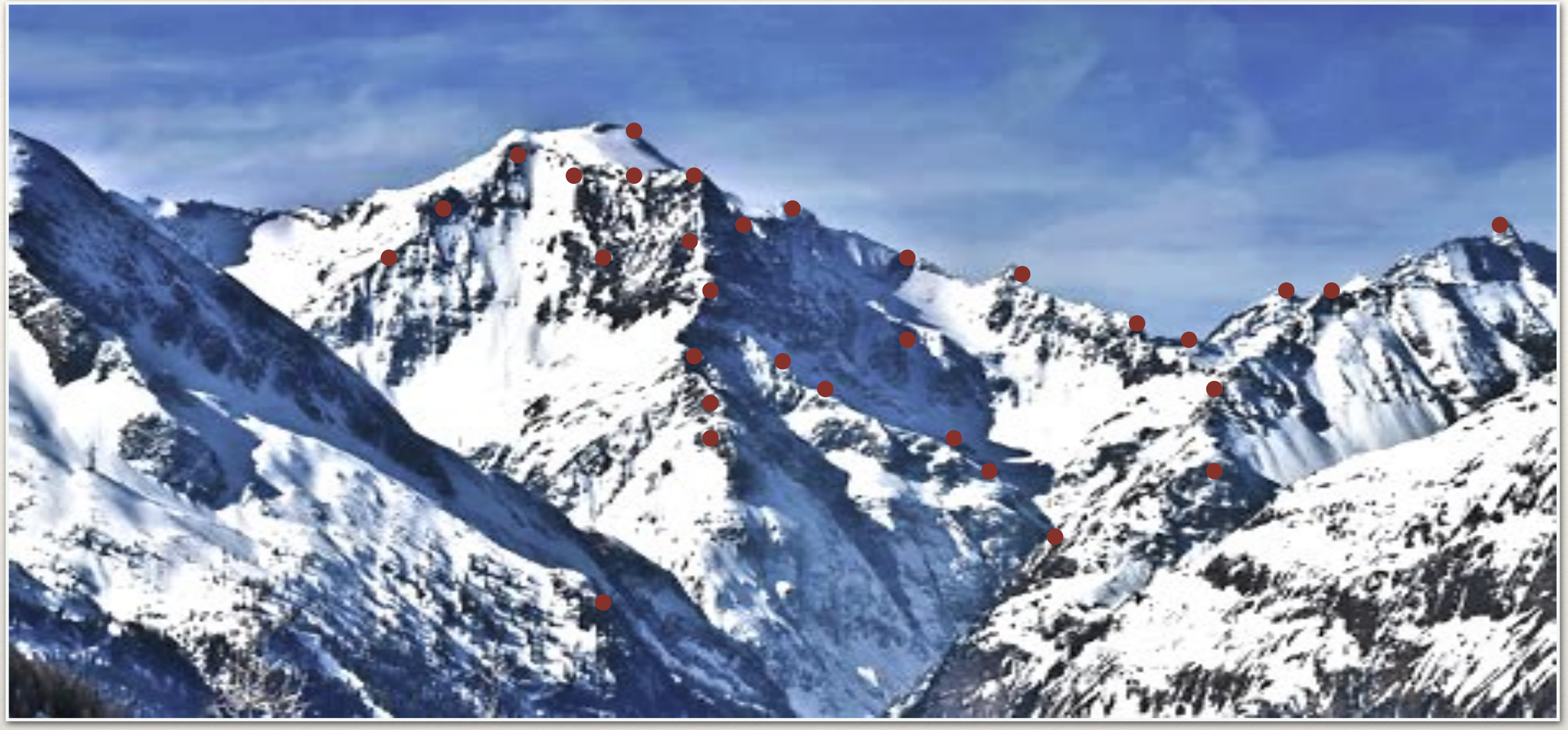


Xu Zhao @ Shanghai Jiao Tong university

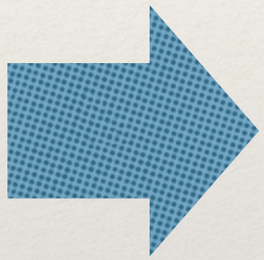
Lecture 3-3: Local Image Features - Part 2

Contents

- **Local feature description**
  - SIFT descriptor
- **Feature matching**
  - Distance based matching
- **Other descriptors**
  - GLOH
  - Shape context
  - HOG
  - Binary features

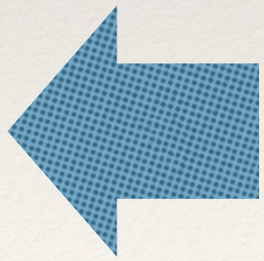
Local image features

- Local, meaningful, detectable parts of an image
- Edge, corner, line, contour, texture, …



Feature detection

Feature description

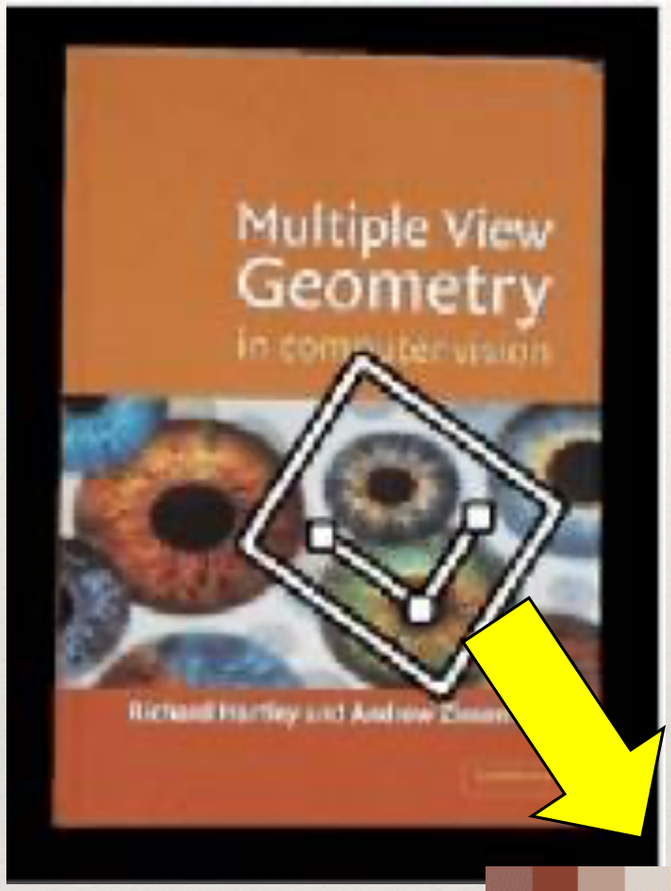


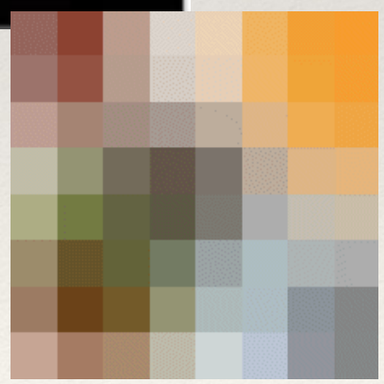
Matching
Indexing
Detection

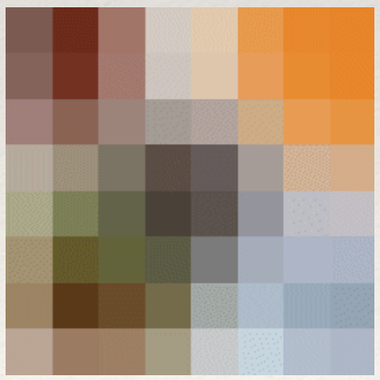
Characteristics of good features

- **Repeatability**
  - The same feature can be found in several images despite geometric and photometric transformations
- **Saliency**
  - Each feature is distinctive
- **Compactness and efficiency**
  - Many fewer features than image pixels
- **Locality**
  - A feature occupies a relatively small area of the image; robust to clutter and occlusion

Geometric transformations







e.g. scale, translation, rotation

Photometric transformations

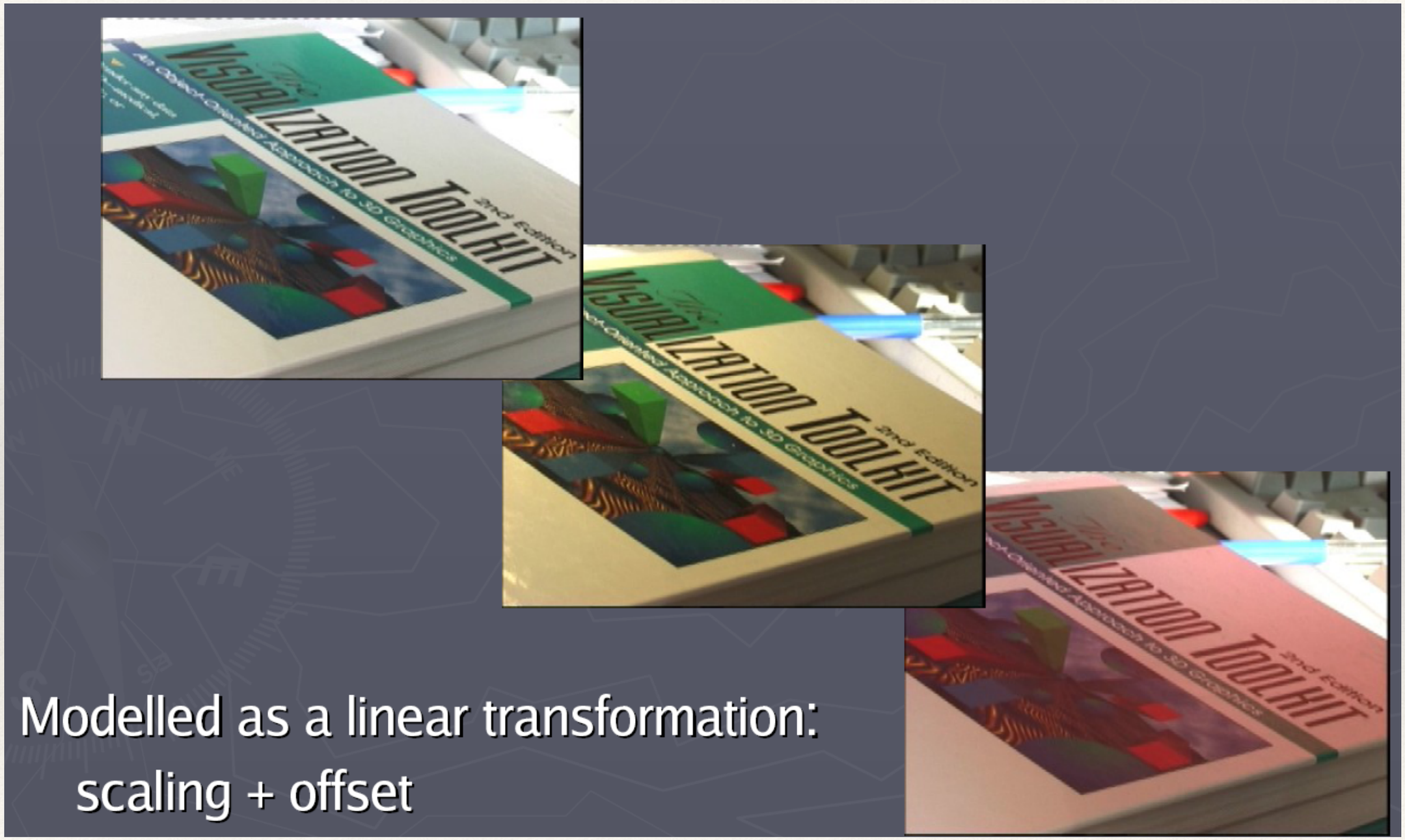
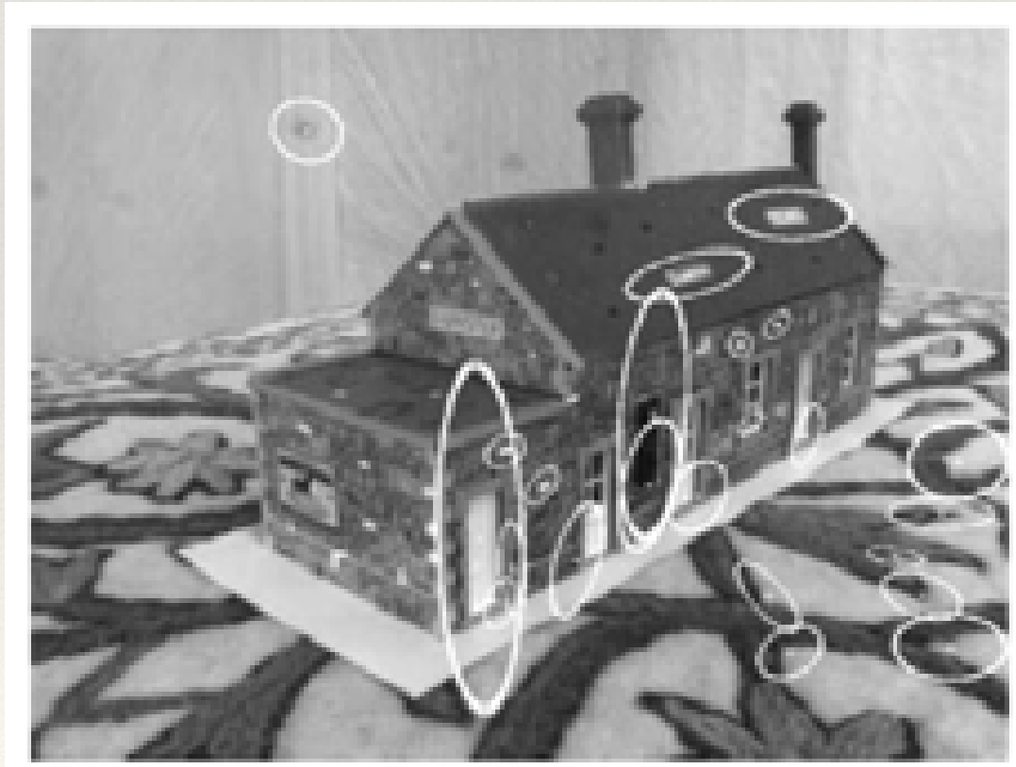
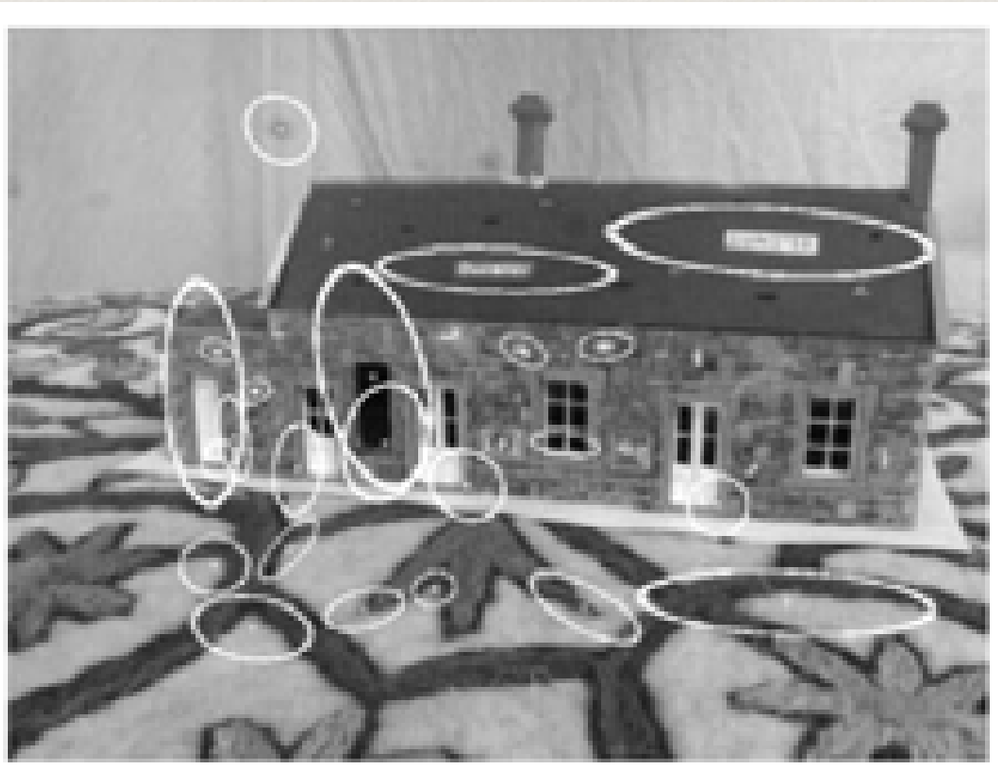


