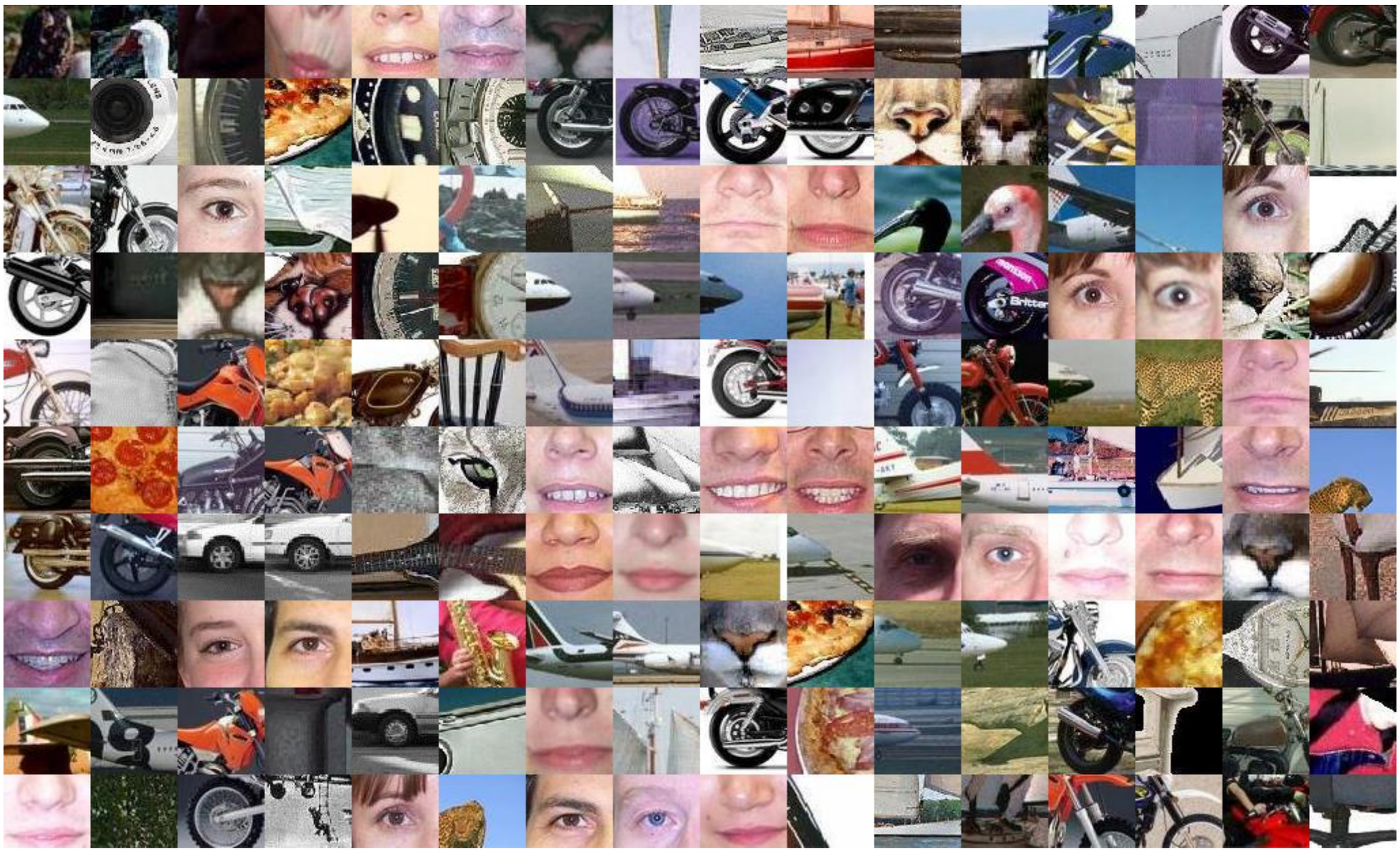
Feature detectors and descriptors



http://www.cs.cmu.edu/~16385/

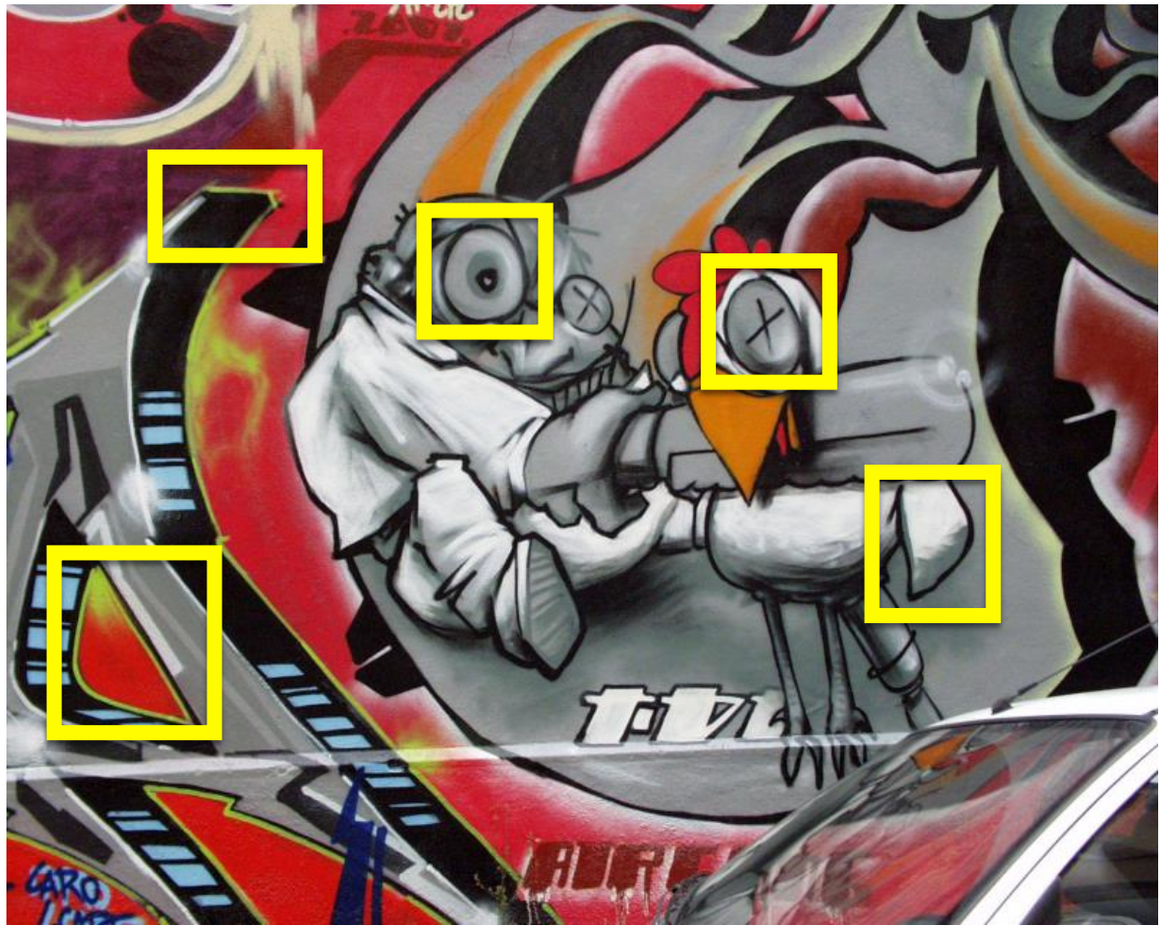
16-385 Computer Vision
Spring 2019, Lecture 6

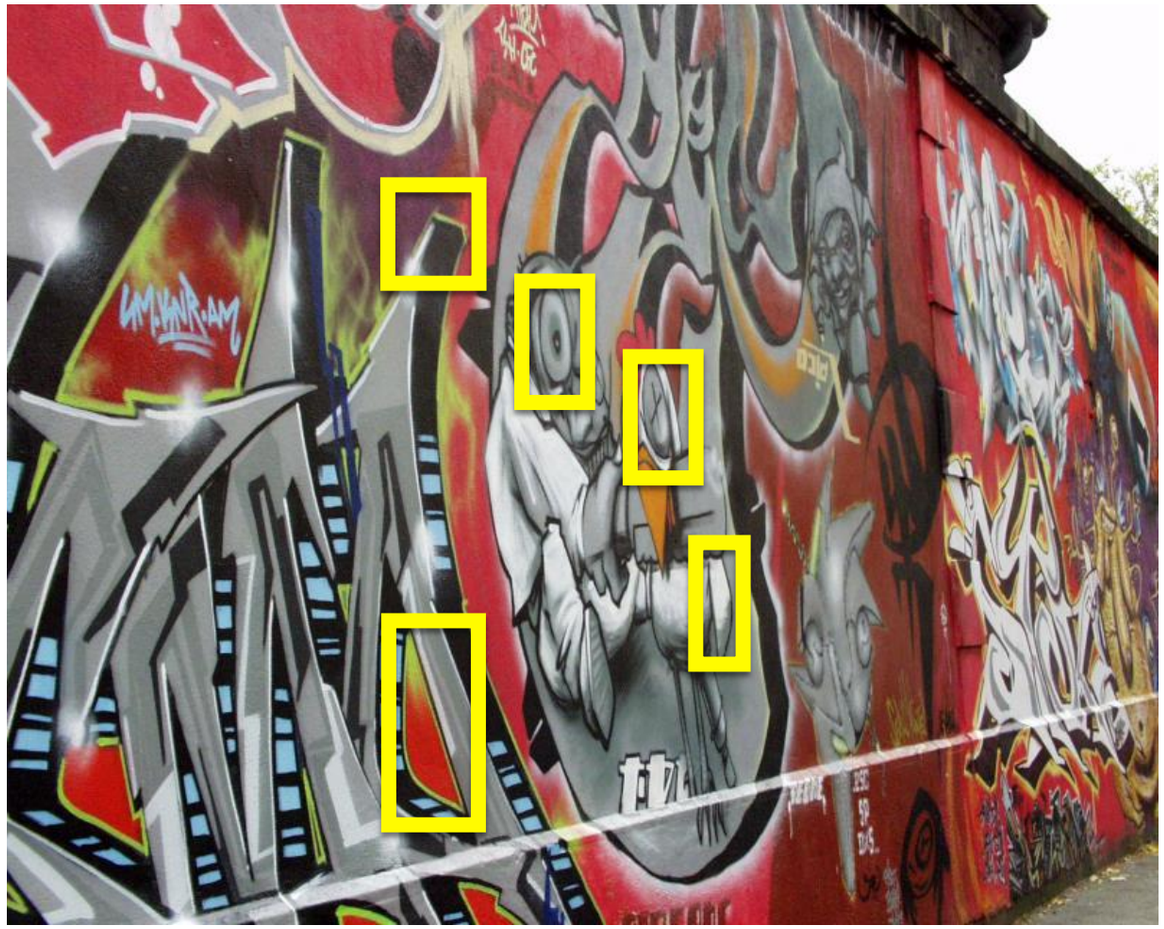
Slide credits

Most of these slides were adapted from:

- Kris Kitani (16-385, Spring 2017).

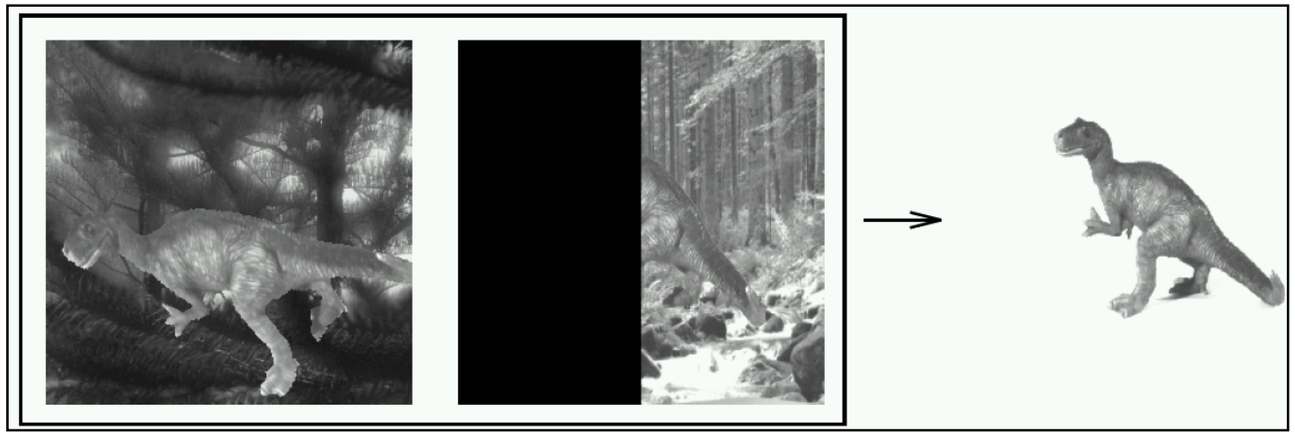
Why do we need feature descriptors?





