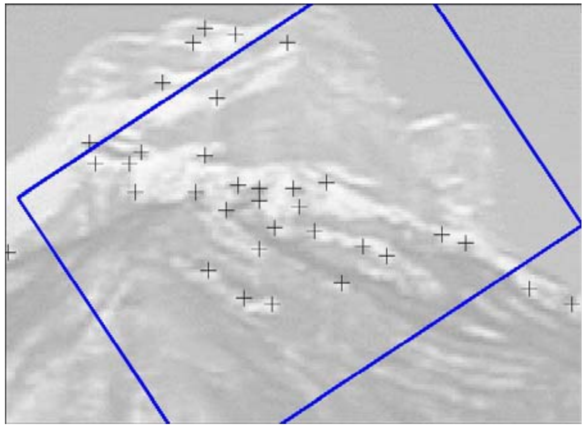
Scale adapted auto-correlation matrix

$$s^2 G(s\tilde{\sigma}) \otimes \begin{bmatrix} I_x^2(s\sigma) & I_x I_y(s\sigma) \\ I_x I_y(s\sigma) & I_y^2(s\sigma) \end{bmatrix}$$

Harris detector – adaptation to scale







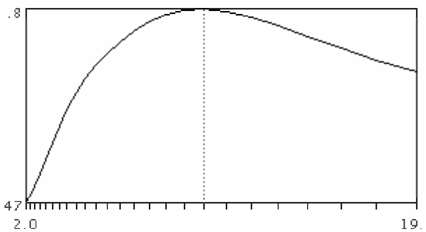
Scale selection

- 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 $\rightarrow$ characteristic scale