Figure from T. Tuytelaars ECCV 2006 tutorial

Affine invariance

- Detectors respond consistently across affine deformations
- Applications: wide baseline stereo matching and location recognition

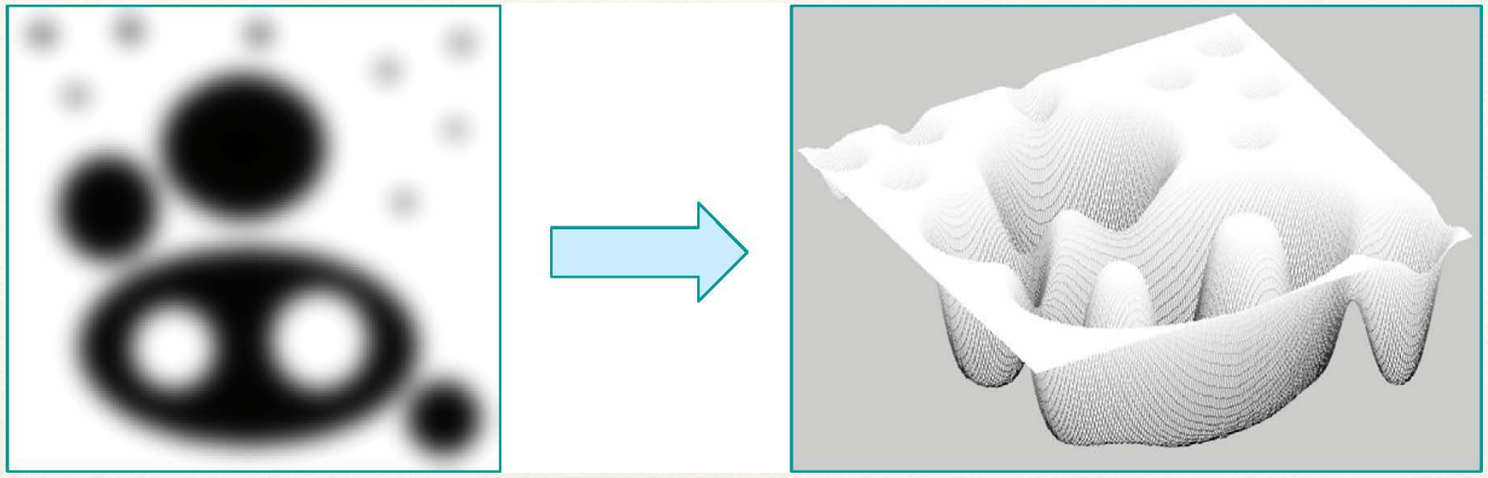




MSER: **M**aximally **S**table **E**xtremal **R**egion

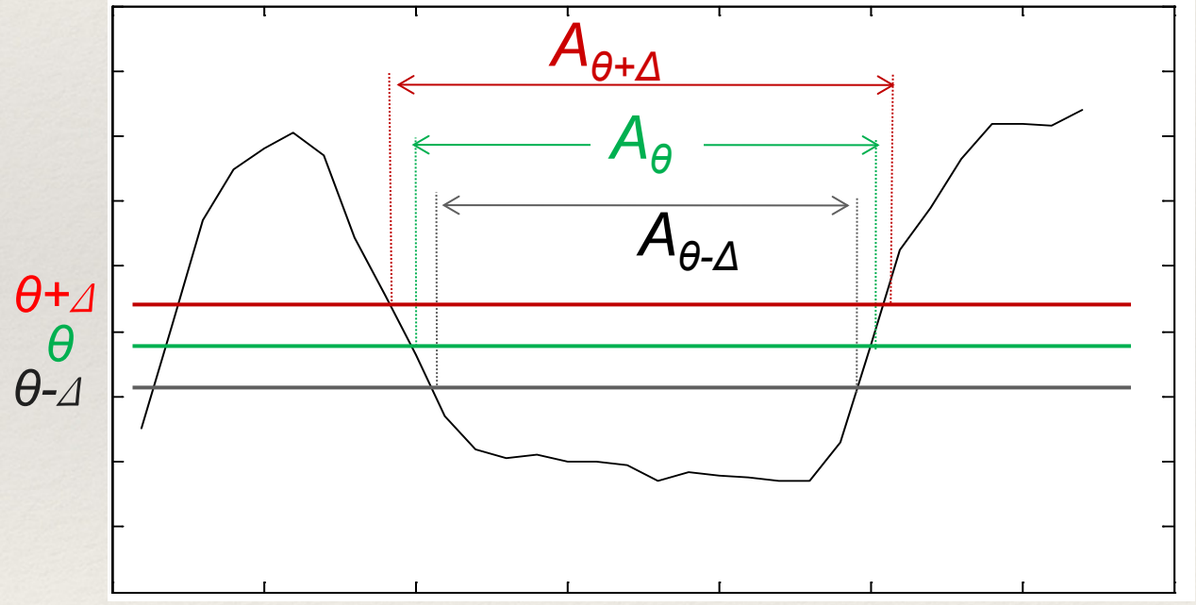
- Binary region extraction by thresholding the image at all possible gray levels
- As the threshold is changed, the area of each connected component is monitored: regions whose rate of change of area with respect to the threshold is minimal are defined as maximally stable

Matas, Chum, Urban *et al.* (2002)



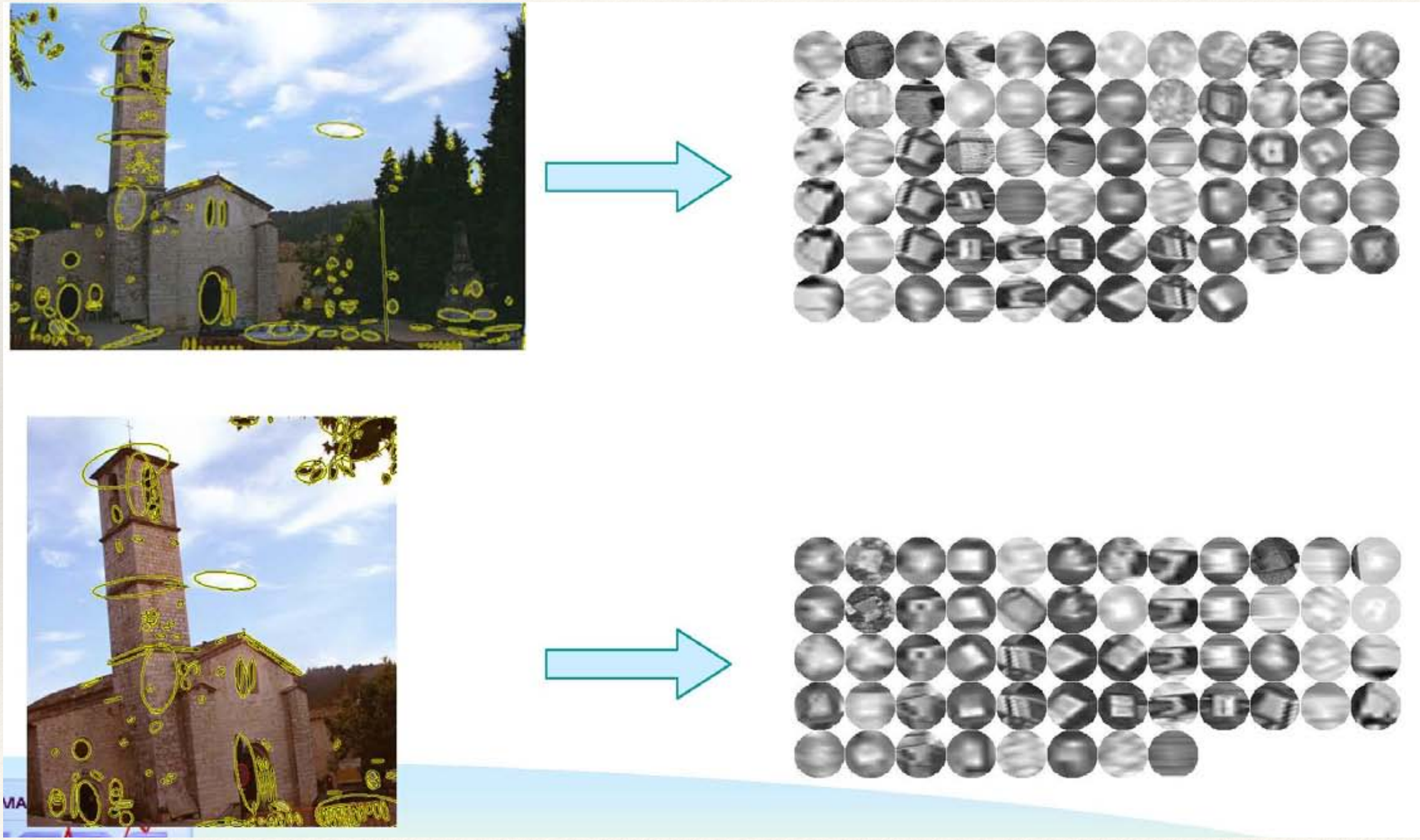
intensity image

shown as a surface function



Local minimum of $\left|\frac{A_{\theta-\Delta} - A_{\theta+\Delta}}{A_{\theta}}\right| \rightarrow$ MSER

MSER example

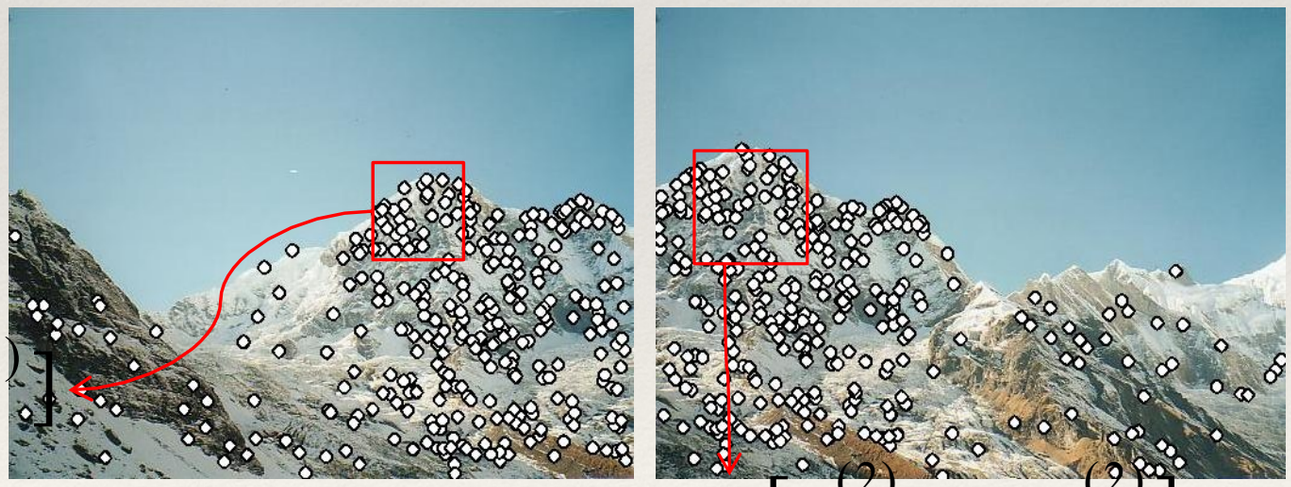


Interest point detection

- Harris corner detector
- Laplacian of Gaussian
  - Automatic scale selection
- Difference of Gaussian
- MSER - affine invariance

Local features: main components

- **Detection**: Find a set of distinctive key points.
- **Description**: **Extract feature descriptor around each interest point as vector.**
- **Matching**: Compute distance between feature vectors to find correspondence.



$$\mathbf{x}_1 = [x_1^{(1)}, \ldots, x_d^{(1)}]$$

$$\mathbf{x}_2 = [x_1^{(2)}, \ldots, x_d^{(2)}]$$

$$d(\mathrm{x}_1, \mathrm{x}_2) < T$$

K. Grauman, B. Leibe

Local feature description

- Raw patches as local descriptors
- The simplest way to describe the neighborhood around an interest point is to write down the list of intensities to form a feature vector.
- But this is very sensitive to even small shifts, rotations.