If we know where the good features are,
how do we match them?

Object instance recognition



Schmid and Mohr 1997

Sivic and Zisserman, 2003

Rothganger et al. 2003

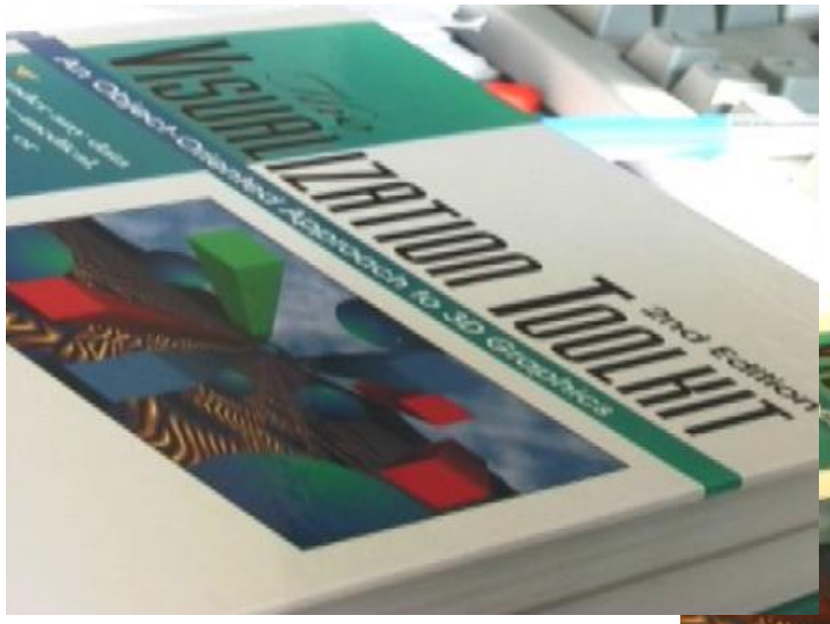
Lowe 2002

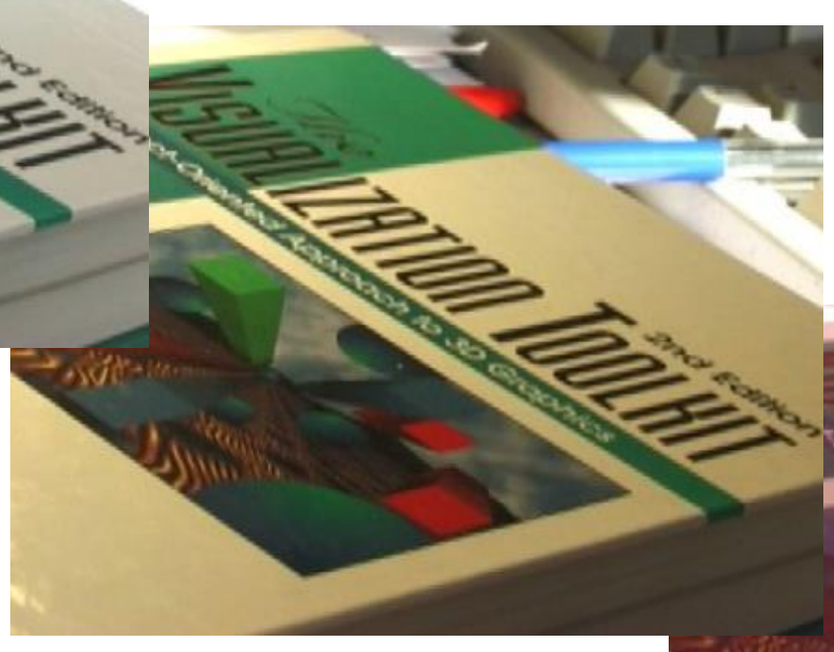
How do we describe an image patch?

Patches with similar content should have similar descriptors.

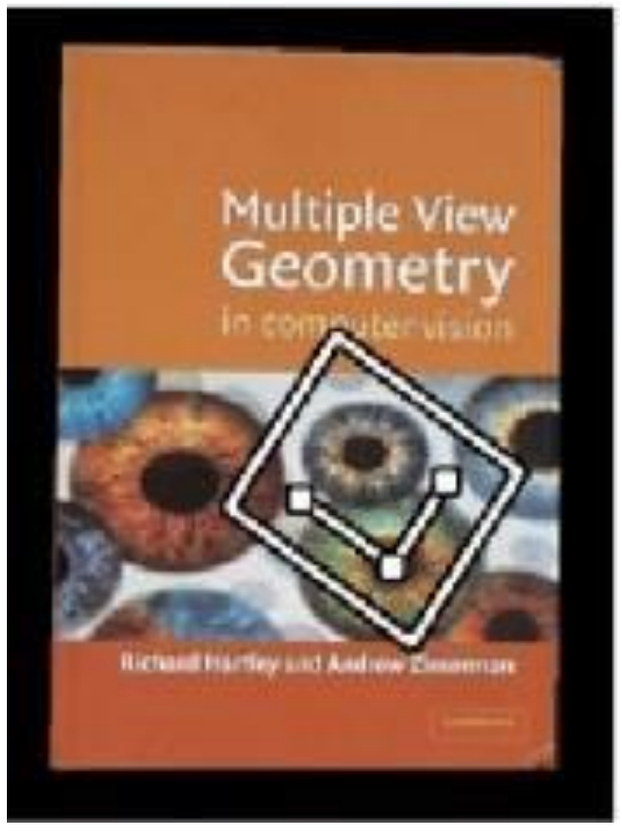
Designing feature descriptors

Photometric transformations





Geometric transformations



objects will appear at different scales, translation and rotation

What is the best descriptor for an image feature?

Image patch

Just use the pixel values of the patch

Perfectly fine if geometry and appearance is unchanged

(a.k.a. template matching)

Tiny Images

Just down-sample it!
Simple, fast, robust to small affine transforms.

Image patch

Just use the pixel values of the patch

Perfectly fine if geometry and appearance is unchanged

(a.k.a. template matching)

What are the problems?

Image patch

Just use the pixel values of the patch

Perfectly fine if geometry and appearance is unchanged

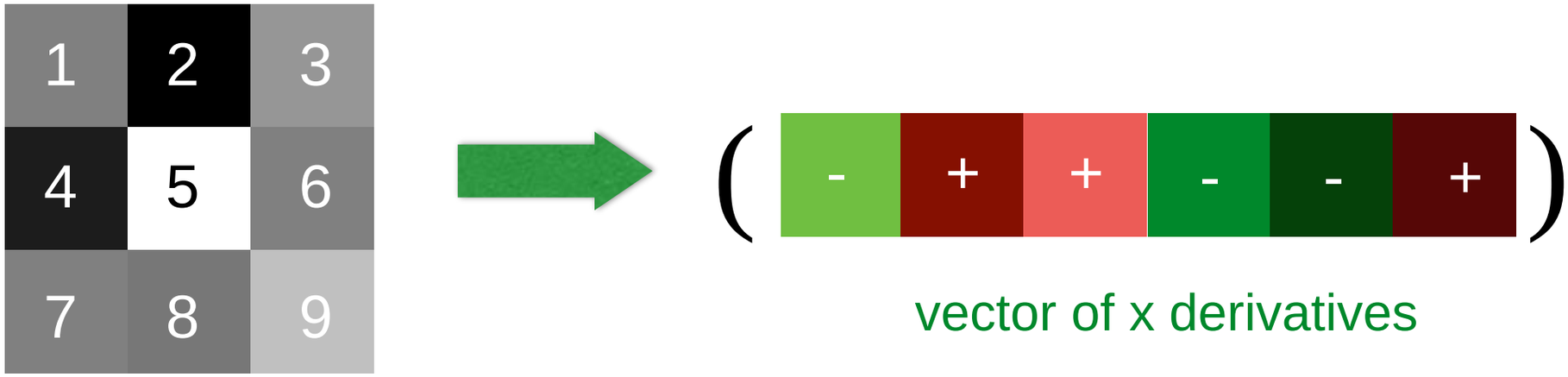
(a.k.a. template matching)

What are the problems?

How can you be less sensitive to absolute intensity values?

Image gradients

Use pixel differences



vector of x derivatives

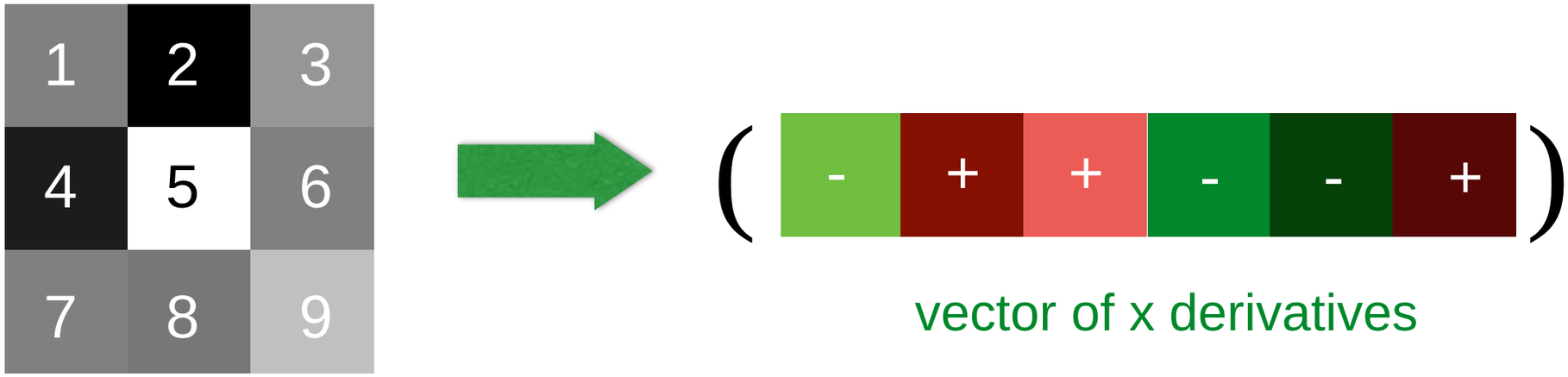
‘binary descriptor’

Feature is invariant to absolute intensity values

What are the problems?

Image gradients

Use pixel differences



vector of x derivatives

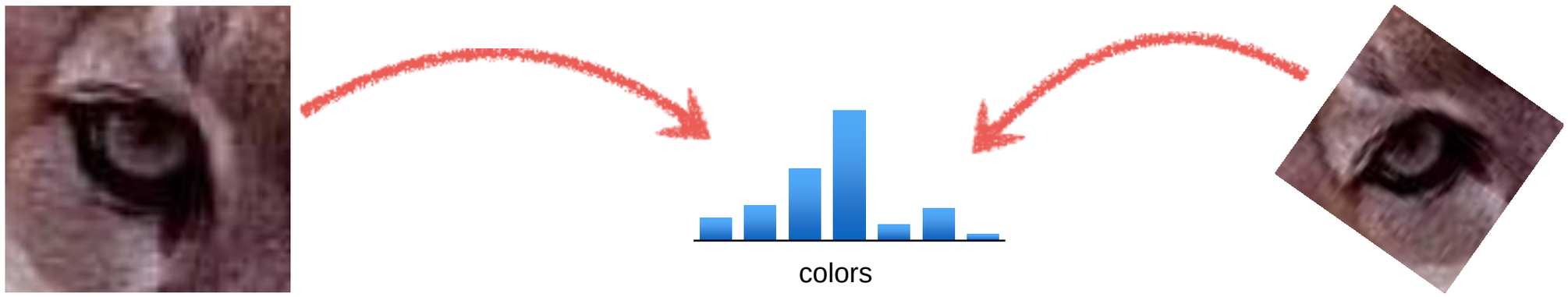
Feature is invariant to absolute intensity values

What are the problems?

How can you be less sensitive to deformations?