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

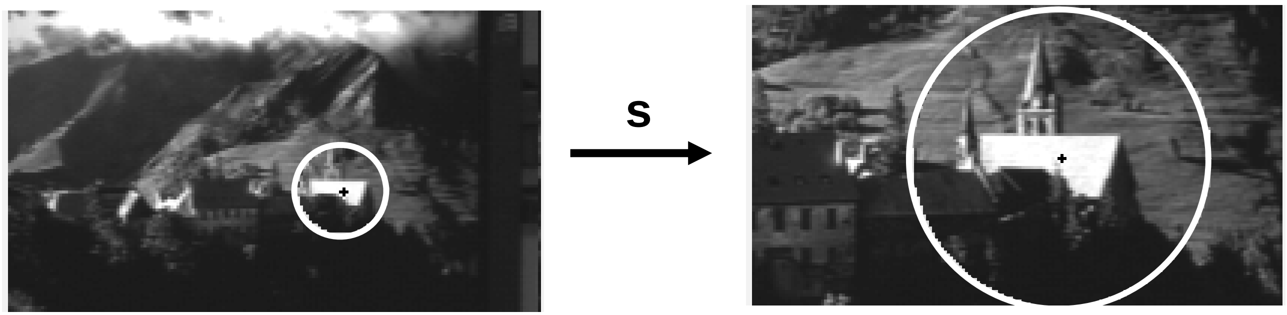


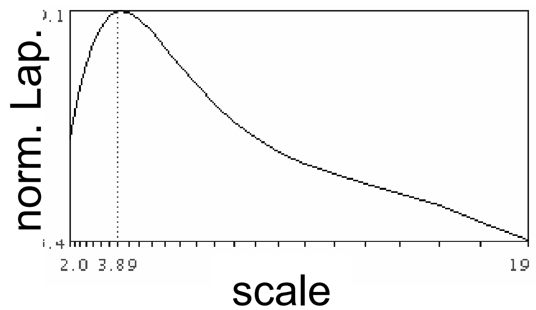
scale

- Exp. results show that the Laplacian gives best results

Scale selection

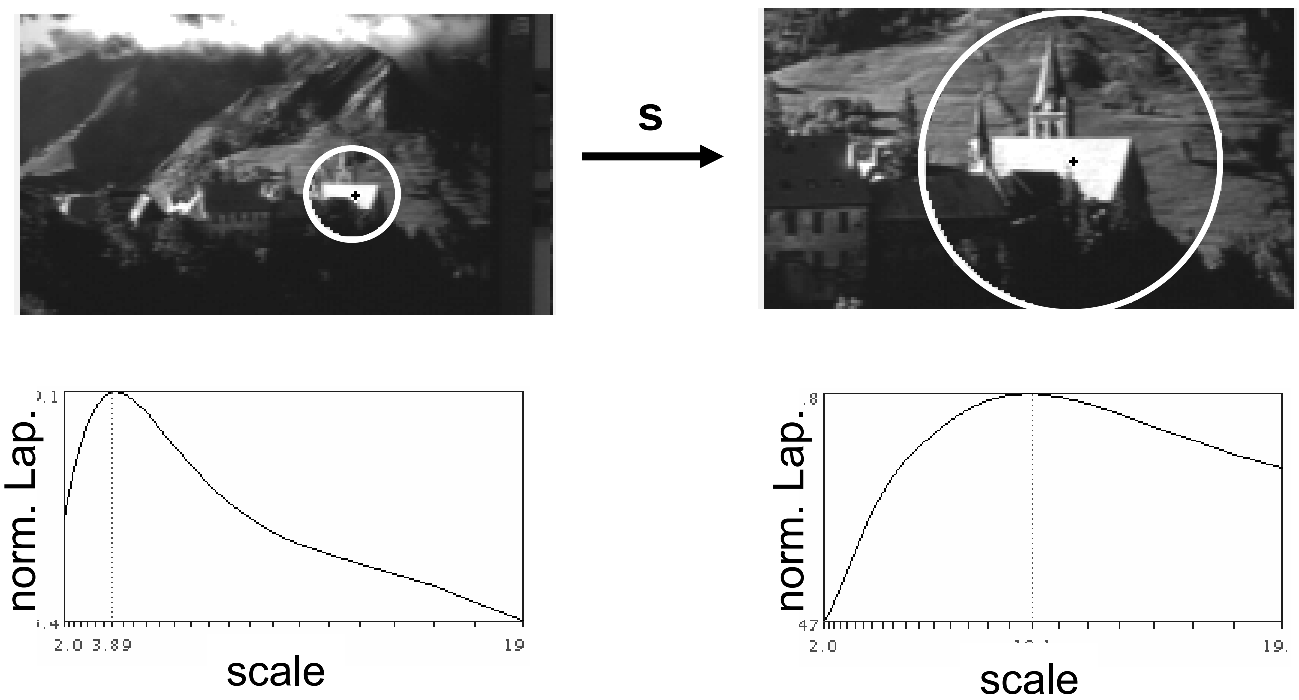
- Scale invariance of the characteristic scale





Scale selection

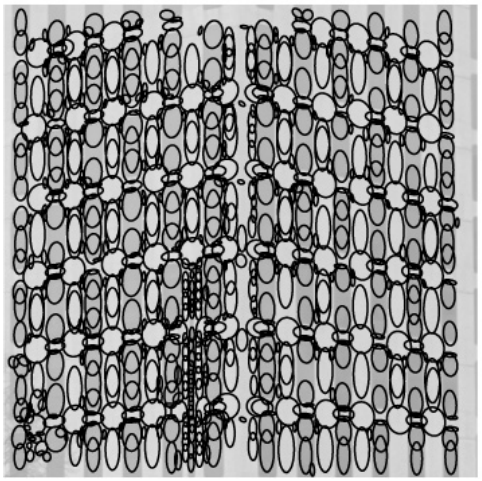
- Scale invariance of the characteristic scale