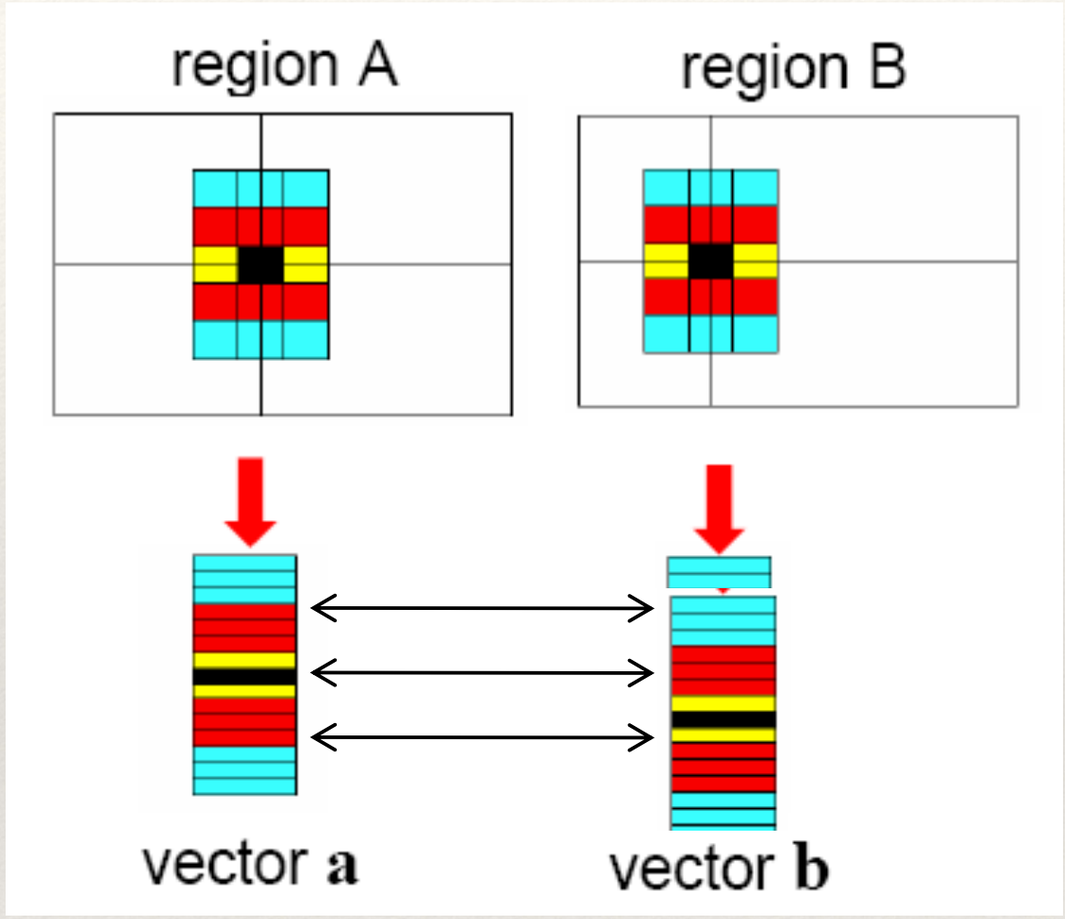
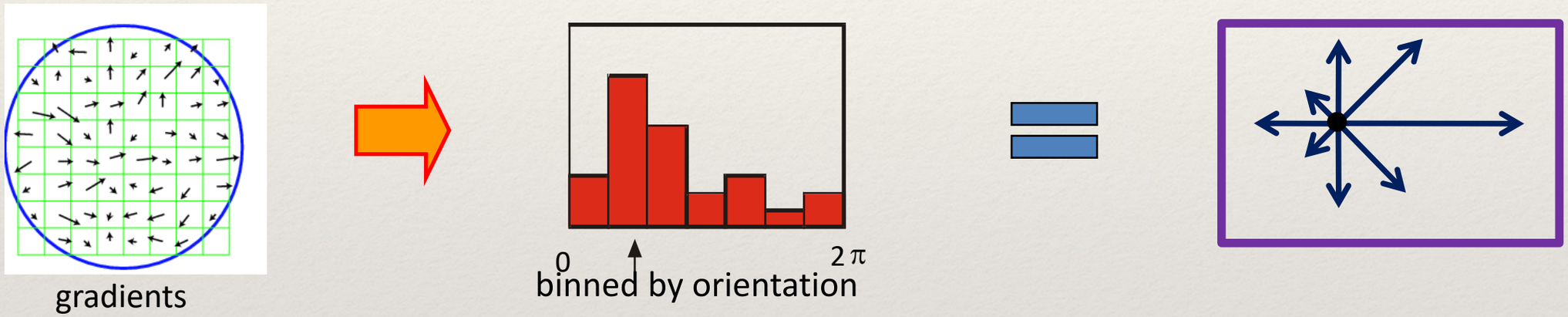
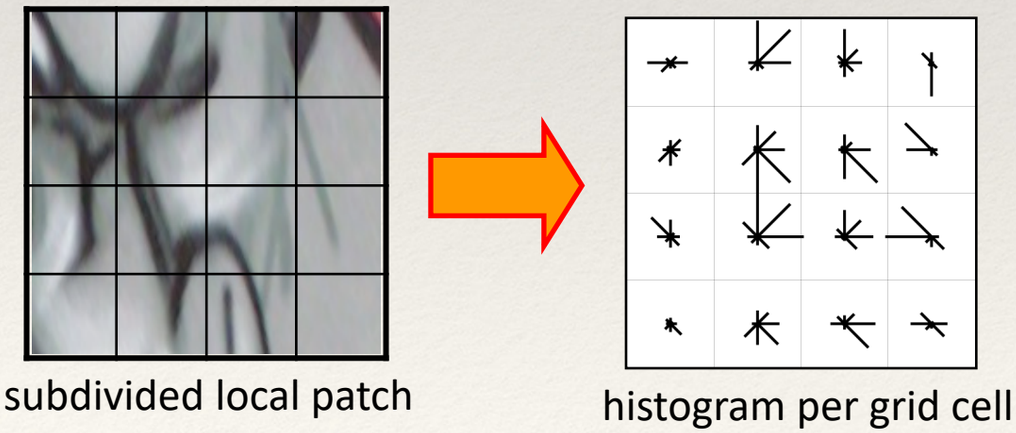


Figure: Andrew Zisserman

Scale Invariant Feature Transform (SIFT) descriptor

- Use histograms to bin pixels within sub-patches according to their orientation.

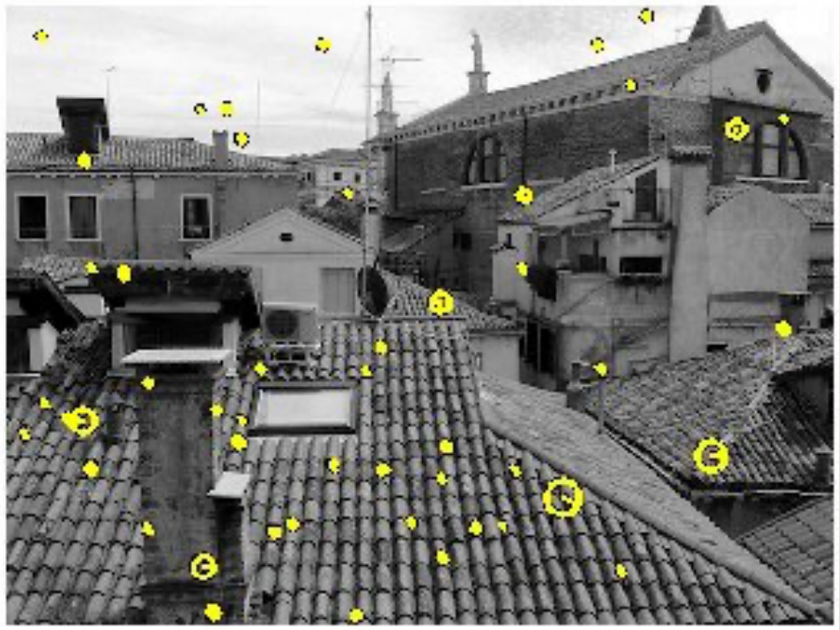


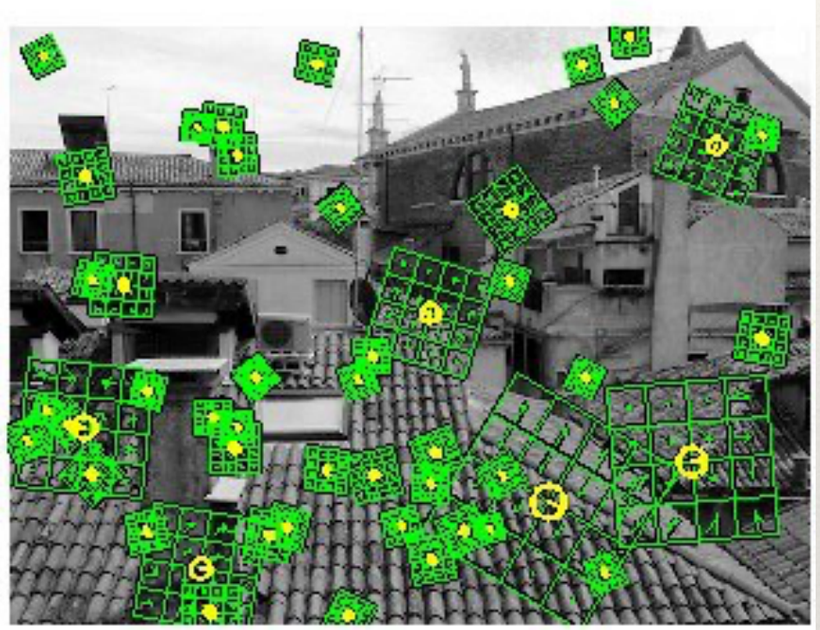


Final descriptor = concatenation of all histograms, normalize

Kristen Grauman

Scale Invariant Feature Transform (SIFT) descriptor





Interest points and their scales
and orientations
(random subset of 50)

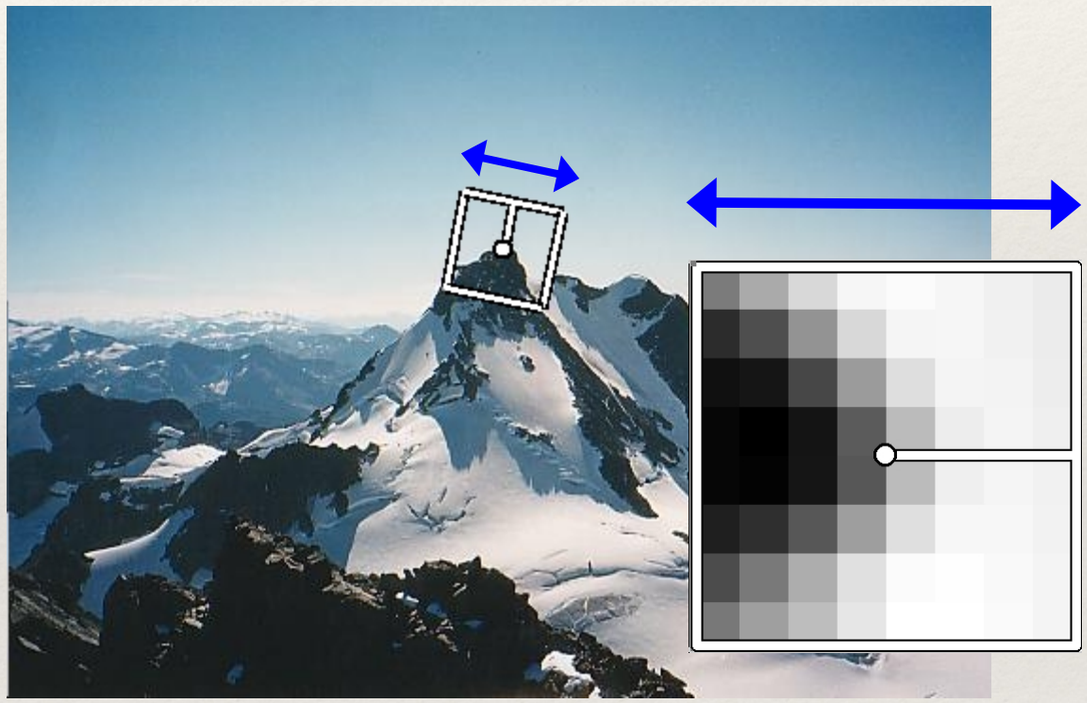
http://www.vlfeat.org/overview/sift.html

Procedures

- Find Difference of Gaussian scale-space extrema as feature point locations
- Post-processing
  - Subpixel position interpolation
  - Discard low-contrast points
  - Eliminate points along edges
- Orientation estimation per feature point

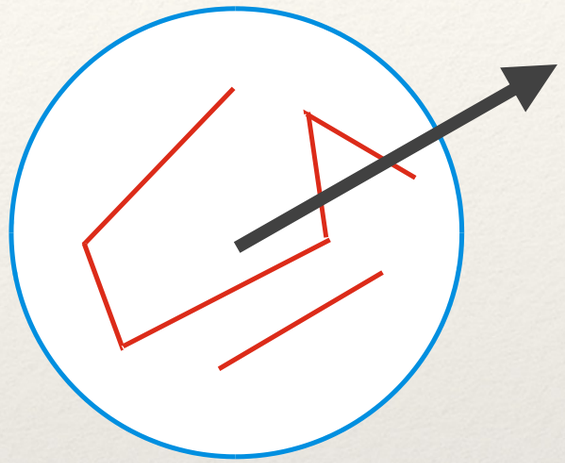
Rotational invariant

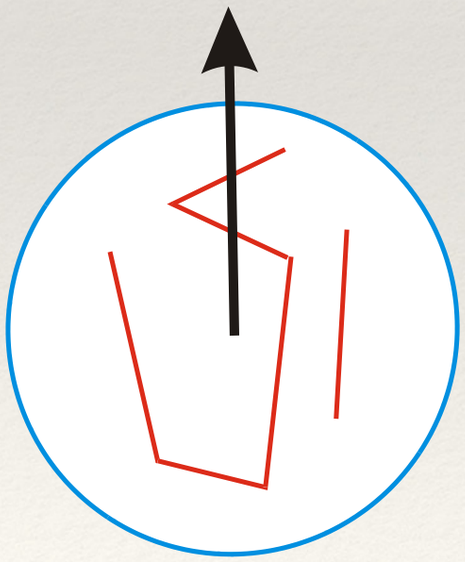
- Rotate patch according to its dominant gradient orientation
- This puts the patches into a canonical orientation.