Color histogram

Count the colors in the image using a histogram

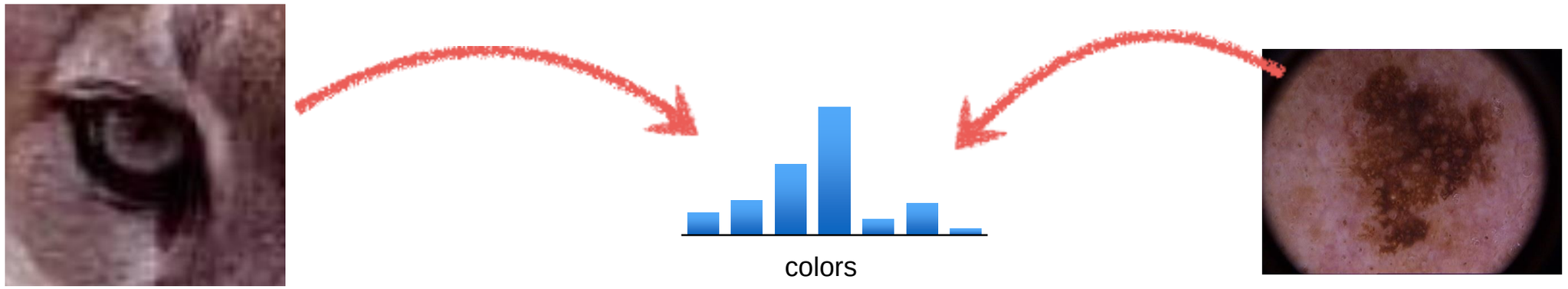


Invariant to changes in scale and rotation

What are the problems?

Color histogram

Count the colors in the image using a histogram

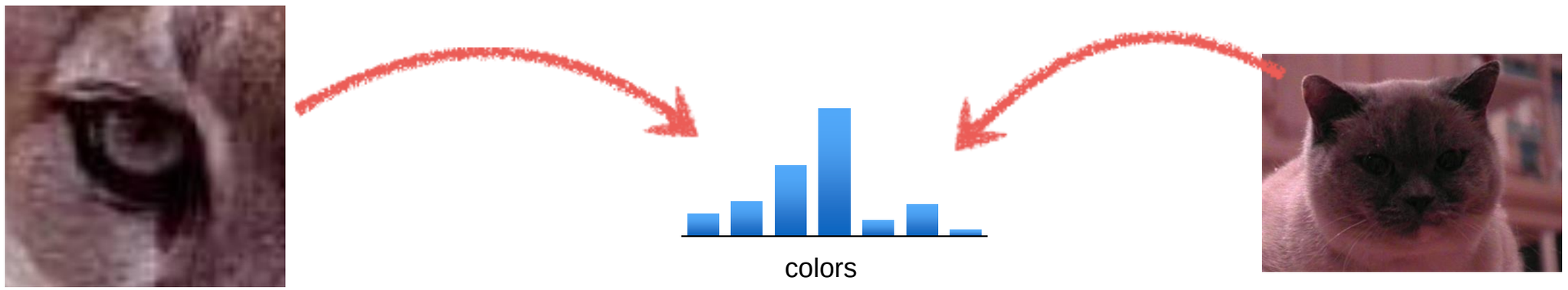


Invariant to changes in scale and rotation

What are the problems?

Color histogram

Count the colors in the image using a histogram



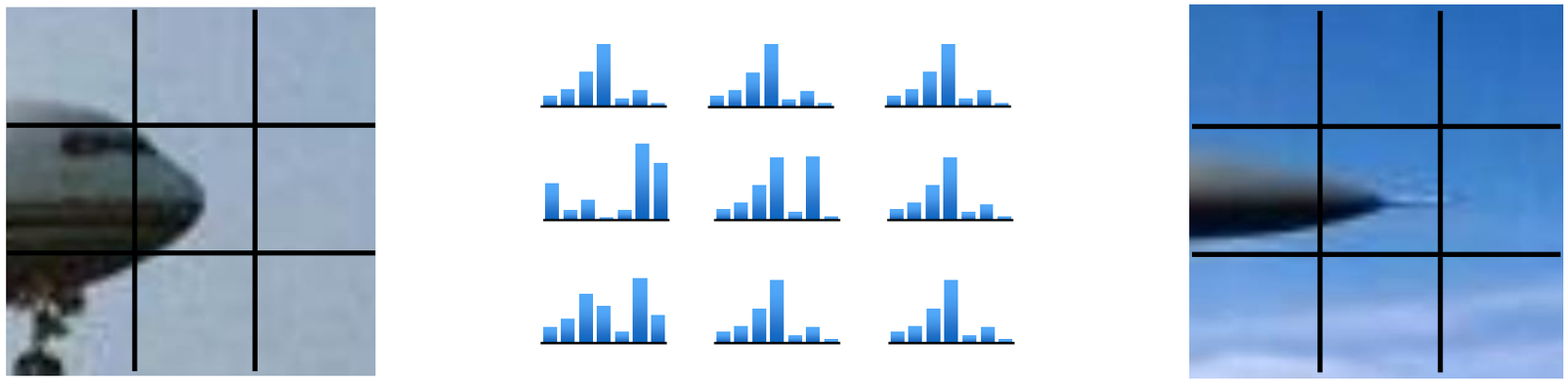
Invariant to changes in scale and rotation

What are the problems?

How can you be more sensitive to spatial layout?

Spatial histograms

Compute histograms over spatial ‘cells’



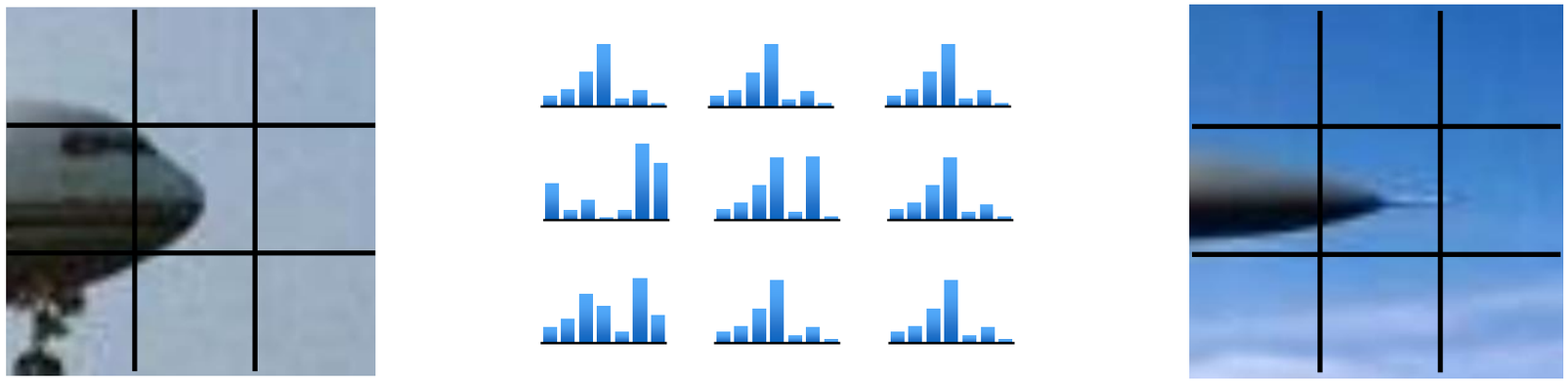
Retains rough spatial layout

Some invariance to deformations

What are the problems?

Spatial histograms

Compute histograms over spatial ‘cells’



Retains rough spatial layout

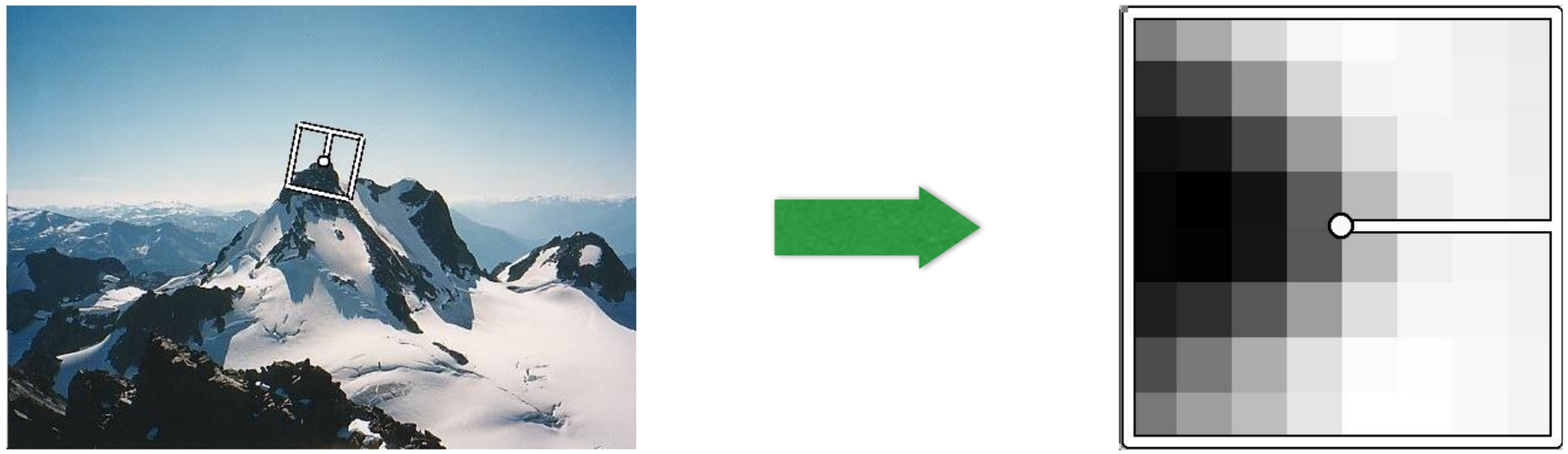
Some invariance to deformations

What are the problems?

How can you be completely invariant to rotation?

Orientation normalization

Use the dominant image gradient direction to normalize the orientation of the patch



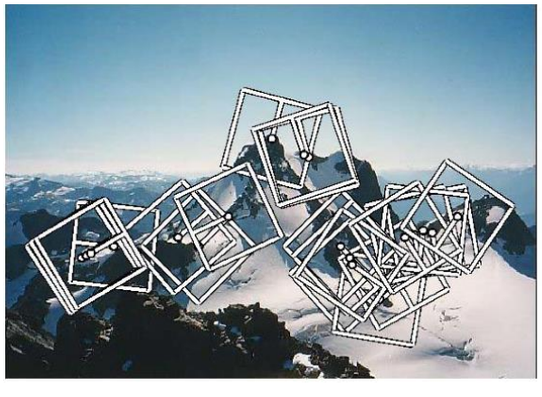
save the orientation angle θ along with (x, y, s)

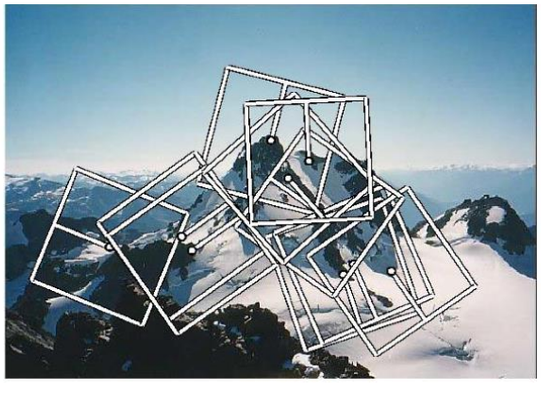
What are the problems?