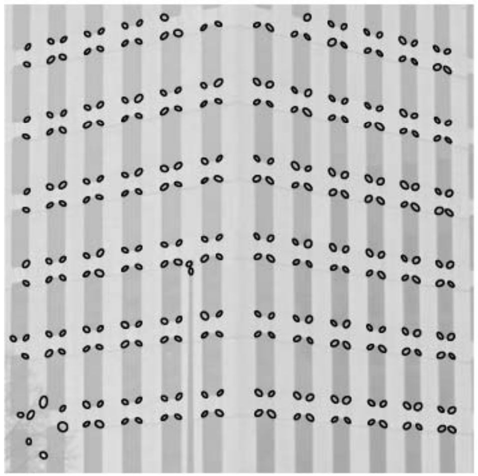
- Relation between characteristic scales $s \cdot s_1^* = s_2^*$

Scale-invariant detectors

- Laplacian detector (LOG) [Lindeberg'98]
- Difference of Gaussian, approximation of LOG [Lowe'99]
- Hessian detector & Harris-Laplace [Mikolajczyk & Schmid'04]
- SURF detector [Bay et al.'08]



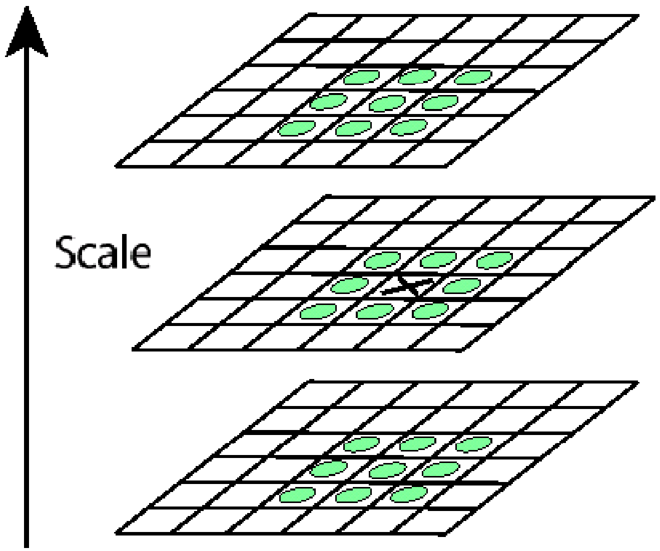
Laplacian



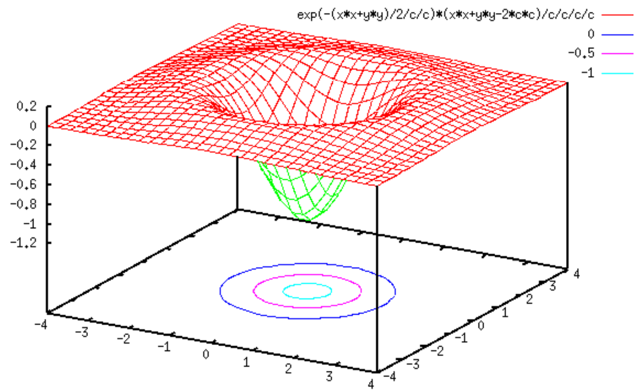
Harris-Laplace

LOG detector

Detection of maxima and minima of Laplacian in scale space



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



Hessian detector

Hessian matrix $H(x) = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix}$

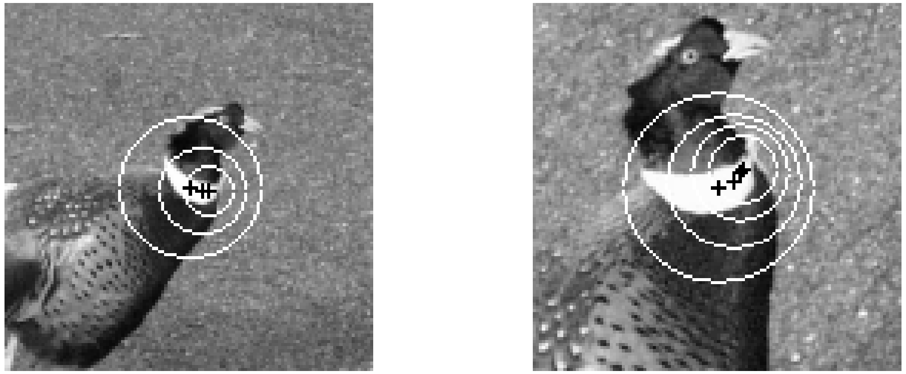
Determinant of Hessian matrix $DET = L_{xx} L_{yy} - L_{xy}{}^{2}$