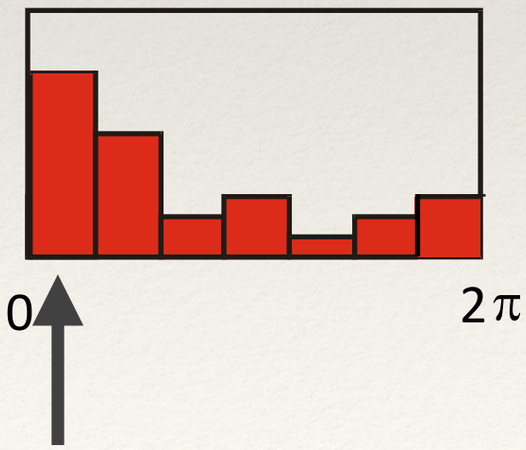


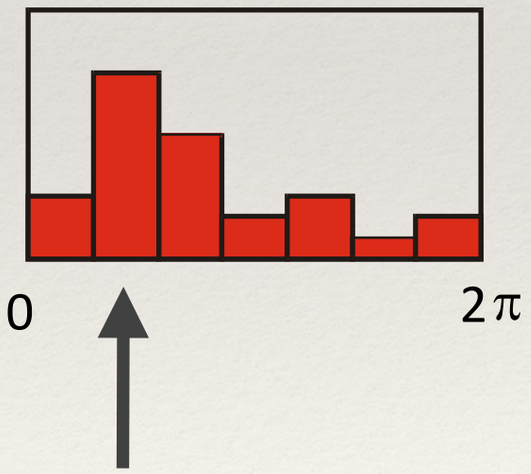
SIFT Orientation estimation and normalization

- Compute gradient orientation histogram
- Select dominant orientation θ
- Normalize: rotate to fixed orientation









Procedures

- Find Difference of Gaussian scale-space extrema as feature point locations
- Post-processing
  - Subpixel position interpolation
  - Discard low-contrast points
  - Eliminate points along edges
- Orientation estimation per feature point
- Descriptor extraction
  - Motivation: We want some sensitivity to spatial layout, but not too much, so blocks of histograms give us that.

SIFT descriptor extraction

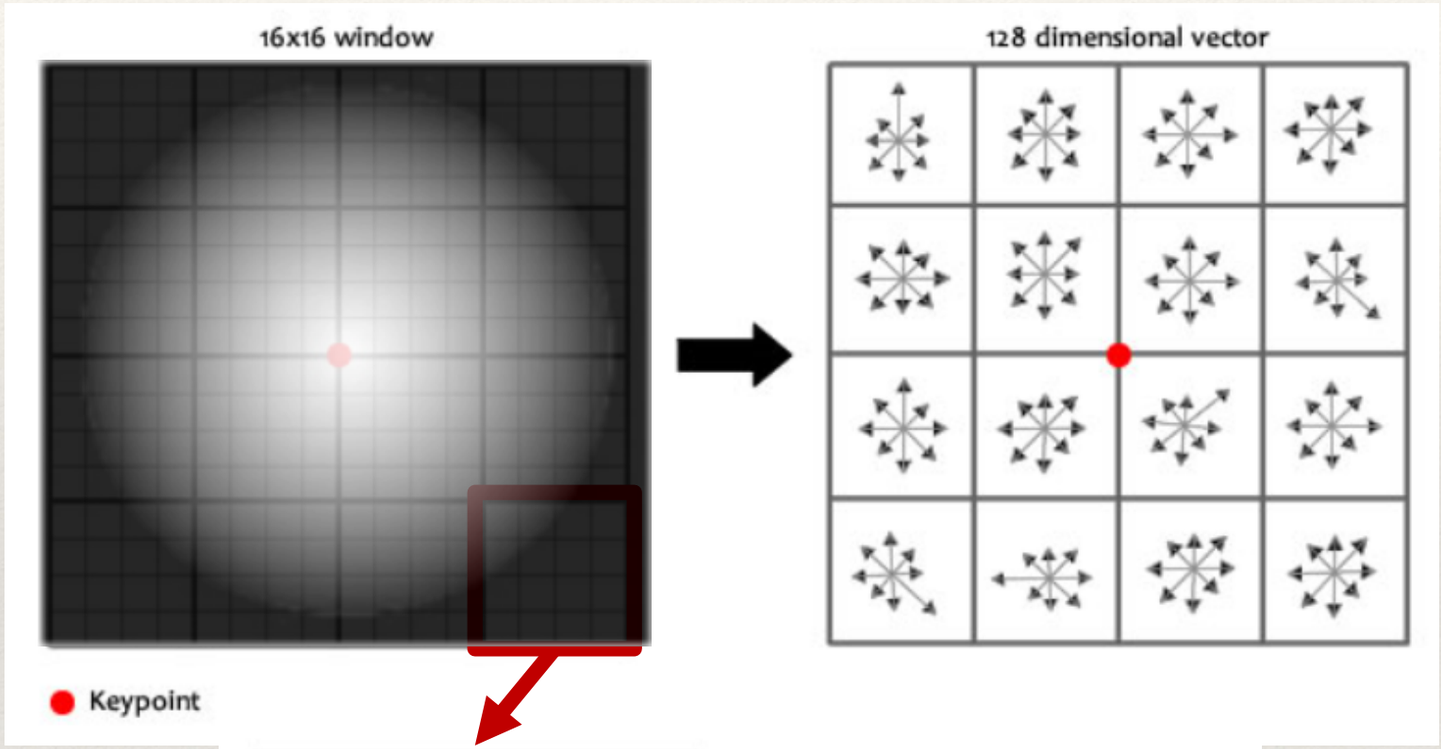
- Given a keypoint with scale and orientation:
  - Pick scale-space image which most closely matches estimated scale
  - Resample image to match orientation OR
  - Subtract detector orientation from vector to give invariance to general image rotation.

SIFT descriptor formation

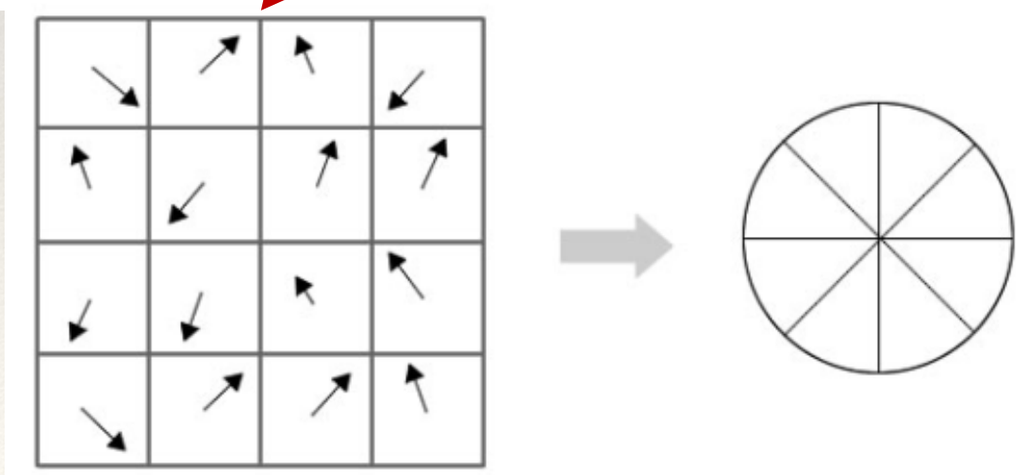
- Given a keypoint with scale and orientation

σ= half window width

Weight 16x16 grid by Gaussian to add location robustness and reduce effect of outer regions



Gradient magnitude and orientation



8 bin ‘histogram’ - add magnitude amounts!

Utkarsh Sinha

SIFT descriptor formation

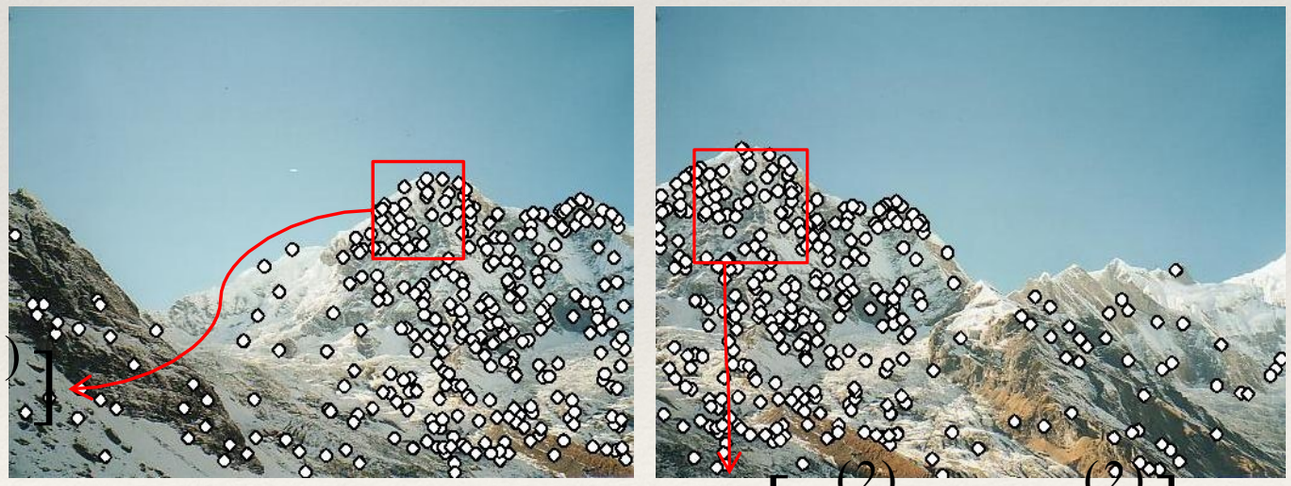
- Extract 8 x 16 values into 128-dim vector
- Illumination invariance:
  - Working in gradient space, so robust to $I = I + b$
  - Normalize vector to [0…1]
    - Robust to $I = \alpha I$ brightness changes
- Clamp all vector values > 0.2 to 0.2.
  - Robust to “non-linear illumination effects”
  - Image value saturation / specular highlights
- Renormalize

SIFT properties

- Invariant to
  - Scale
  - Rotation
- Partially invariant to
  - Illumination changes
  - Camera viewpoint
  - Occlusion, clutter

Local features: main components

- **Detection**: Find a set of distinctive key points.
- **Description**: Extract feature descriptor around each interest point as vector.
- **Matching: Compute distance between feature vectors to find correspondence.**



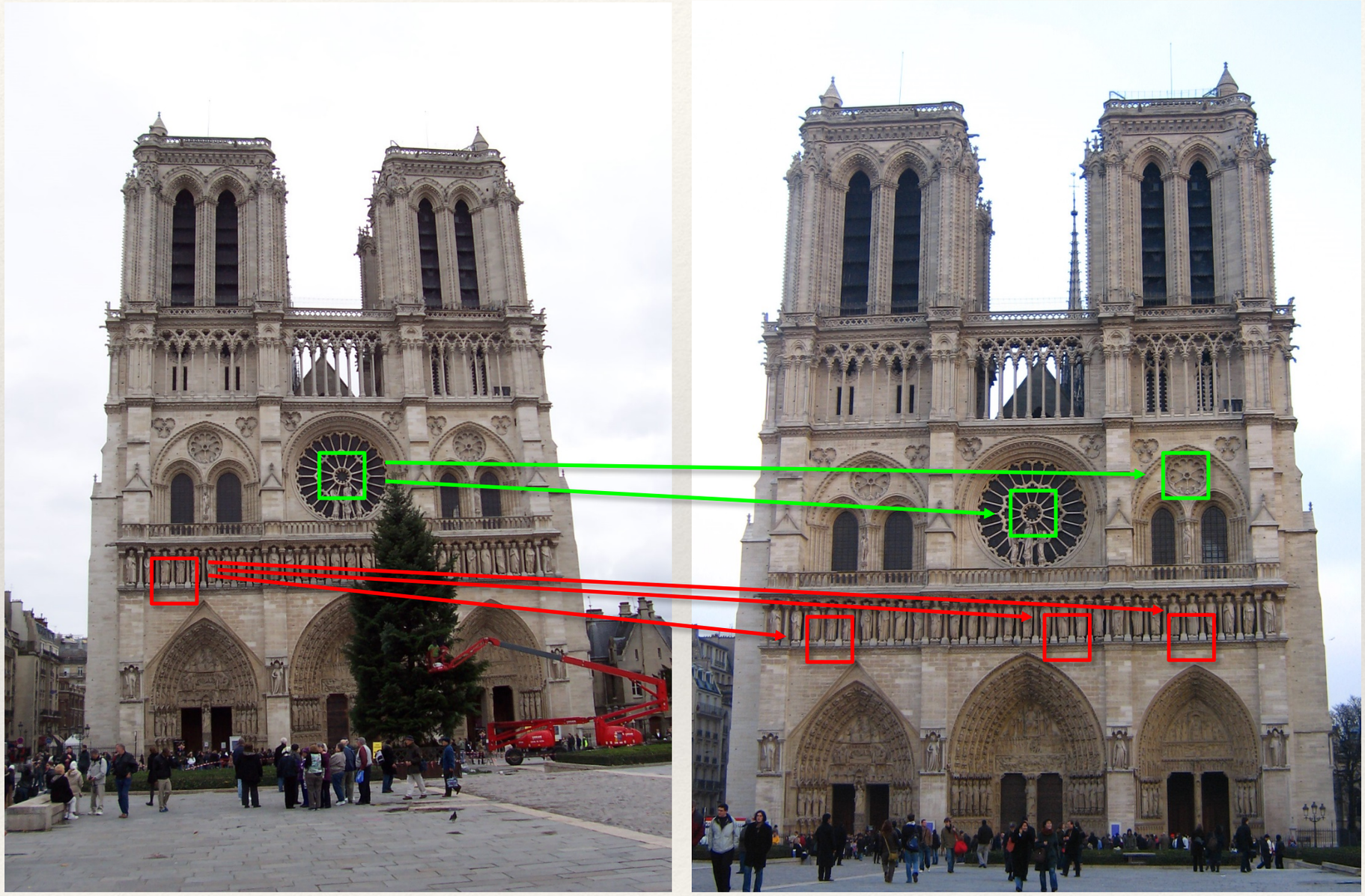
$$\mathbf{x}_1 = [x_1^{(1)}, \ldots, x_d^{(1)}]$$

$$\mathbf{x}_2 = [x_1^{(2)}, \ldots, x_d^{(2)}]$$

$$d(\mathbf{x}_1, \mathbf{x}_2) < T$$

K. Grauman, B. Leibe

How do we decide which features match?