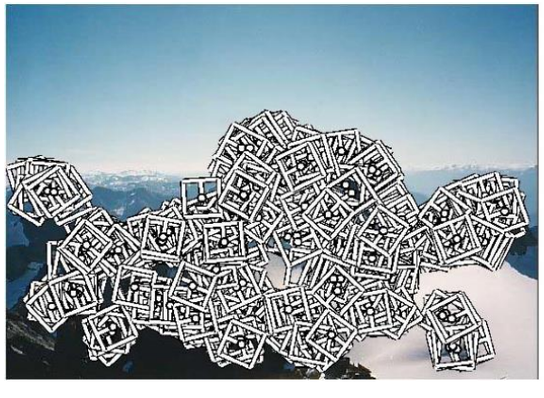
MOPS descriptor

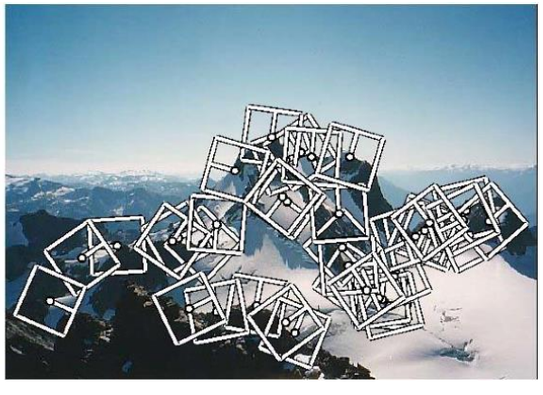
Multi-Scale Oriented Patches (MOPS)

Multi-Image Matching using Multi-Scale Oriented Patches. M. Brown, R. Szeliski and S. Winder. International Conference on Computer Vision and Pattern Recognition (CVPR2005). pages 510-517









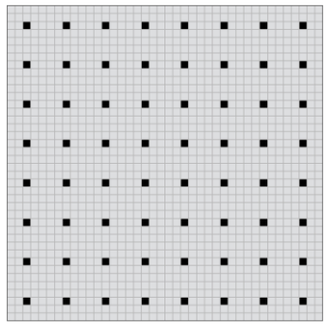
Multi-Scale Oriented Patches (MOPS)

Multi-Image Matching using Multi-Scale Oriented Patches. M. Brown, R. Szeliski and S. Winder. International Conference on Computer Vision and Pattern Recognition (CVPR2005). pages 510-517

Given a feature (x, y, s, θ)

Get 40 x 40 image patch, subsample every 5th pixel

(*what's the purpose of this step?*)



Subtract the mean, divide by standard deviation

(*what's the purpose of this step?*)

Haar Wavelet Transform

(*what's the purpose of this step?*)

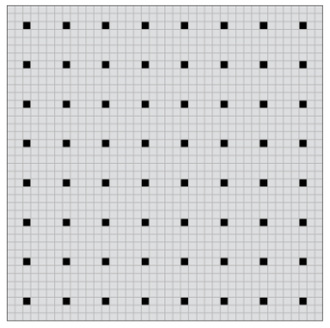
Multi-Scale Oriented Patches (MOPS)

Multi-Image Matching using Multi-Scale Oriented Patches. M. Brown, R. Szeliski and S. Winder. International Conference on Computer Vision and Pattern Recognition (CVPR2005). pages 510-517

Given a feature (x, y, s, θ)

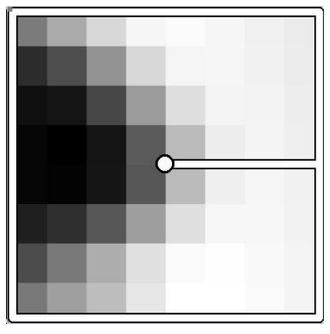
Get 40 x 40 image patch, subsample every 5th pixel

(low frequency filtering, absorbs localization errors)



Subtract the mean, divide by standard deviation

(*what's the purpose of this step?*)



Haar Wavelet Transform

(*what's the purpose of this step?*)

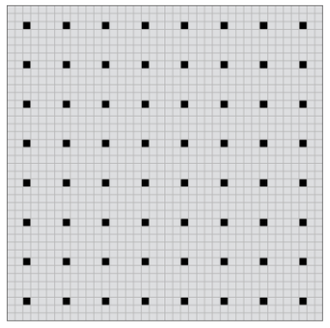
Multi-Scale Oriented Patches (MOPS)

Multi-Image Matching using Multi-Scale Oriented Patches. M. Brown, R. Szeliski and S. Winder. International Conference on Computer Vision and Pattern Recognition (CVPR2005). pages 510-517

Given a feature (x, y, s, θ)

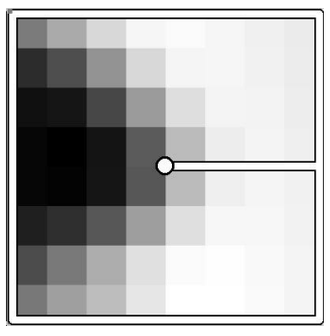
Get 40 x 40 image patch, subsample every 5th pixel

(low frequency filtering, absorbs localization errors)



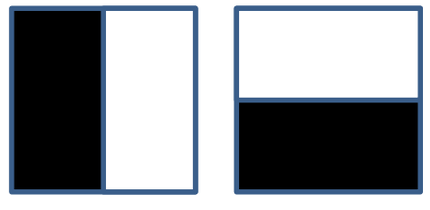
Subtract the mean, divide by standard deviation

(removes bias and gain)



Haar Wavelet Transform

(*what's the purpose of this step?*)

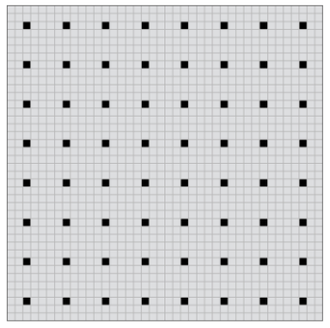


Multi-Scale Oriented Patches (MOPS)

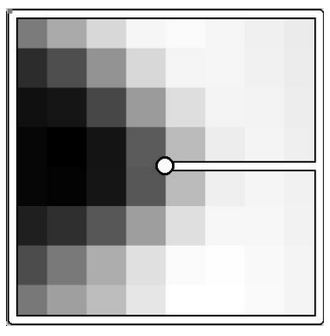
Multi-Image Matching using Multi-Scale Oriented Patches. M. Brown, R. Szeliski and S. Winder. International Conference on Computer Vision and Pattern Recognition (CVPR2005). pages 510-517

Given a feature (x, y, s, θ)

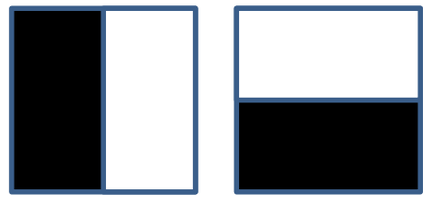
Get 40 x 40 image patch, subsample every 5th pixel
(low frequency filtering, absorbs localization errors)



Subtract the mean, divide by standard deviation
(removes bias and gain)



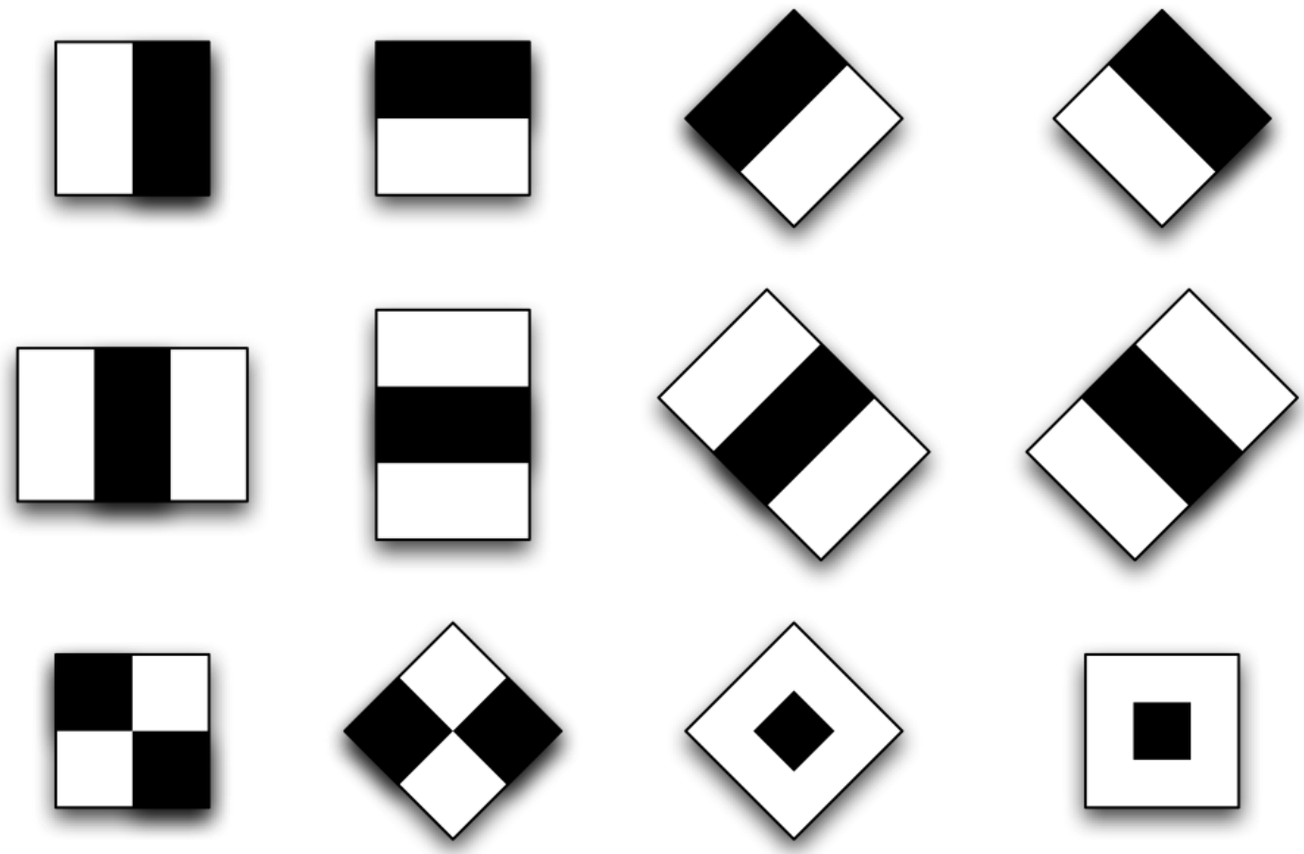
Haar Wavelet Transform
(low frequency projection)



Haar Wavelets

(actually, Haar-like features)

Use responses of a bank of filters as a descriptor



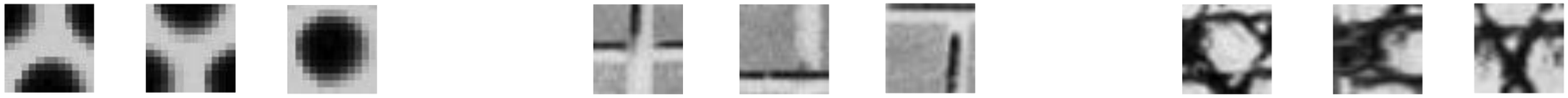
We will see later in class how to compute Haar wavelet responses **efficiently** (in constant time) with integral images

Histogram of Textons descriptor

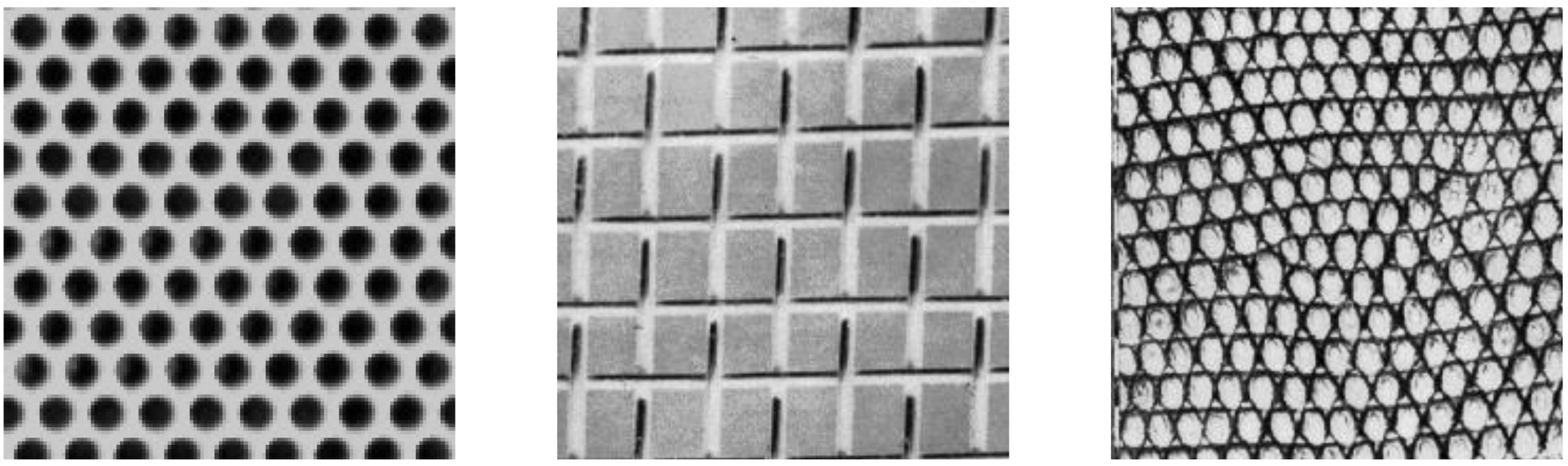
Textons

Julesz. Textons, the elements of texture perception, and their interactions. Nature 1981