Penalizes/eliminates long structures

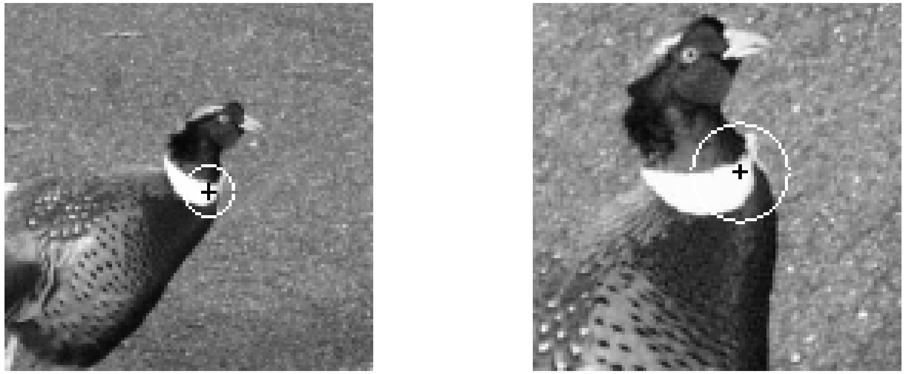
- with small derivative in a single direction

Harris-Laplace

multi-scale Harris points



selection of points at maximum of Laplacian



➡ invariant points + associated regions

Matching results

213 / 190 detected interest points

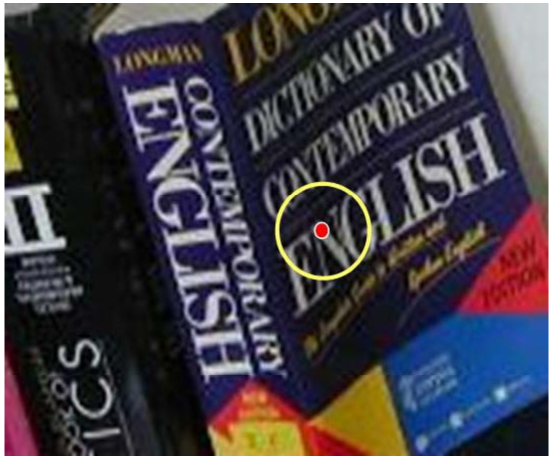
Matching results

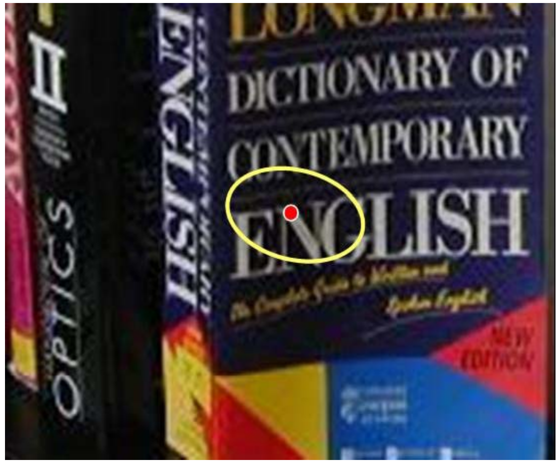
58 points are initially matched

Matching results

32 points are matched after verification – all correct

Affine invariant regions - Motivation





Scale invariance is not sufficient for large baseline changes

Affine invariant regions - Motivation

Example for wide baseline matching (22 correct matches)

Affine invariant regions - Motivation

Example for wide baseline matching (33 correct matches)