Distance: 0.34, 0.30, 0.40

Distance: 0.61, 1.22

Euclidean distance vs. Cosine similarity

- Euclidean distance:

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \mathrm{d}(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

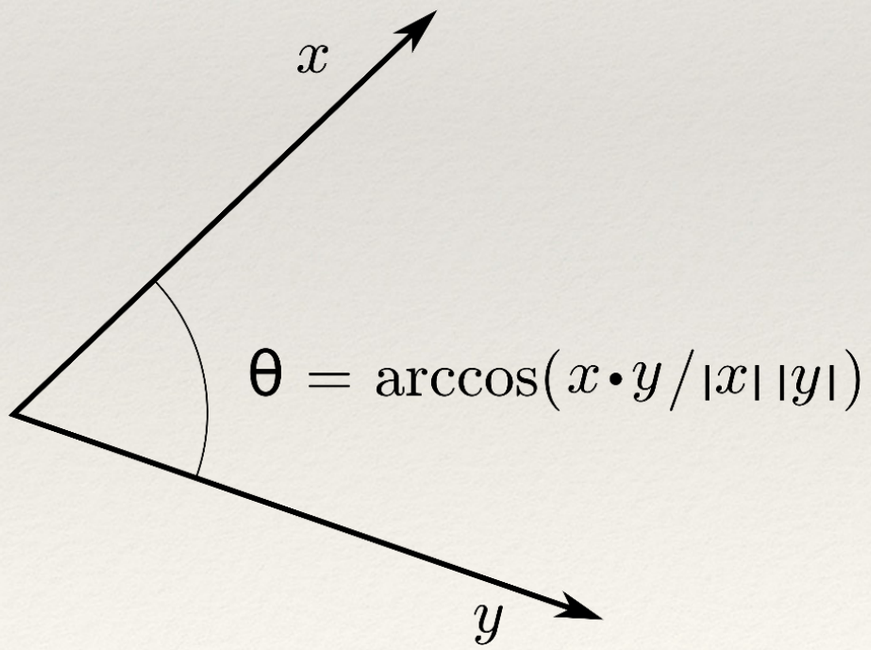
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

$$\|\mathbf{q} - \mathbf{p}\| = \sqrt{(\mathbf{q} - \mathbf{p}) \cdot (\mathbf{q} - \mathbf{p})}.$$

- Cosine similarity:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2 \cos\theta$$

$$\mathbf{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2}$$



Wikipedia

Feature matching

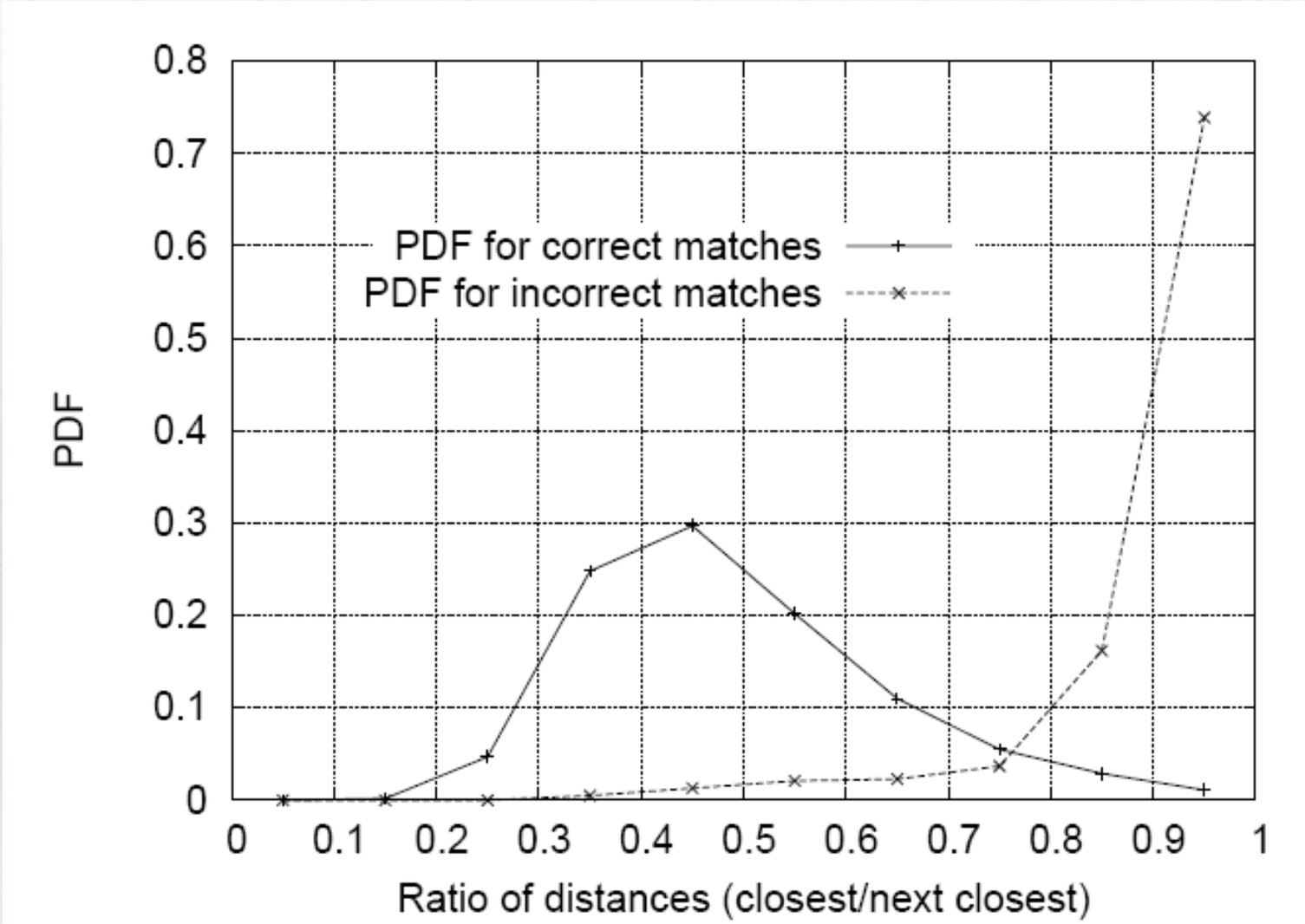
- Criteria:
  - Compute distance in feature space, e.g., Euclidean distance between 128-dim SIFT descriptors
  - Match point to lowest distance (nearest neighbor)
  - Ignore anything higher than threshold
- Problems:
  - Threshold is hard to pick
  - Non-distinctive features could have lots of close matches, only one of which is correct

Nearest neighbor distance ratio

- Compare distance of closest (NN1) and second-closest (NN2) feature vector neighbor.
  - If NN1 ≈ NN2, ratio $\mathbf{NN1/NN2}$ will be ≈ 1 -> matches too close.
  - As NN1 << NN2, ratio $\mathbf{NN1/NN2}$ tends to 0.
- Sorting by this ratio puts matches in order of confidence. Threshold ratio – but how to choose?

Matching SIFT descriptors

- Nearest neighbor (Euclidean distance)
- Threshold ratio of nearest to 2nd nearest descriptor
- 40,000 keypoints with hand-labeled ground truth

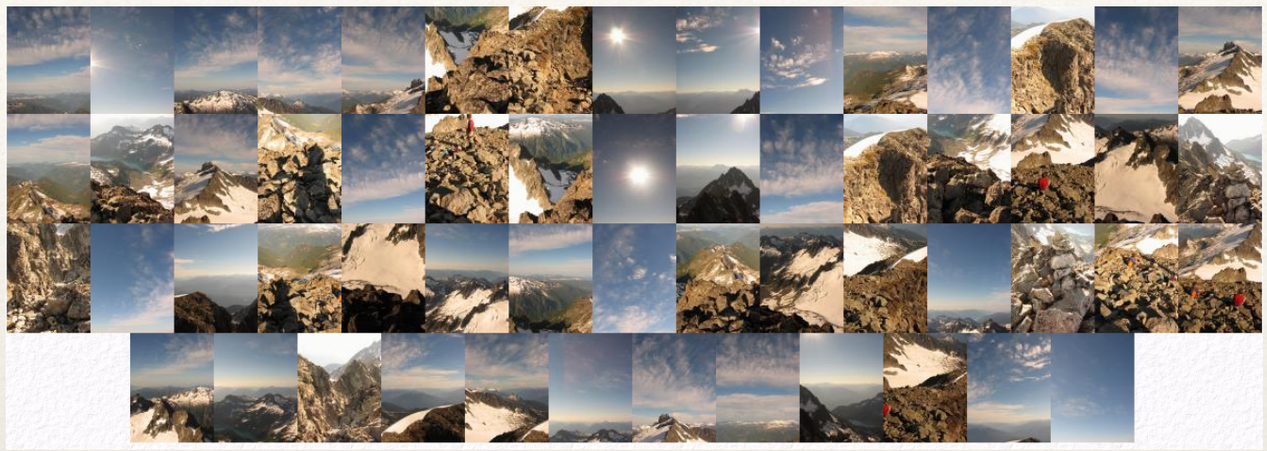


Lowe IJCV 2004

Value of local (invariant) features

- Complexity reduction via selection of distinctive points
  - Describe images, objects, parts without requiring segmentation
  - Local character means robustness to clutter, occlusion
  - Robustness: similar descriptors in spite of noise, blur, etc.

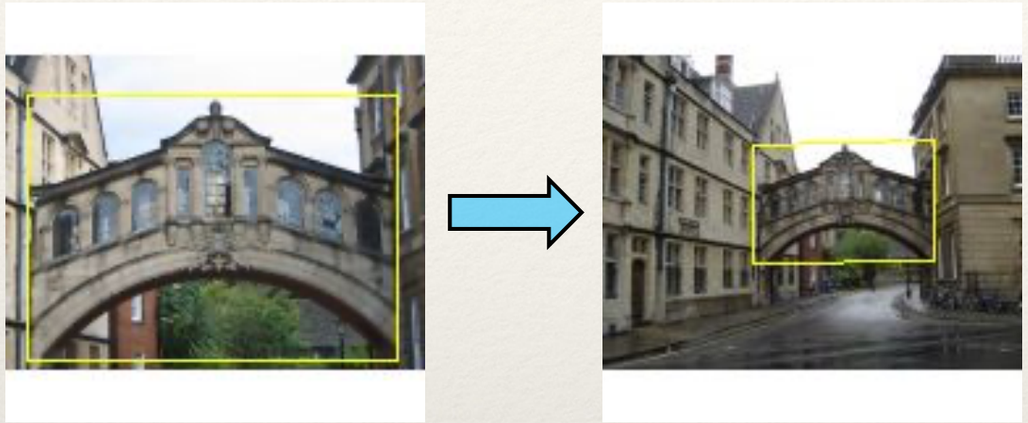
Applications: automatic mosaicing



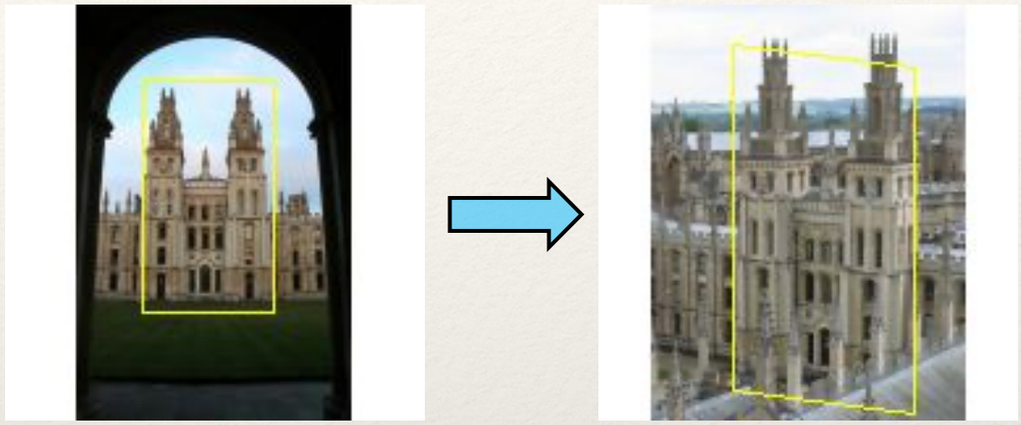
Matthew Brown

http://matthewalunbrown.com/autostitch/autostitch.html

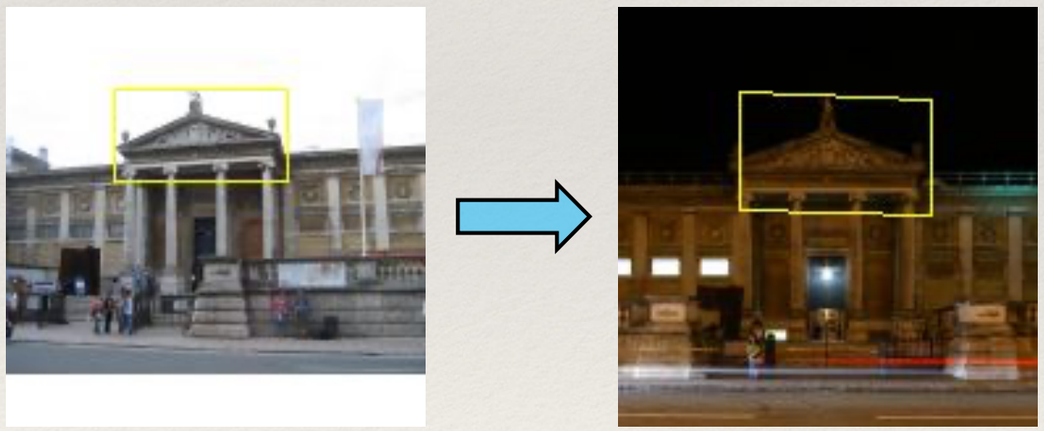
Applications: recognition of specific objects, scenes