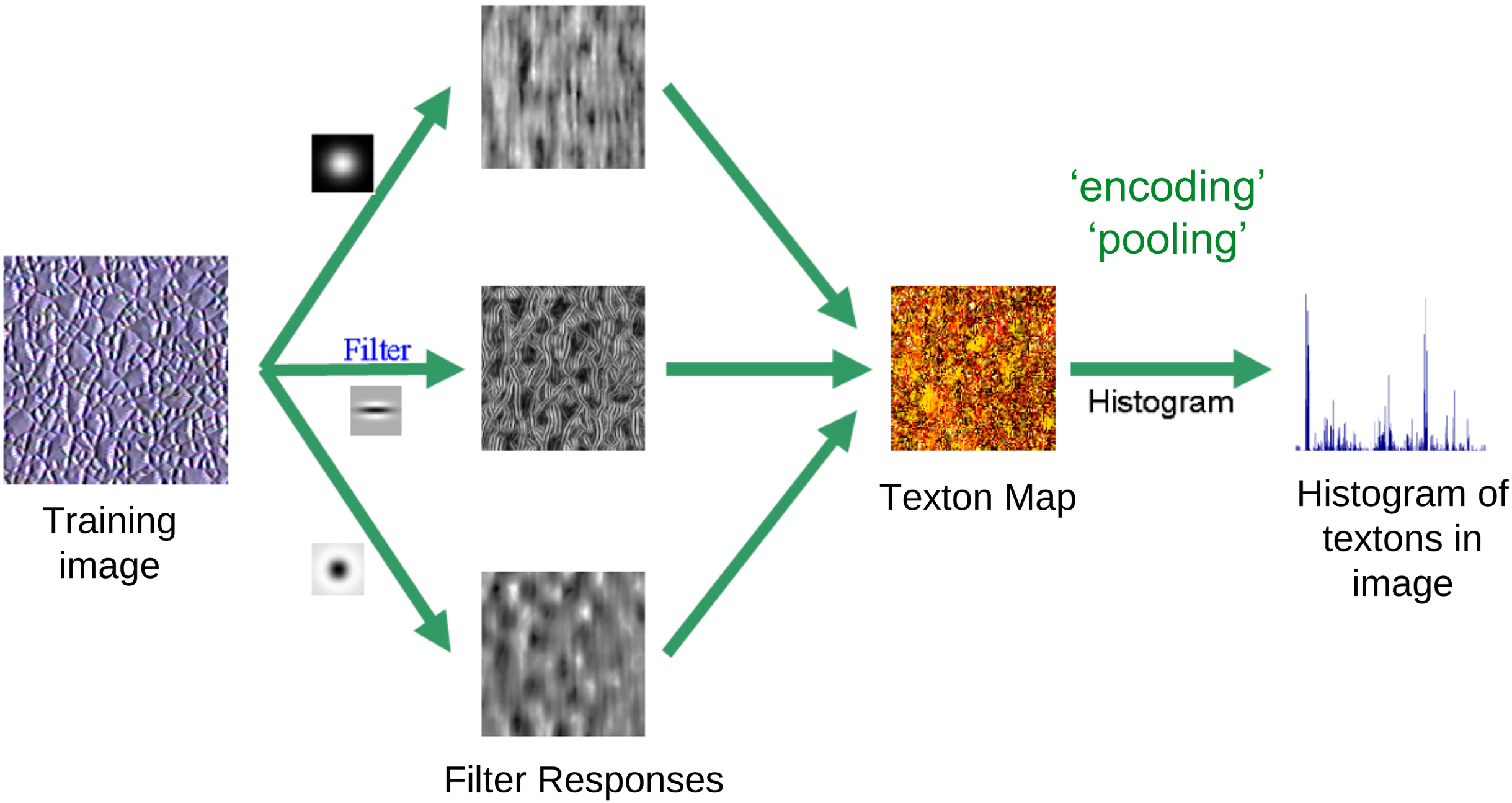
Texture is characterized by the repetition of basic elements or ***textons***



For stochastic textures, it is the identity of the ***textons***, not their spatial arrangement, that matters

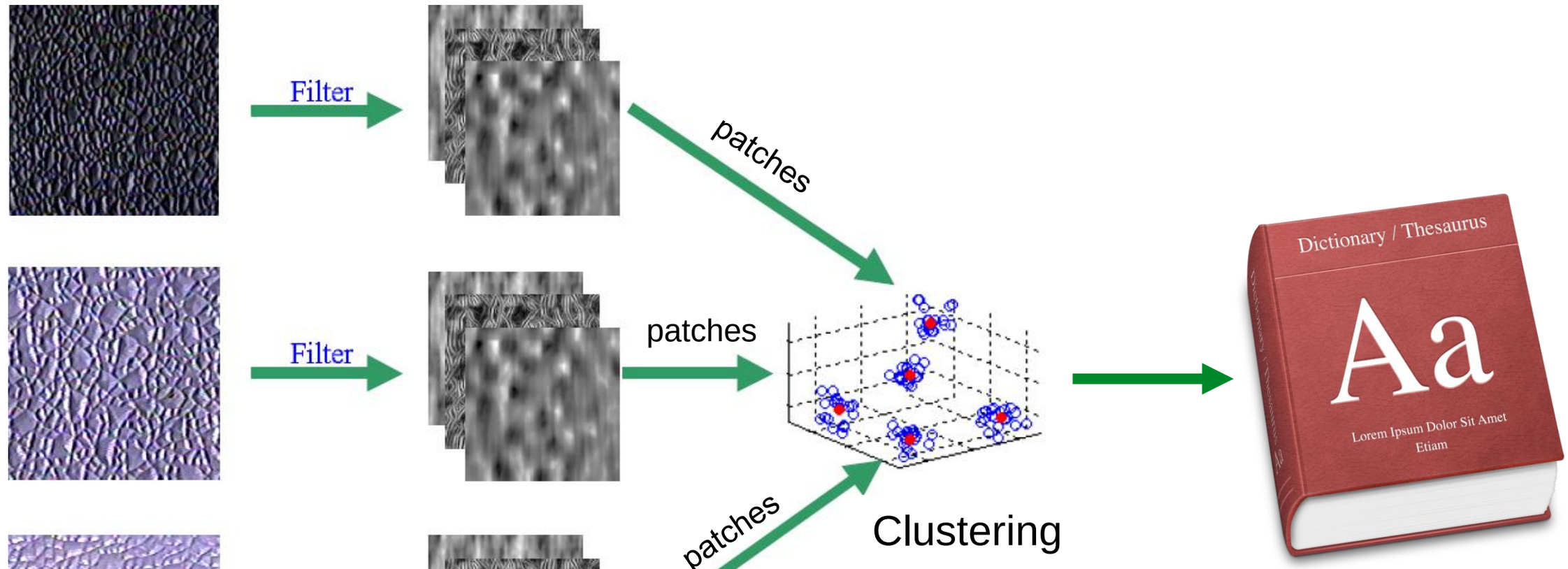


Histogram of Textons descriptor



Malik, Belongie, Shi, Leung. Textons, Contours and Regions: Cue Integration in Image Segmentation. ICCV 1999.

Learning Textons from data



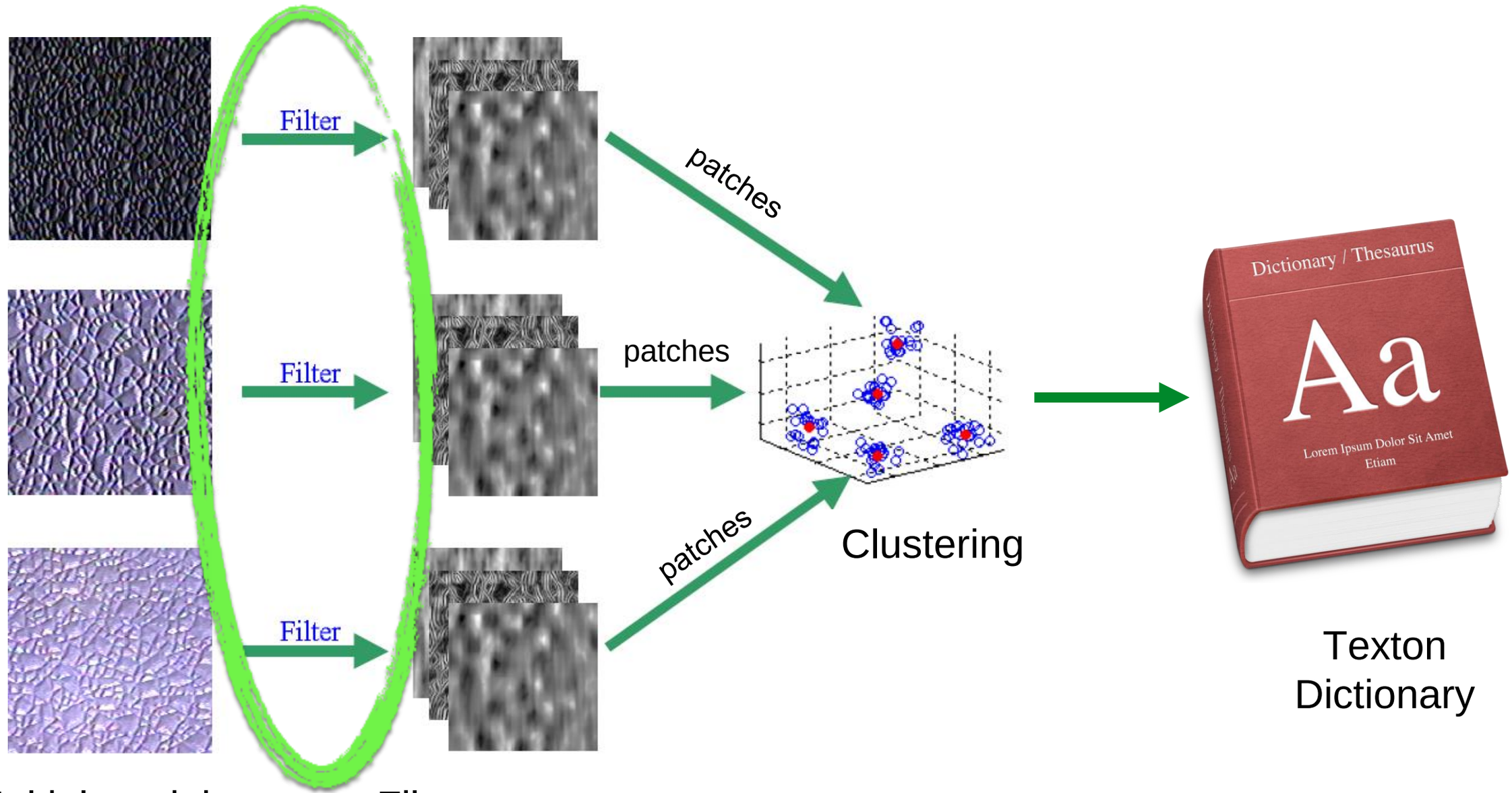
Multiple training images of the **same texture**