Harris/Hessian/Laplacian-Affine

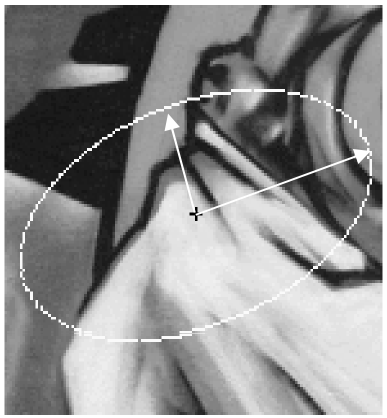
- Initialize with scale-invariant Harris/Hessian/Laplacian points
- Estimation of the affine neighbourhood with the second moment matrix [Lindeberg’94]
- Apply affine neighbourhood estimation to the scale-invariant interest points [Mikolajczyk & Schmid’02, Schaffalitzky & Zisserman’02]
- Excellent results in a comparison [Mikolajczyk et al.’05]

Affine invariant regions

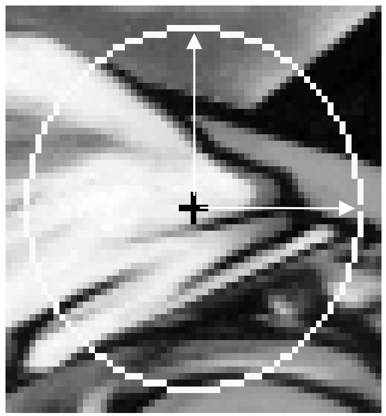
- Based on the second moment matrix (Lindeberg'94)

$$M = \mu(\mathbf{x}, \sigma_I, \sigma_D) = \sigma_D^{\ 2} G(\sigma_I) \otimes \begin{bmatrix} I_x^{\ 2}(\mathbf{x}, \sigma_D) & I_x I_y(\mathbf{x}, \sigma_D) \\ I_x I_y(\mathbf{x}, \sigma_D) & I_y^{\ 2}(\mathbf{x}, \sigma_D) \end{bmatrix}$$

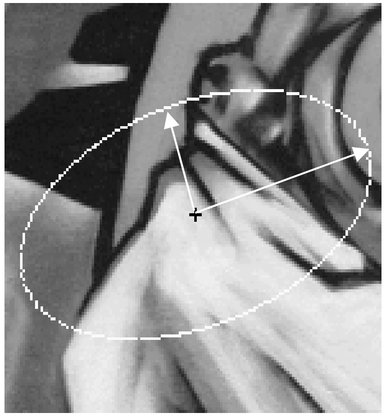
- Normalization with eigenvalues/eigenvectors



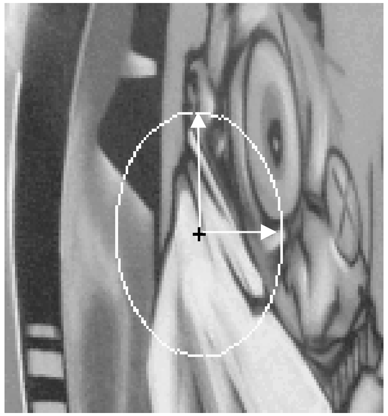
$$\mathrm{x}' = M^{\frac{1}{2}}\mathrm{x}$$