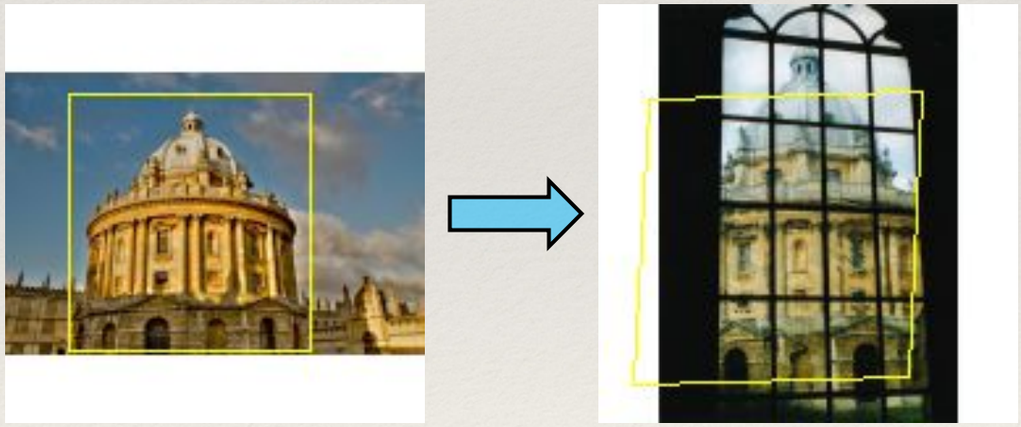
Scale



Viewpoint



Lighting



Occlusion

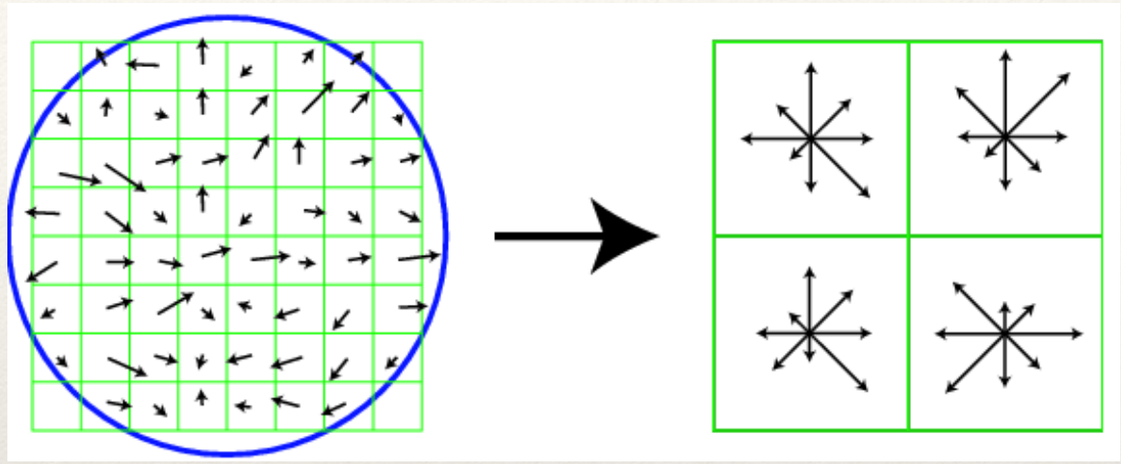
Slide credit: J. Sivic

Applications: wide baseline stereo

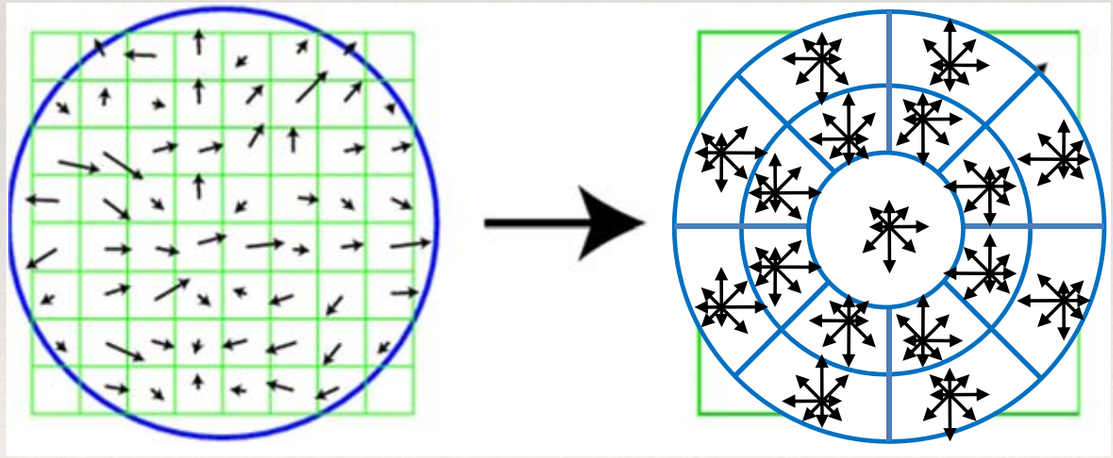
[Image from T. Tuytelaars ECCV 2006 tutorial]

Descriptors - GLOH

- Uses a log-polar binning structure instead of the four quadrants
- The spatial bins are of radius 6, 11 and 15: 3 bins in radius direction. 17 location bins in total and 16 orientation bins results in 272 bin histogram.
- Uses PCA to reduce the dimensionality to 128.



(a) image gradients (b) keypoint descriptor

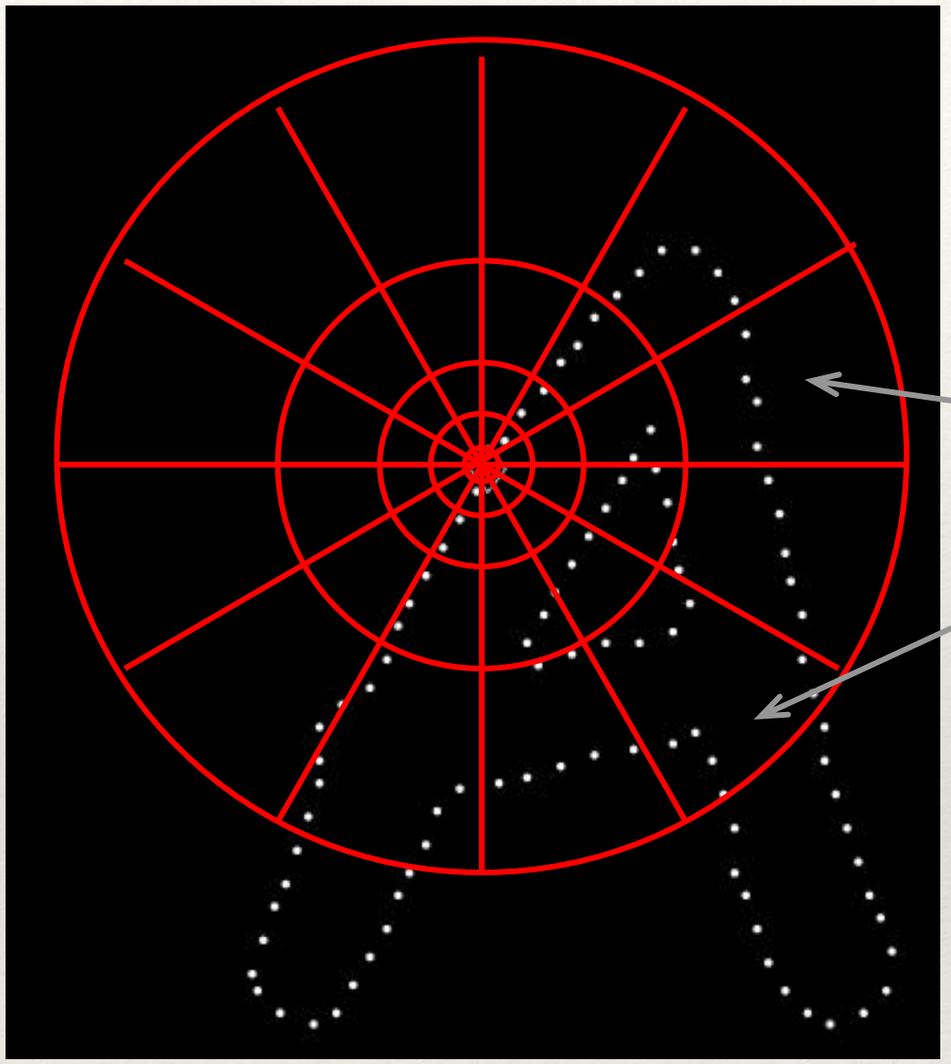


(a) image gradients (b) keypoint descriptor

Descriptors - shape context

- Originally designed for shape matching
- A set of vectors originating from a point to all other sample points on a shape.
- For a point p_i on the shape, compute a coarse histogram h_i of the relative coordinates of the remaining n-1 points

$$h_i(k) = \#\{q \neq p_i \quad : \quad (q - p_i) \in \mathrm{bin}(k)\}$$



Count the number of points inside each bin, e.g.:

Count = 4

⋮

Count = 10

Log-polar binning:
More precision for nearby points, more flexibility for farther points.

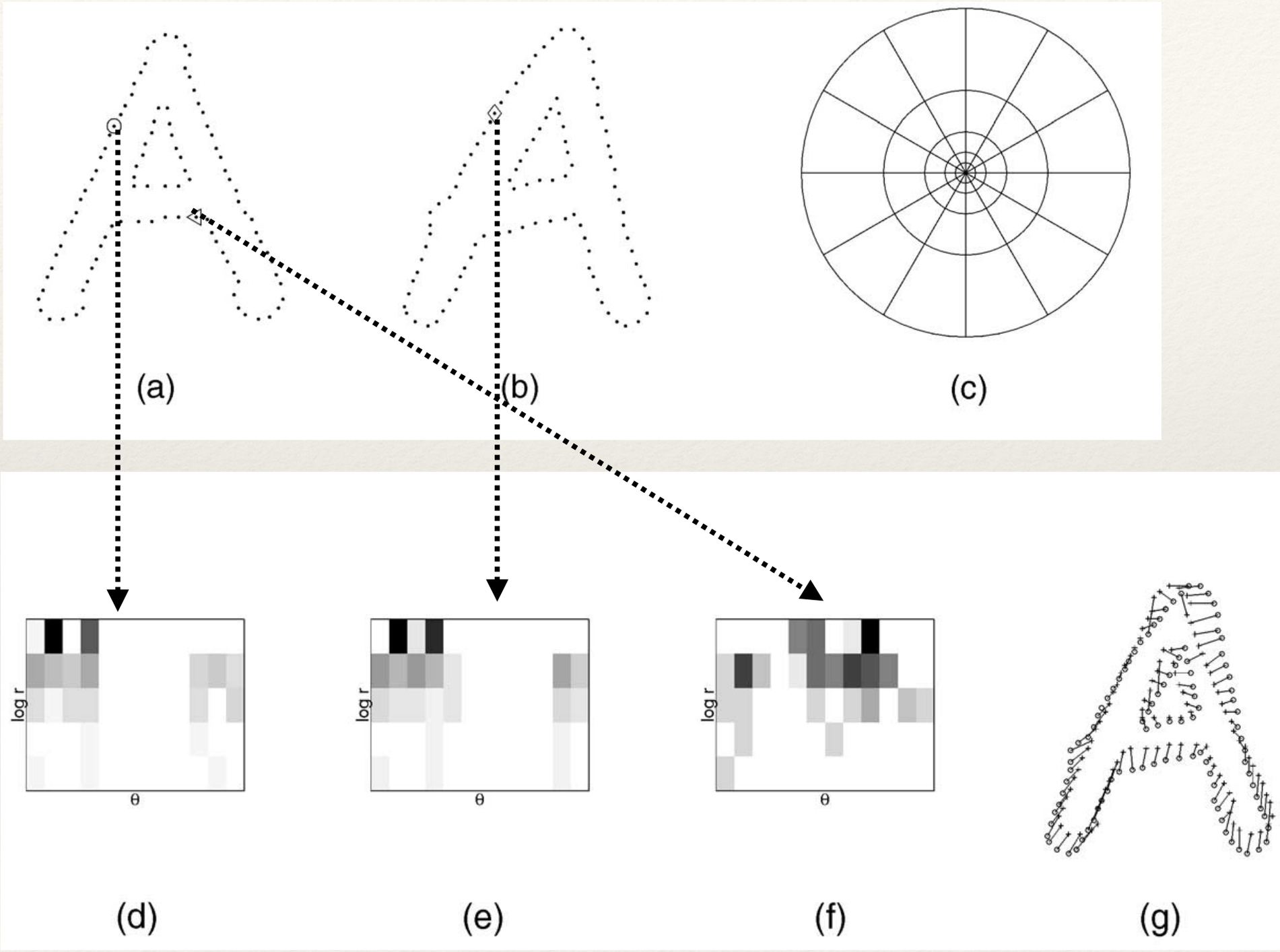
Descriptors - shape context

- Shape matching: X^2 distance

$$C_{ij} \equiv C(p_i, q_j) = \frac{1}{2}\sum_{k=1}^{K}\frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)}$$