Filter response over a bank of filters

Clustering

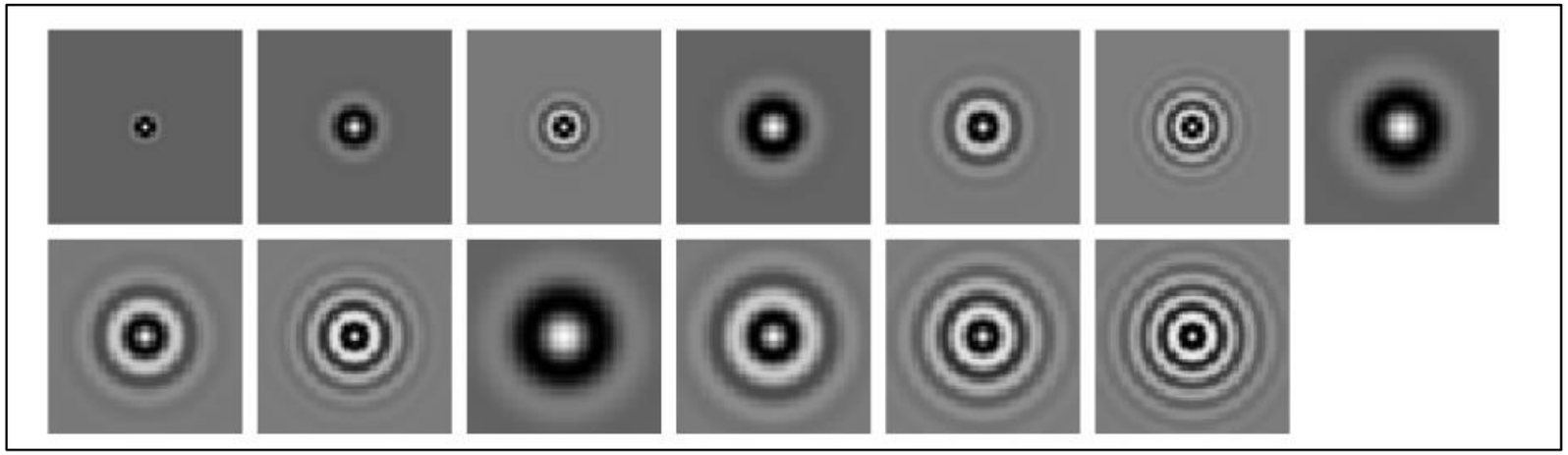
Texton Dictionary

Learning Textons from data



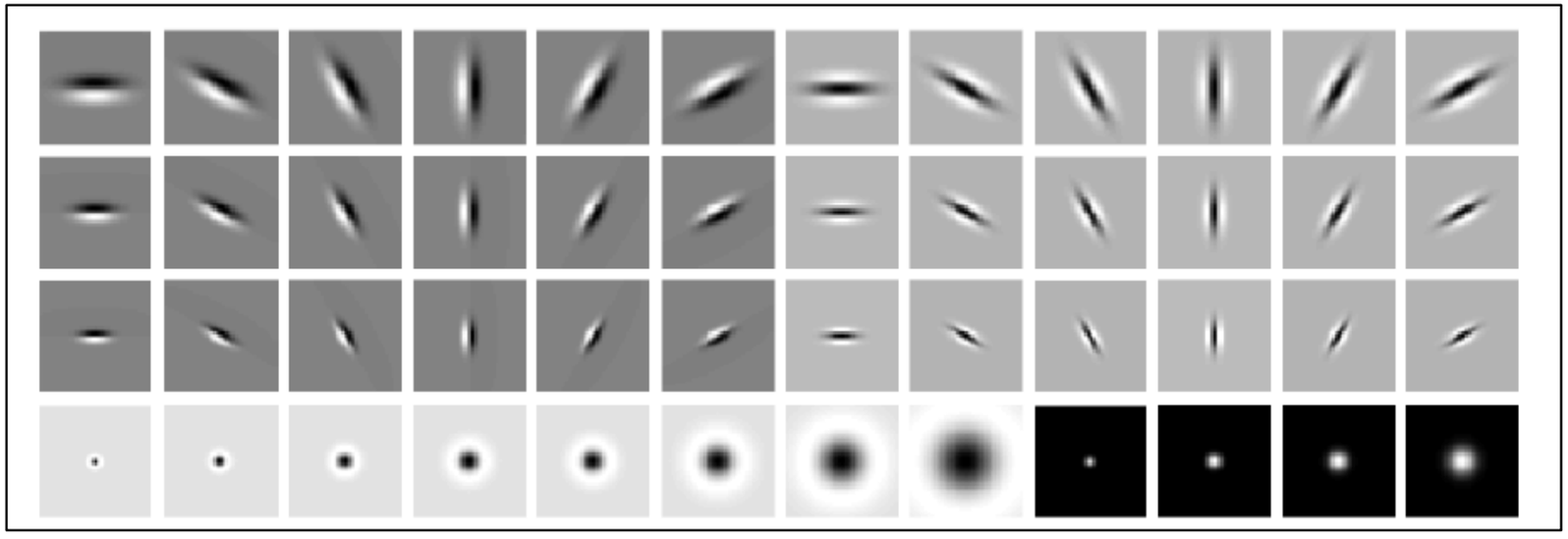
Example of Filter Banks

Isotropic Gabor

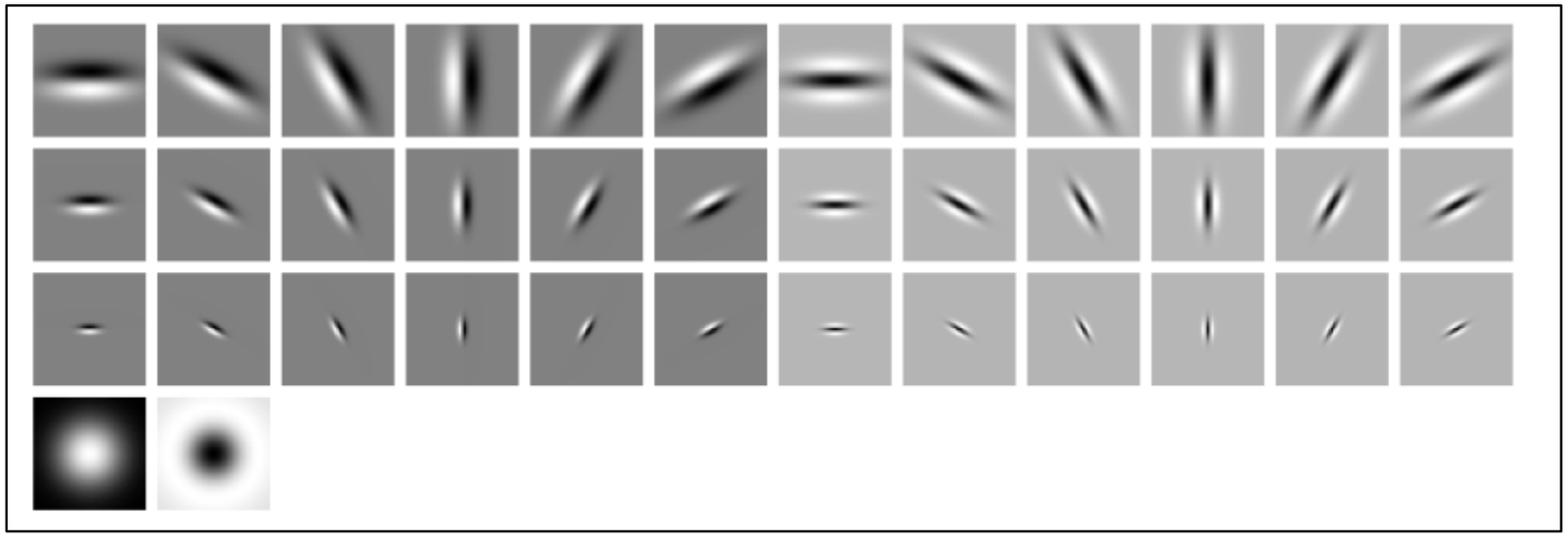


‘S’

Gaussian derivatives at different scales and orientations

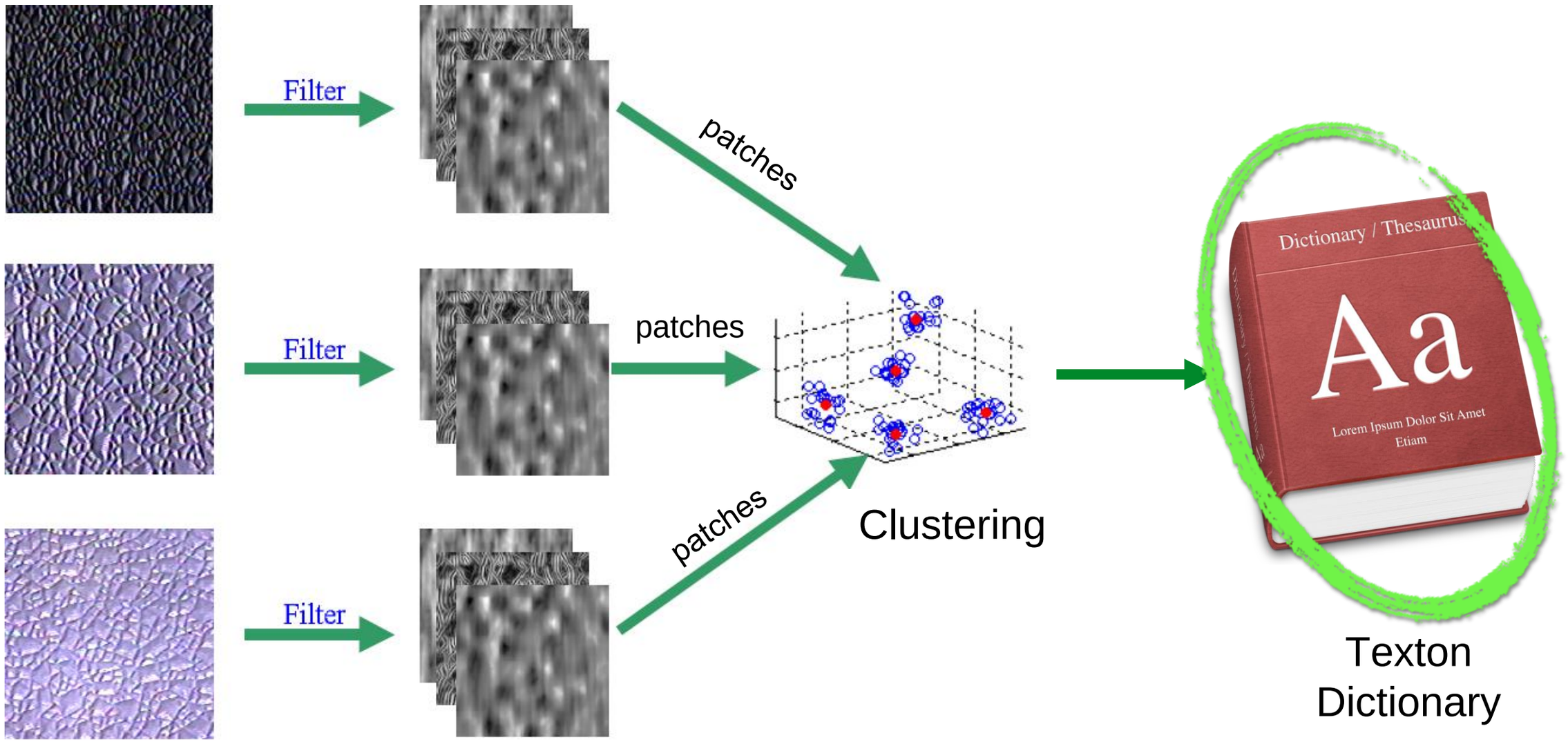


‘LM’



‘MR8’

Learning Textons from data

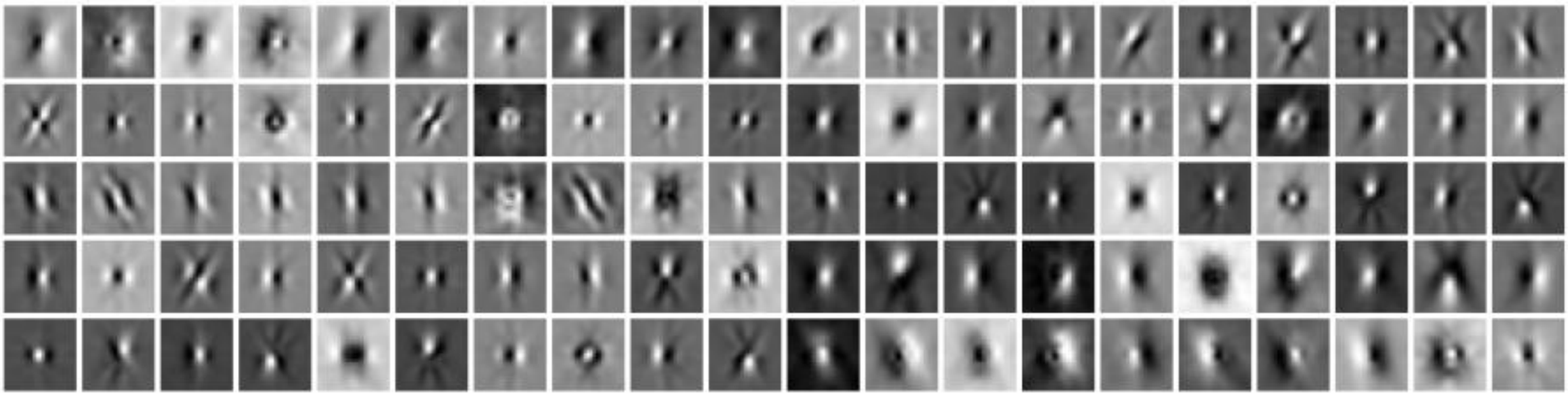


Multiple training images of the same texture

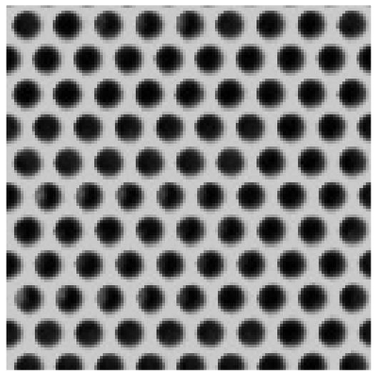
Filter response over a bank of filters

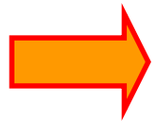
We will learn more about clustering later in class (Bag of Words lecture).