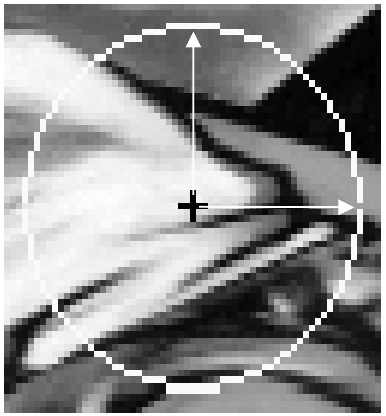
Affine invariant regions



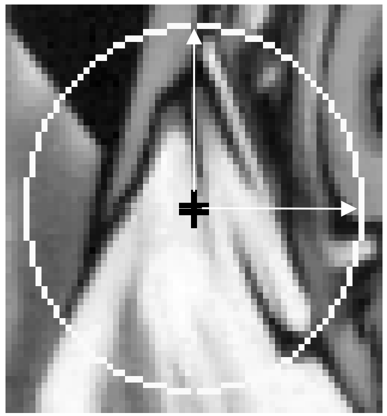
$$x_R = A x_L$$



$$x_L' = M_L^{\frac{1}{2}} x_L \qquad x_R' = M_R^{\frac{1}{2}} x_R$$

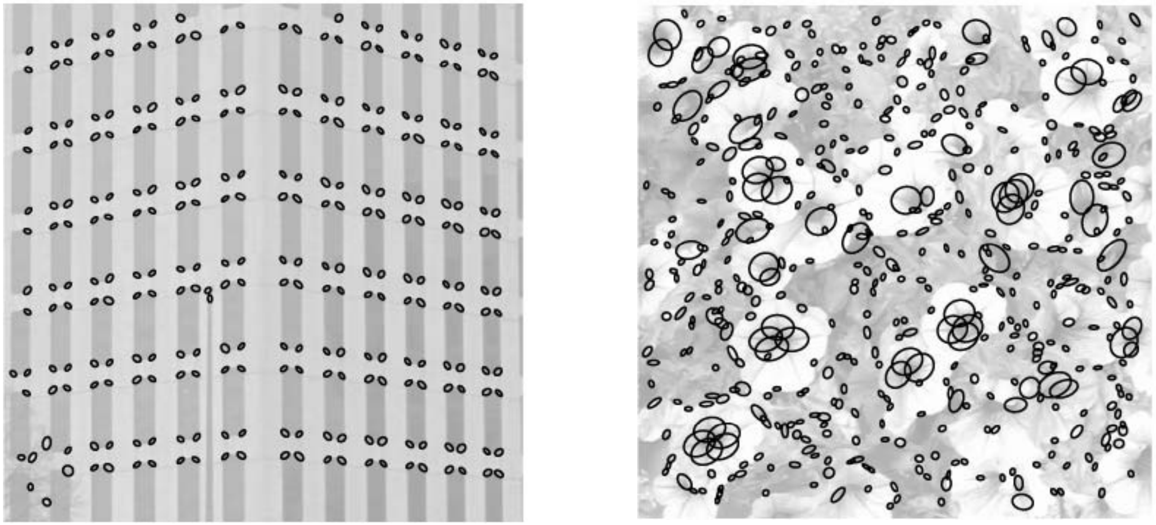


$$x_R' = R x_L'$$

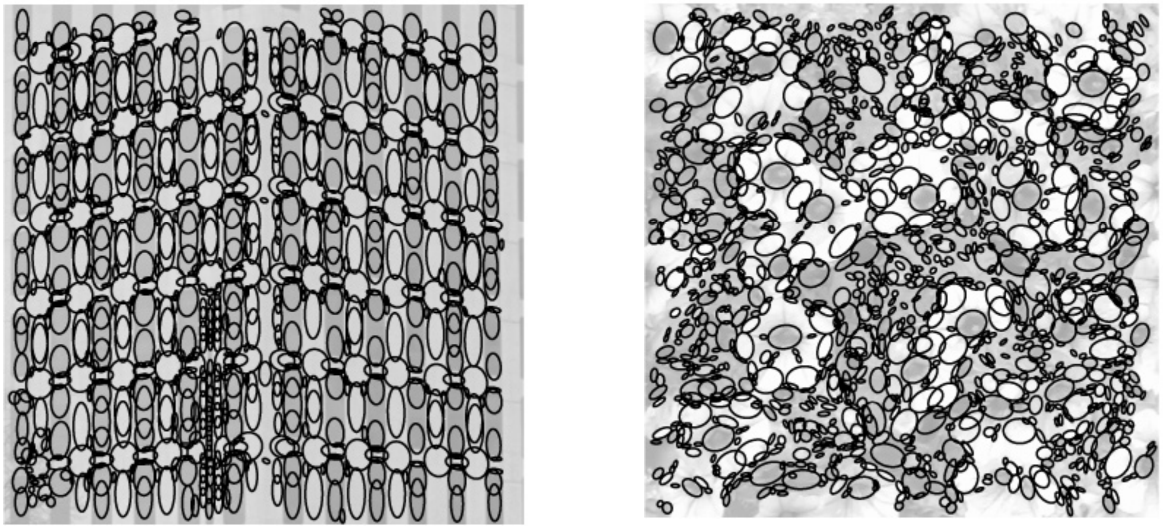


Isotropic neighborhoods related by image rotation

Harris/Hessian-Affine