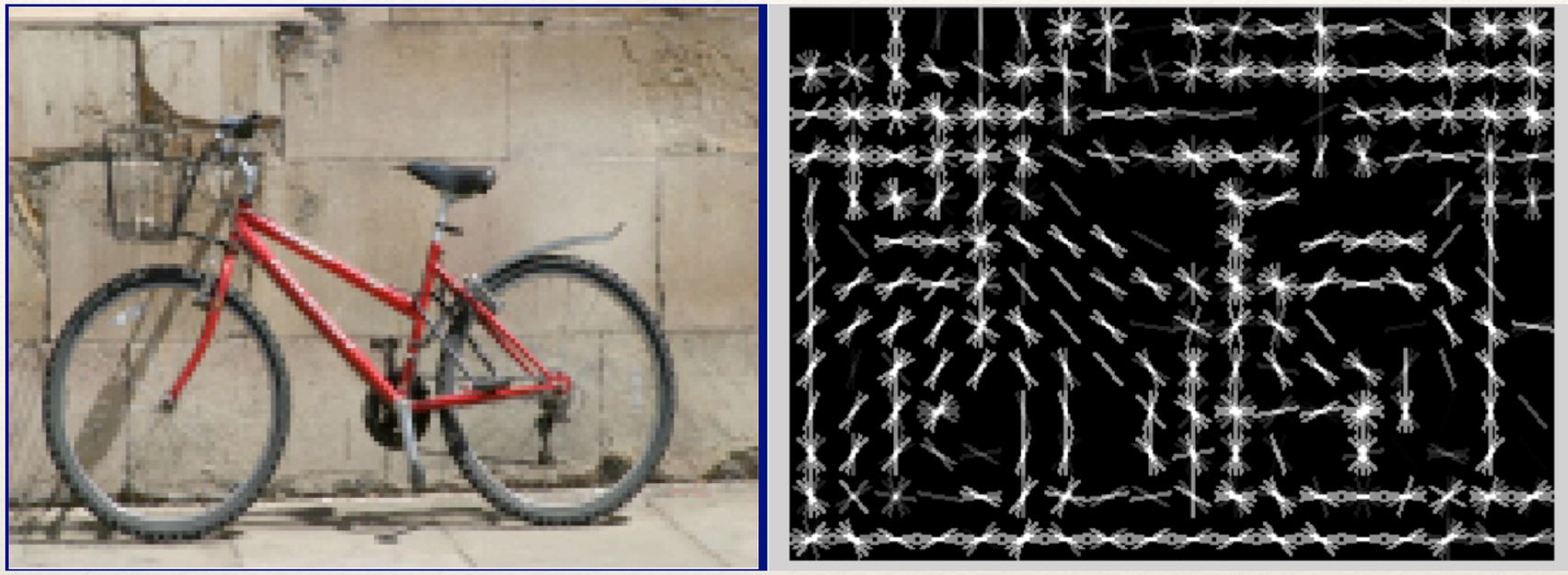
- Bipariate graph matching: used for minimising the total cost of matching between two sets of points

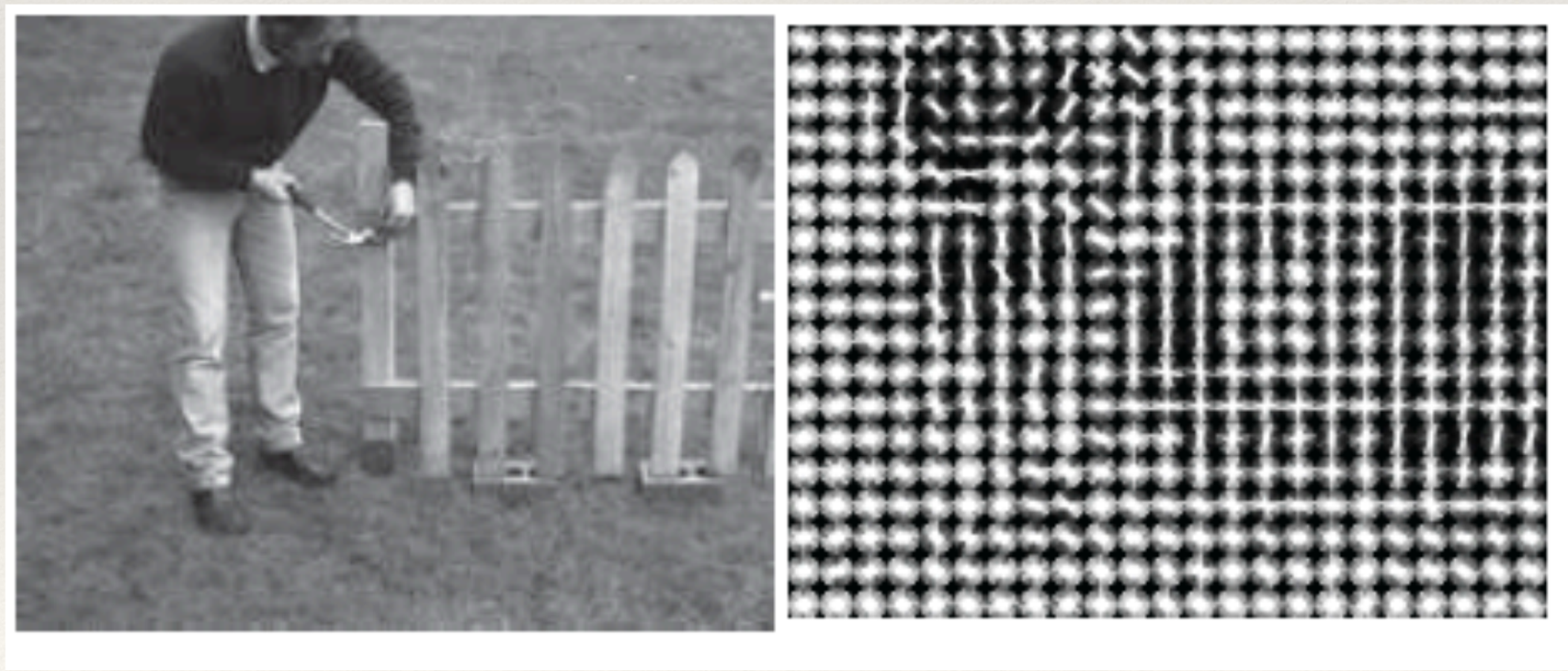
$$H(\pi) = \sum_i C\left(p_i, q_{\pi(i)}\right)$$


(a)
(b)
(c)
log r
θ
log r
θ
log r
θ
(d)
(e)
(f)
(g)

Descriptors - HOG

- Strategy
  - break patch up into blocks
  - construct histogram representing gradients in that block, which won't change much if the patch moves slightly
- Variants
  - histogram of angles
  - histogram of gradient vectors, length normalized by block averages





D. A. Forsyth

HOG and SIFT - Crucial points

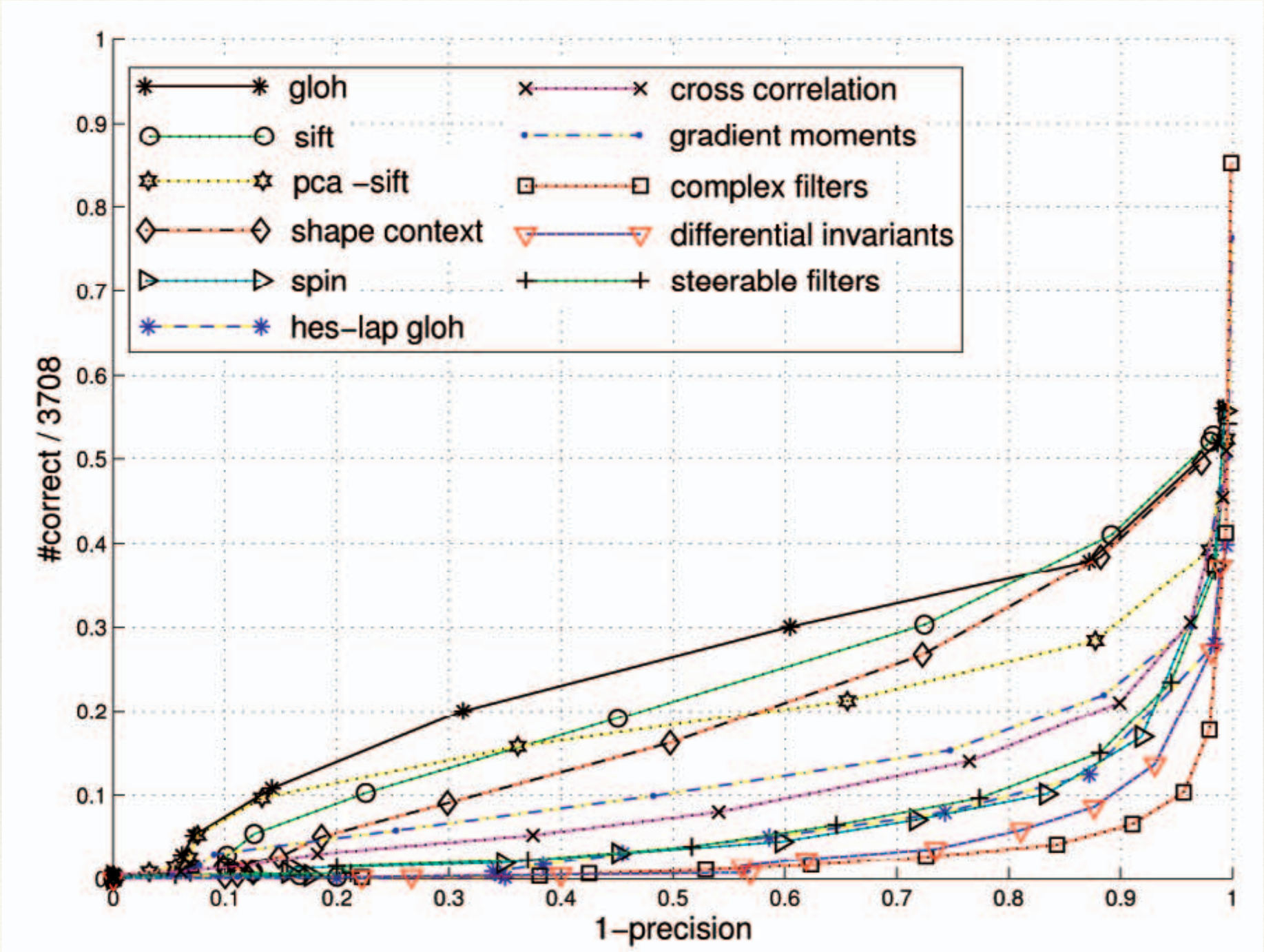
- Orientation based descriptors are very powerfu
  - because robust to changes in brightness
- HOG feature
  - known window, make histogram of orientations
- SIFT feature
  - find domain: patch center and radius
  - compute descriptor: histogram of orientations
- Numerous powerful variants
- Software available

Descriptor evaluation

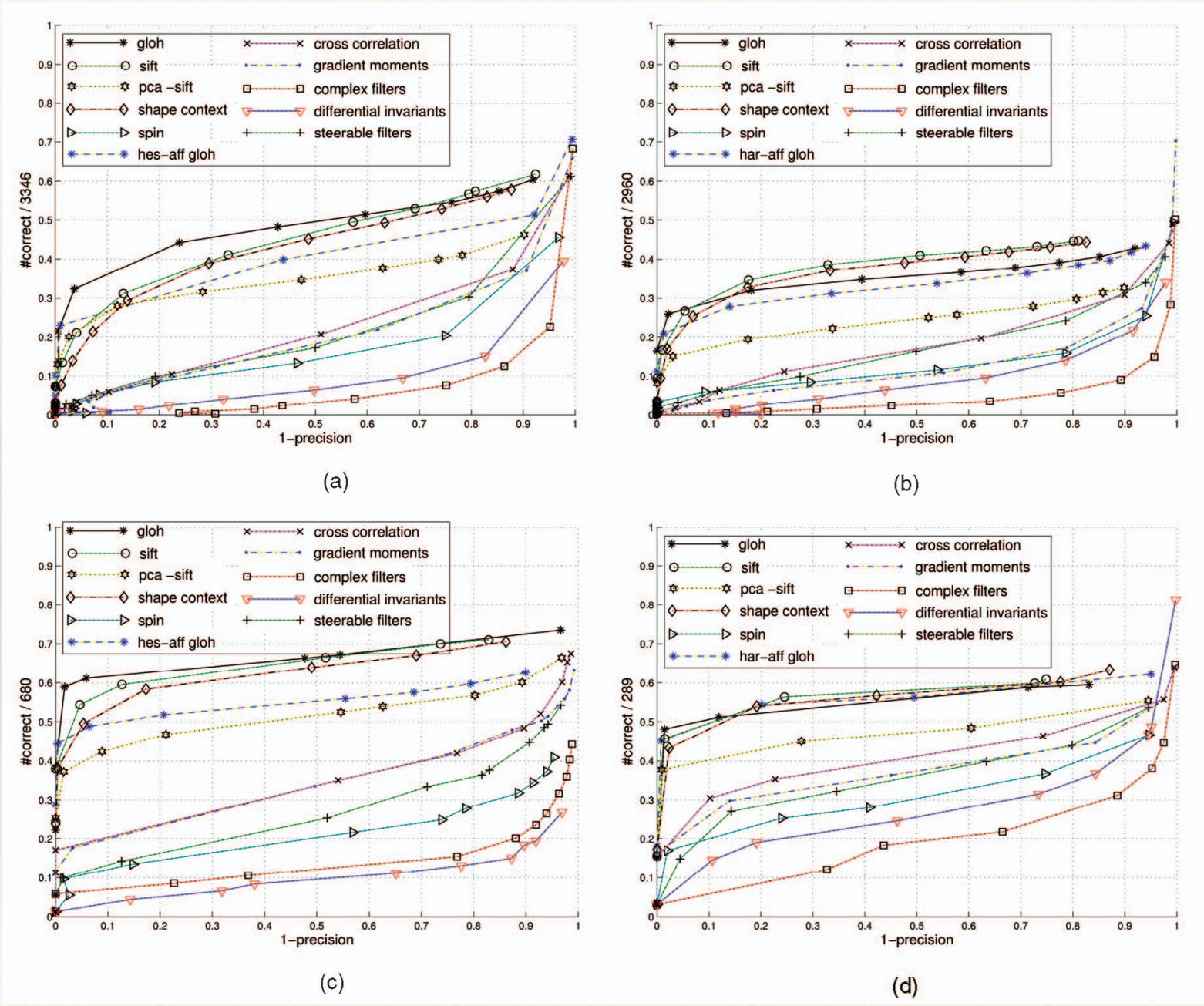
- Evaluate the robustness to:
  - Affine transform
  - Scale changes
  - Image rotation
  - Image blur
  - JPEG compression
  - Illumination change