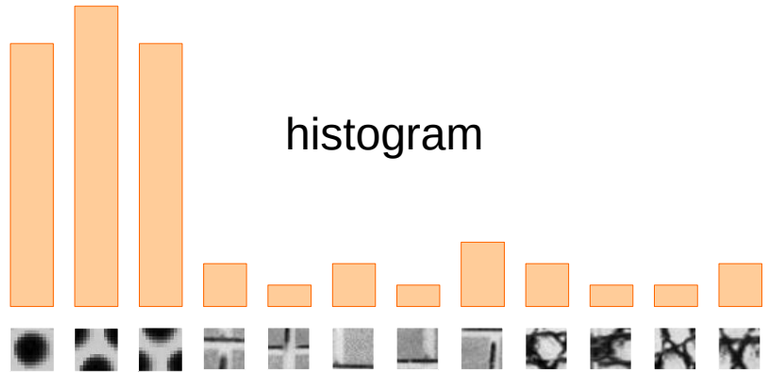
Texton Dictionary



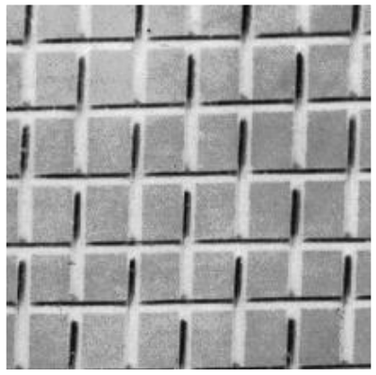
Malik, Belongie, Shi, Leung. Textons, Contours and Regions: Cue Integration in Image Segmentation. ICCV 1999.

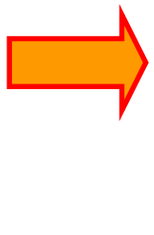


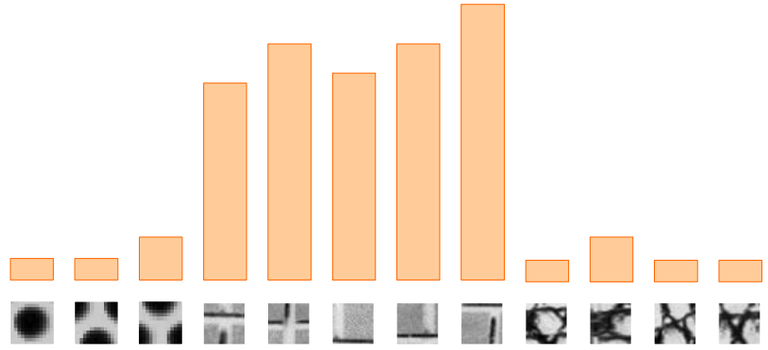


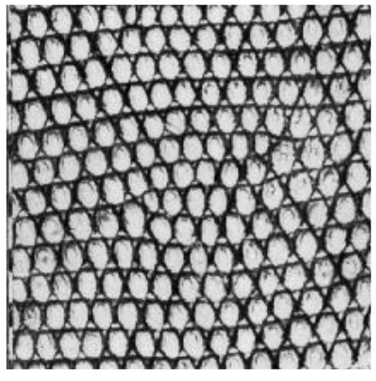


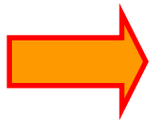
Universal texton dictionary

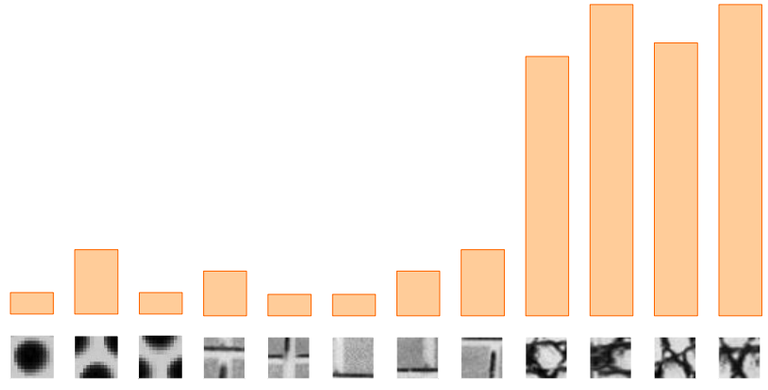










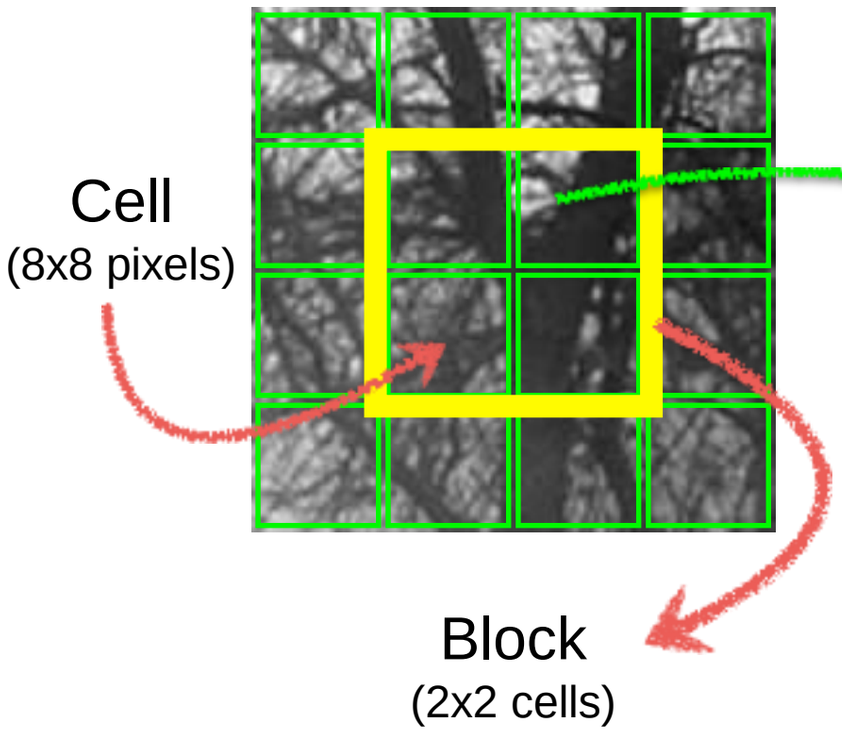


Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001; Mori, Belongie & Malik, 2001; Schmid 2001; Varma & Zisserman, 2002, 2003; Lazebnik, Schmid & Ponce, 2003

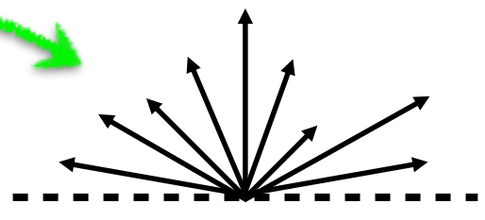
HOG descriptor

HOG

Dalal, Triggs. **Histograms of Oriented Gradients** for Human Detection. CVPR, 2005



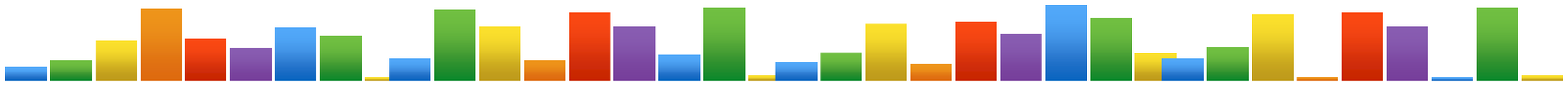
histogram of
‘unsigned’ gradients



soft binning

gradient magnitude histogram
(one for each cell)

Concatenate and L-2 normalization

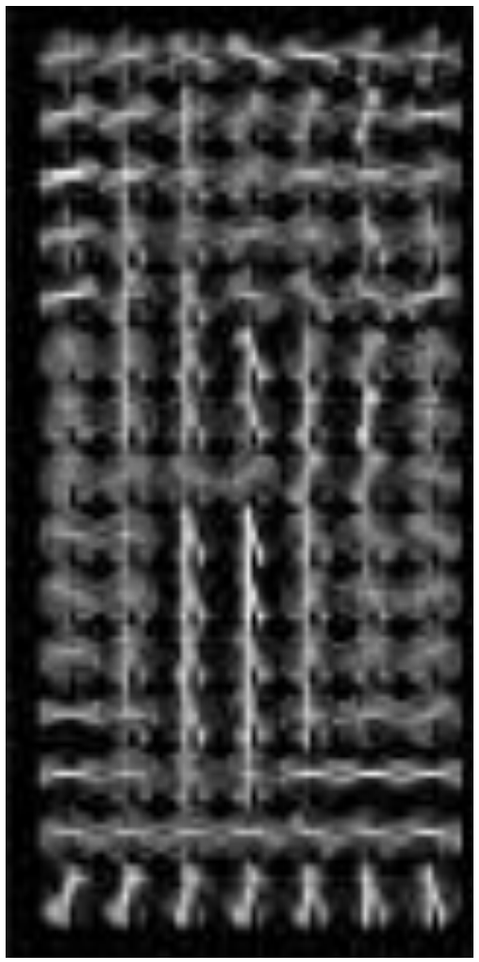


Single scale, no dominant orientation

Pedestrian detection

1 cell step size

visualization



128 pixels
16 cells
15 blocks

15 x 7 x 4 x 36 =
3780

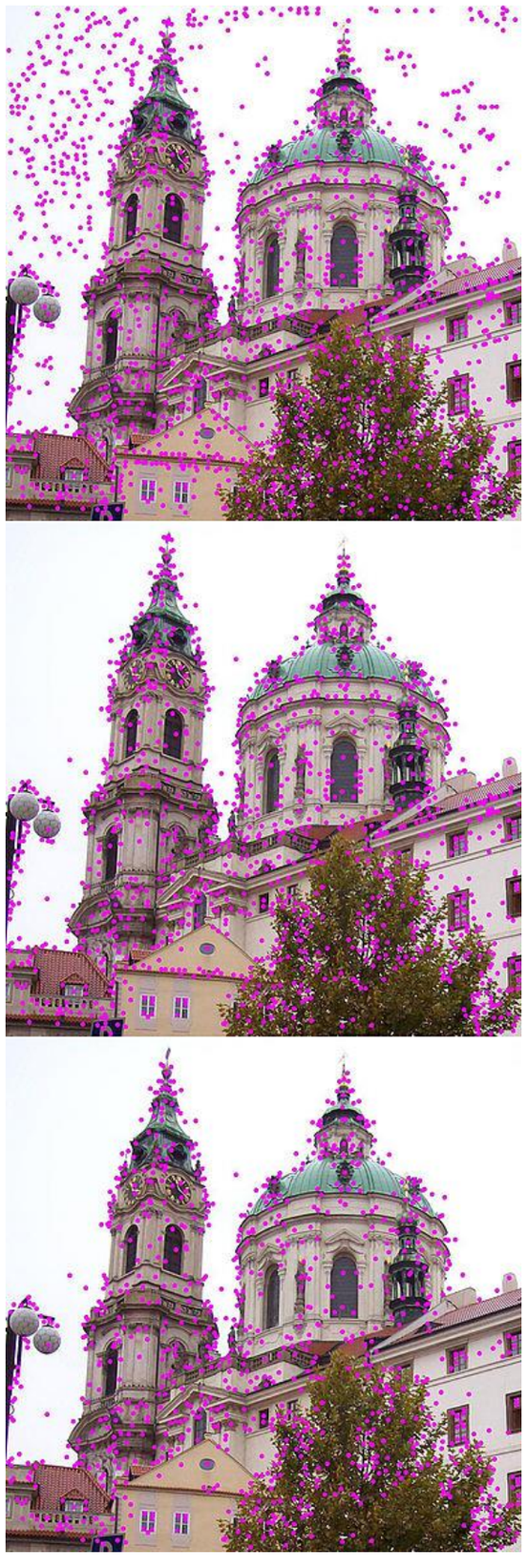
64 pixels
8 cells
7 blocks

Redundant representation due to overlapping blocks

How many times is each inner cell encoded?