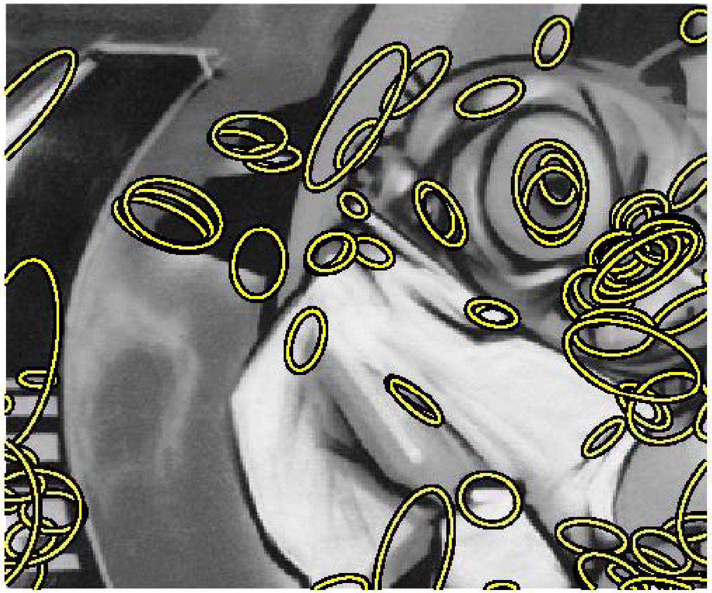


Harris-Affine



Hessian-Affine

Harris-Affine



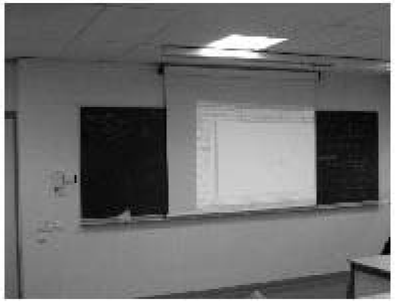
Hessian-Affine

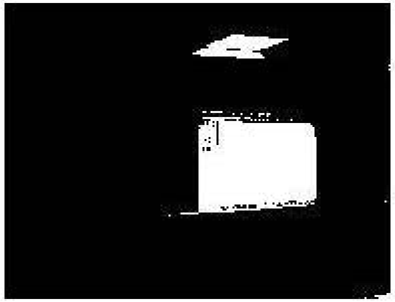
Maximally stable extremal regions (MSER) [Matas'02]

- 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 comparison [Mikolajczyk et al.'05]

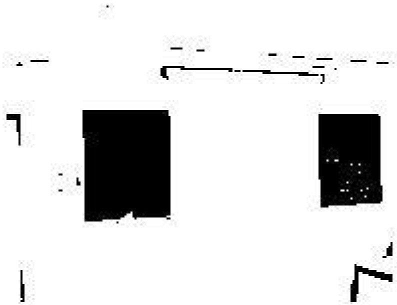
Maximally stable extremal regions (MSER)