Mikolajczyk and Schmid (2005)

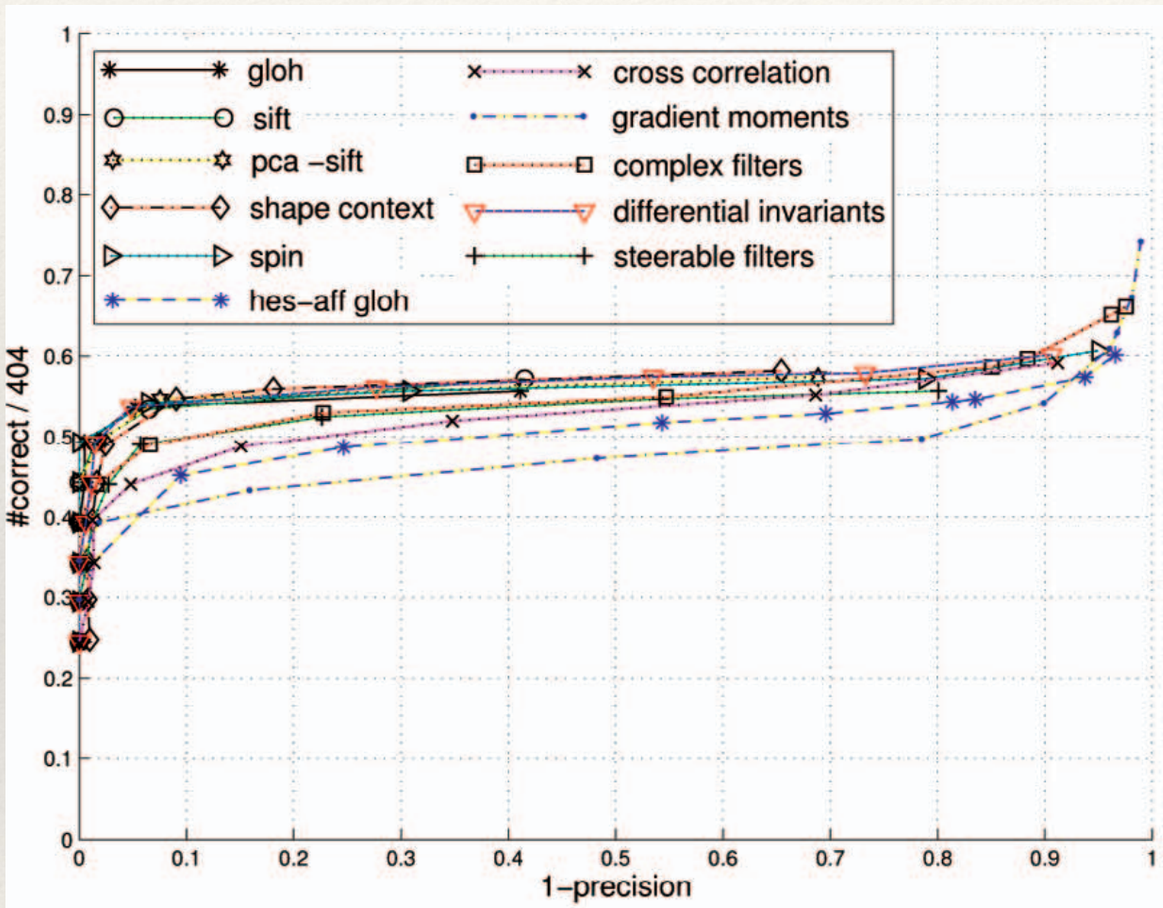
Affine


gloh
sift
pca –sift
shape context
spin
hes–lap gloh
cross correlation
gradient moments
complex filters
differential invariants
steerable filters
#correct / 3708
1–precision
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

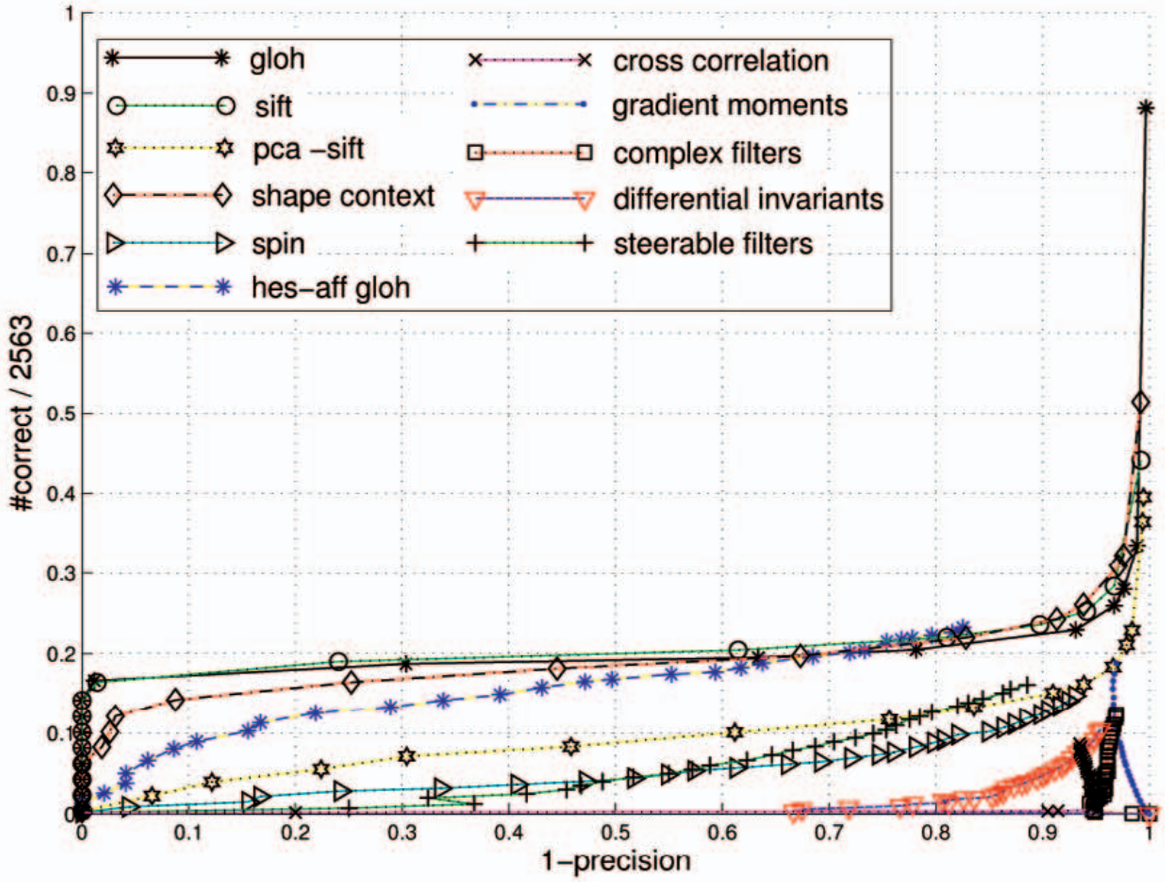
Scale


gloh
sift
pca –sift
shape context
spin
hes–aff gloh
cross correlation
gradient moments
complex filters
differential invariants
steerable filters
#correct / 3346
1–precision
(a)
gloh
sift
pca –sift
shape context
spin
har–aff gloh
cross correlation
gradient moments
complex filters
differential invariants
steerable filters
#correct / 2960
1–precision
(b)
gloh
sift
pca –sift
shape context
spin
hes–aff gloh
cross correlation
gradient moments
complex filters
differential invariants
steerable filters
#correct / 680
1–precision
(c)
gloh
sift
pca –sift
shape context
spin
har–aff gloh
cross correlation
gradient moments
complex filters
differential invariants
steerable filters
#correct / 289
1–precision
(d)

Rotation


gloh
sift
pca –sift
shape context
spin
hes–aff gloh
cross correlation
gradient moments
complex filters
differential invariants
steerable filters
#correct / 404
1–precision
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

(a)


gloh
sift
pca –sift
shape context
spin
hes–aff gloh
cross correlation
gradient moments
complex filters
differential invariants
steerable filters
#correct / 2563
1–precision
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

(b)

Conclusion

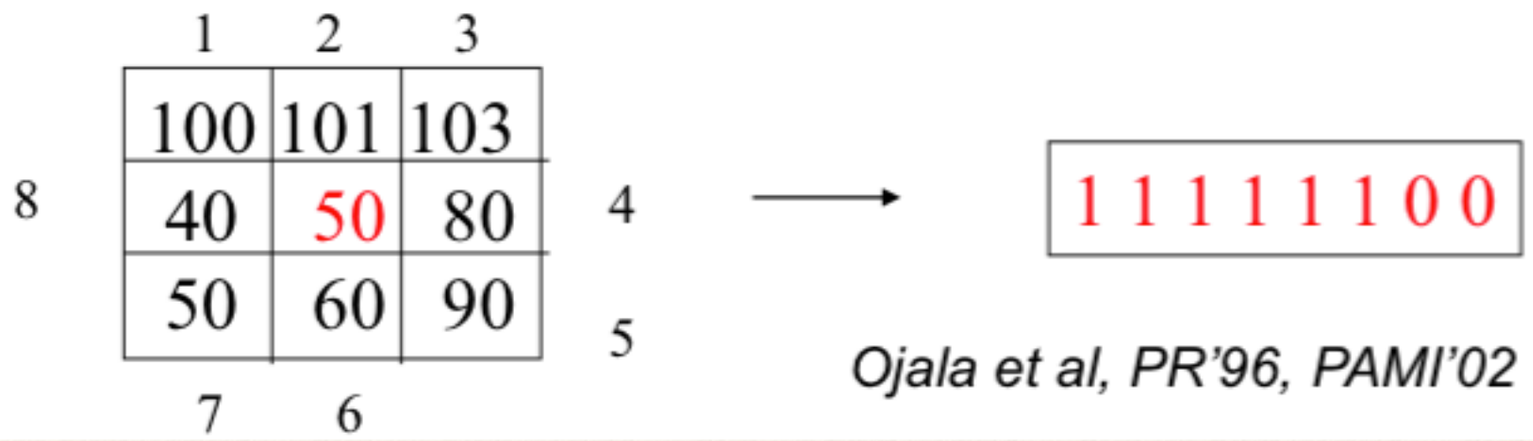
- In most of the tests, GLOH obtains the best results, closely followed by SIFT. This shows the robustness and the distinctive character of the region-based SIFT descriptor.
- Shape context also shows a high performance. However, for textured scenes or when edges are not reliable, its score is lower.

Binary features

- Local Binary Pattern (LBP)

• For each pixel p, create an 8-bit number $b_1\ b_2\ b_3\ b_4\ b_5\ b_6\ b_7\ b_8$, where $b_i = 0$ if neighbor i has value less than or equal to p's value and 1 otherwise.

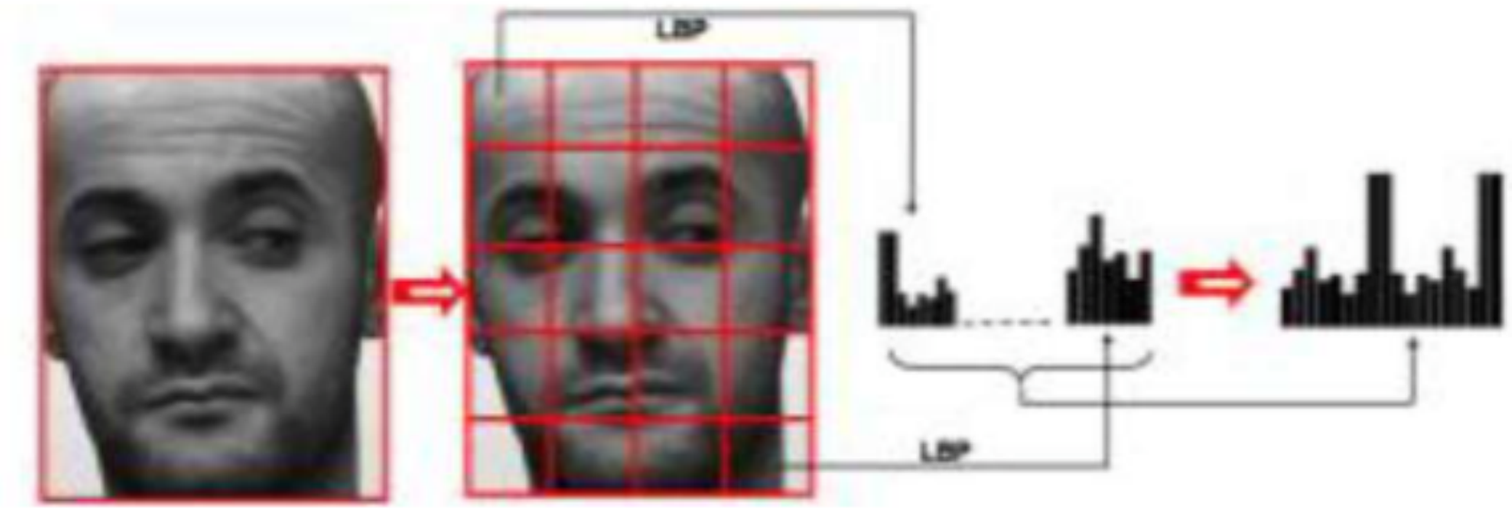
• Represent the texture in the image (or a region) by the histogram of these numbers.



Ojala et al, PR'96, PAMI'02

LBP histograms

- Divide the examined window to cells (e.g. 16x16 pixels for each cell).
- Compute the histogram, over the cell, of the frequency of each "number" occurring.
- Optionally normalize the histogram.
- Concatenate normalized histograms of all cells.



Other binary features

- BRIEF - Binary Robust Independent Elementary Features
- ORB - Oriented FAST and Rotated BRIEF
- BRISK - Binary Robust Invariant Scalable Keypoints

Summary

- Interest point detection
  - Harris corner detector
  - Laplacian of Gaussian, automatic scale selection
- Invariant descriptors
  - Rotation according to dominant gradient direction
  - Histograms for robustness to small shifts and translations (SIFT descriptor)