Examples of thresholded images



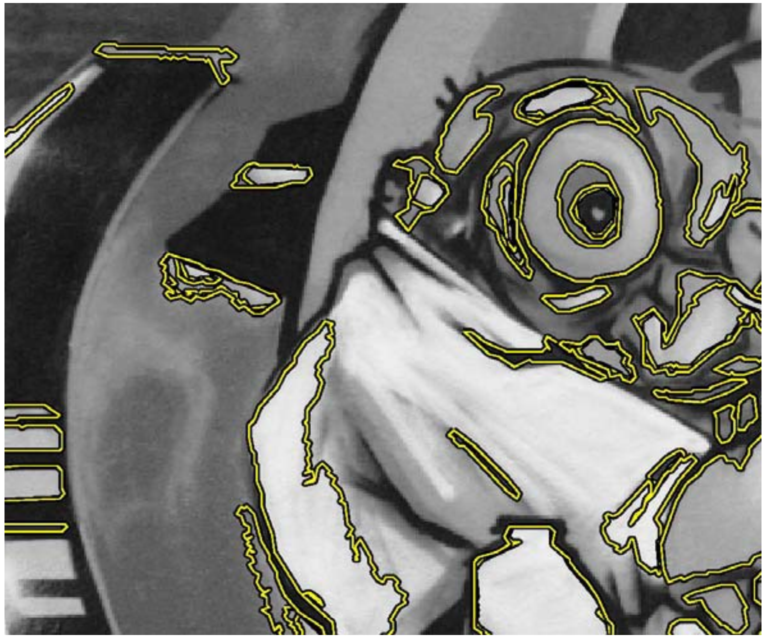


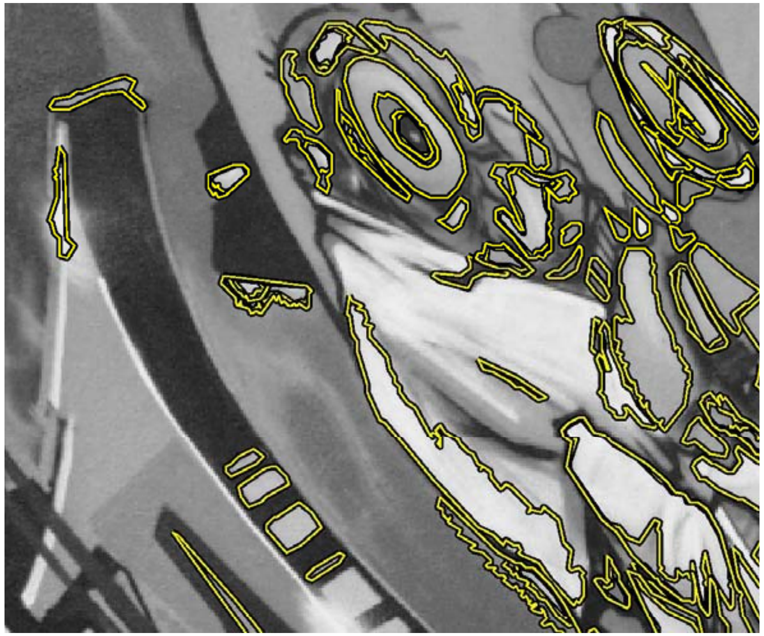
high threshold



low threshold

MSER





Conclusion – detectors [Mikolajczyk & al. ’05]

- Good performance for large viewpoint and scale changes
- Results depend on transformation and scene type, no one best detector
- Detectors are complementary
  - MSER adapted to structured scenes
  - Harris and Hessian adapted to textured scenes
- Performance of the different scale invariant detectors is very similar (Harris-Laplace, Hessian and LoG)
- Scale-invariant detector sufficient up to 40 degrees of viewpoint change

Conclusion

- Excellent performance for wide baseline matching
- Binaries for detectors and descriptors on-line available
  - for example at http://lear.inrialpes.fr/software
- On-line available evaluation setup
  - Dataset with transformations
  - Evaluation code in matlab
  - Benchmark for new detectors and descriptors