http://chrisjmccormick.wordpress.com/2013/05/09/hog-person-detector-tutorial/

SIFT



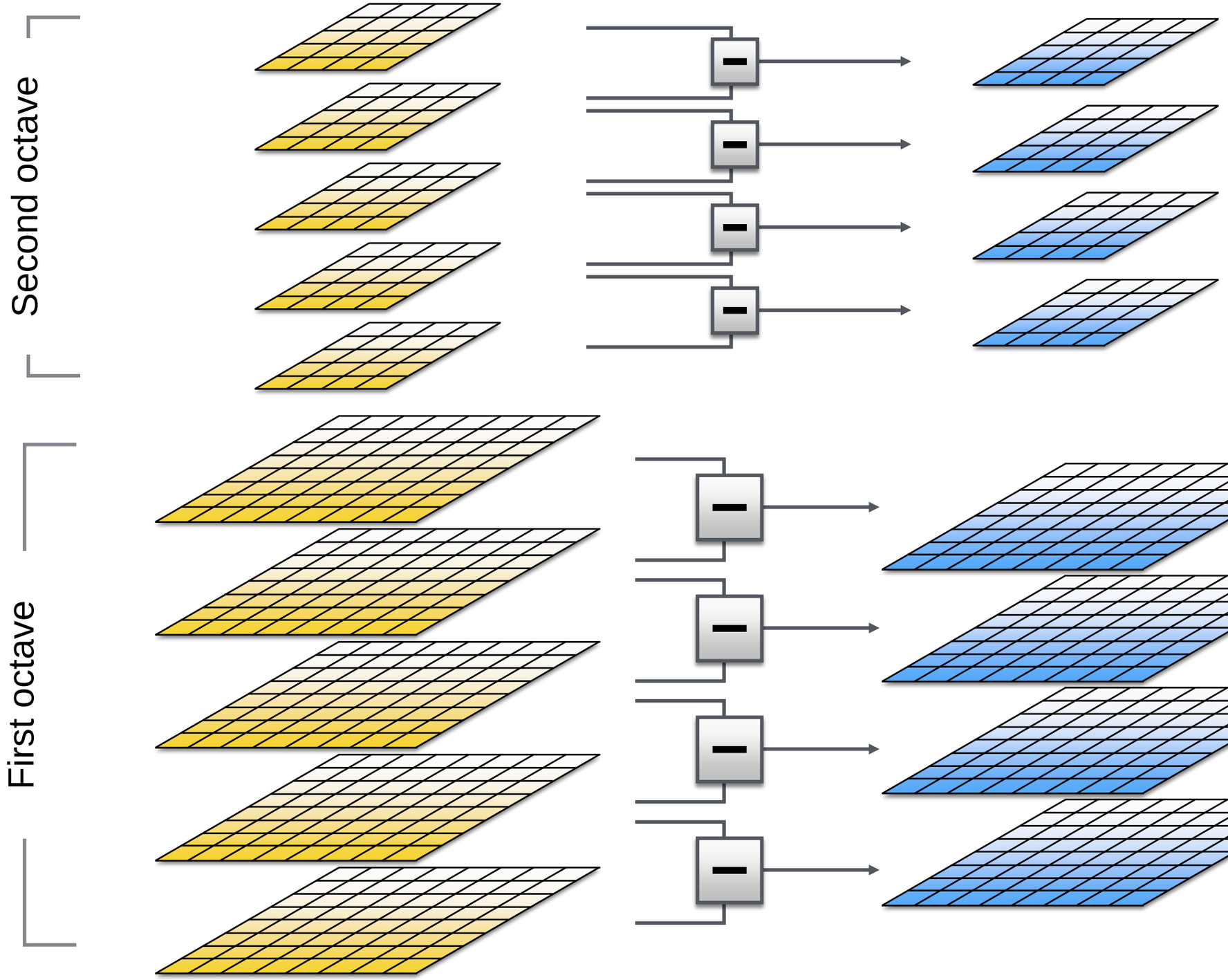
SIFT

(Scale Invariant Feature Transform)

SIFT describes both a **detector** and **descriptor**

1. Multi-scale extrema detection
2. Keypoint localization
3. Orientation assignment
4. Keypoint descriptor

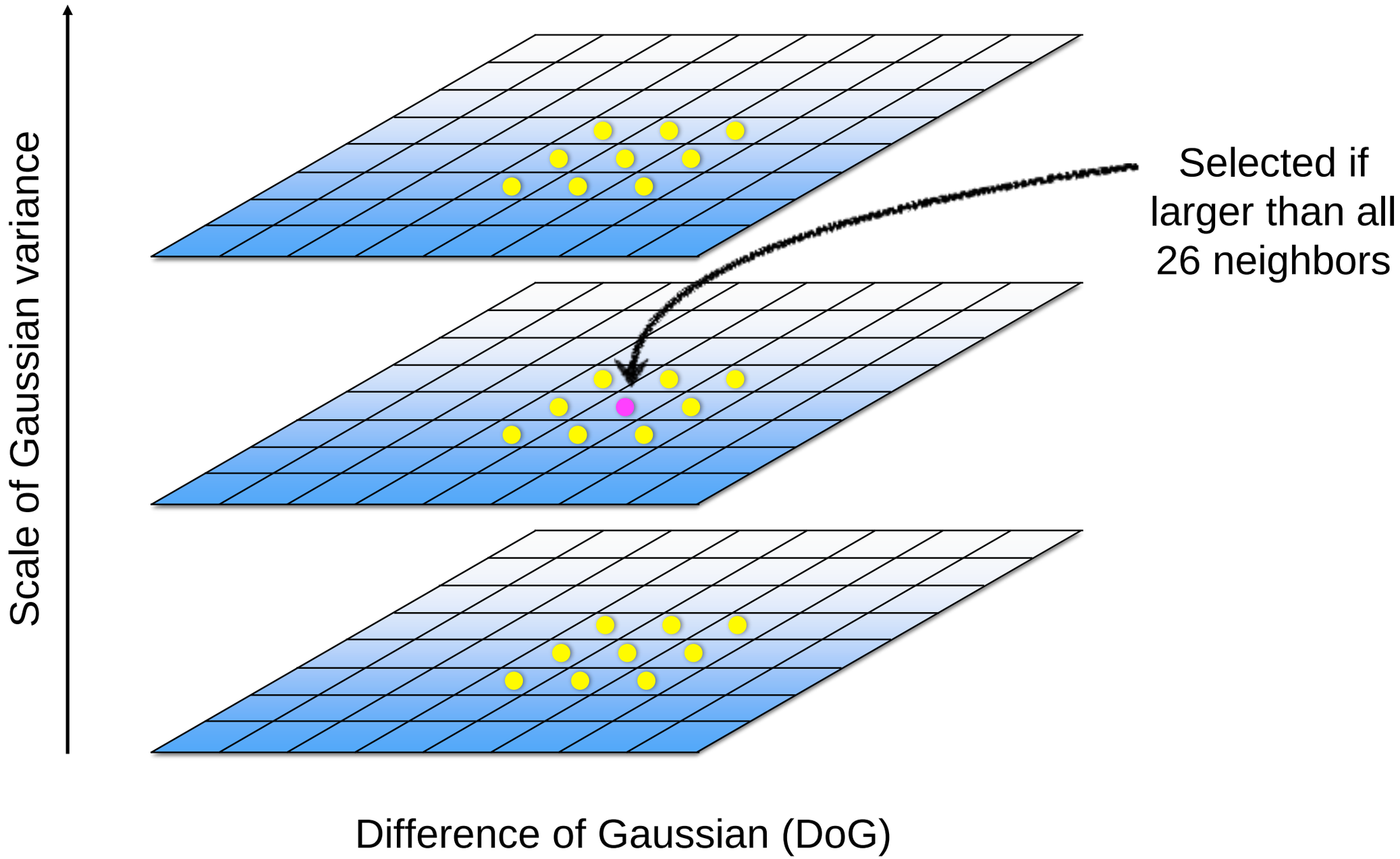
1. Multi-scale extrema detection



Gaussian

Laplacian

Scale-space extrema



2. Keypoint localization

2nd order Taylor series approximation of DoG scale-space

$$f(\mathbf{x}) = f + \frac{\partial f}{\partial \mathbf{x}}^T \mathbf{x} + \frac{1}{2}\mathbf{x}^T \frac{\partial^2 f}{\partial \mathbf{x}^2}\mathbf{x}$$

$$\mathbf{x} = \{x, y, \sigma\}$$

Take the derivative and solve for extrema

$$\mathbf{x}_m = -\frac{\partial^2 f}{\partial \mathbf{x}^2}^{-1} \frac{\partial f}{\partial \mathbf{x}}$$

Additional tests to retain only strong features

3. Orientation assignment

For a keypoint, **L** is the **Gaussian-smoothed** image with the closest scale,

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2}$$

x-derivative **y-derivative**

$$\theta(x,y) = \tan^{-1}((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y)))$$

Detection process returns

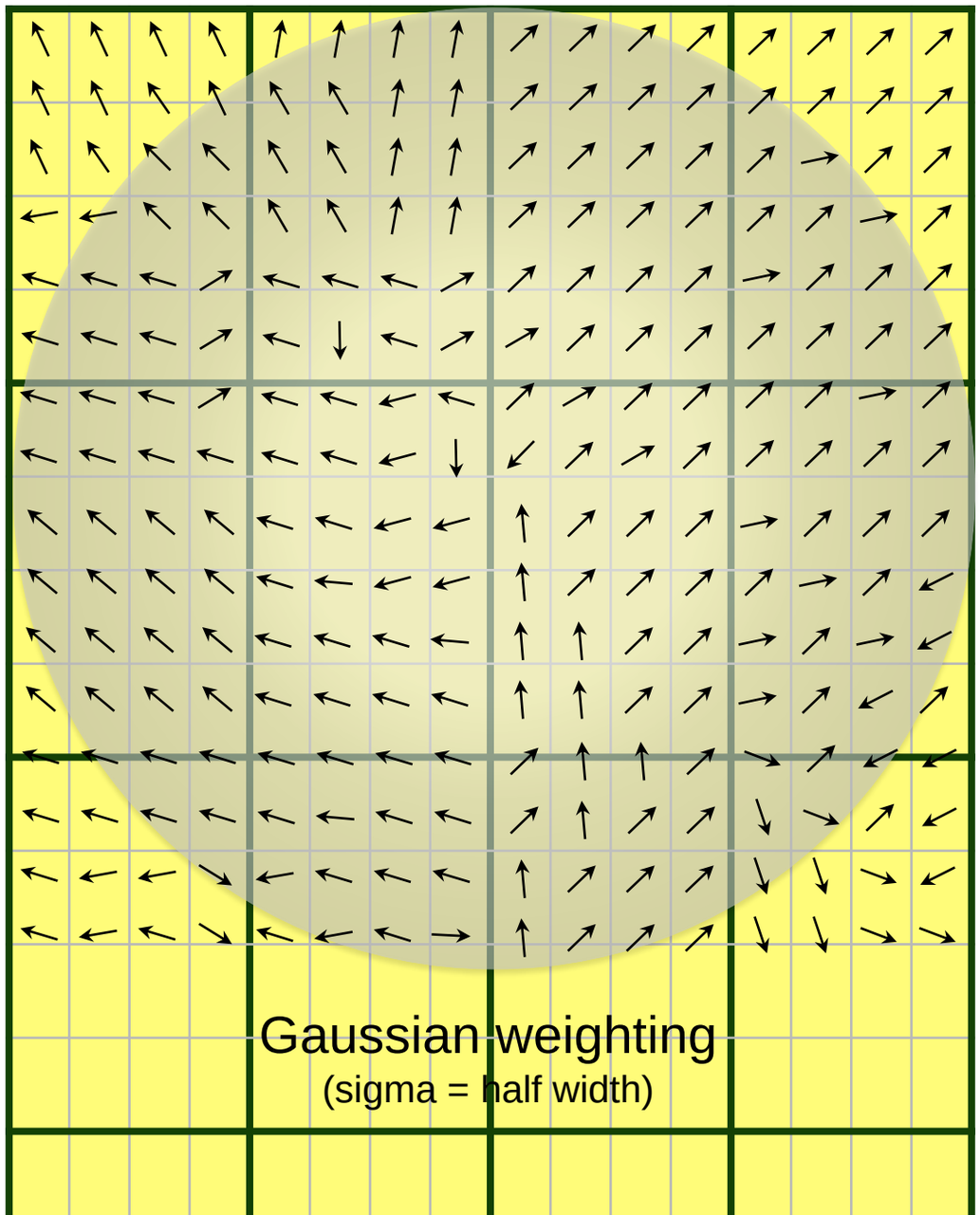
$$\{x, y, \sigma, \theta\}$$

location scale orientation

4. Keypoint descriptor

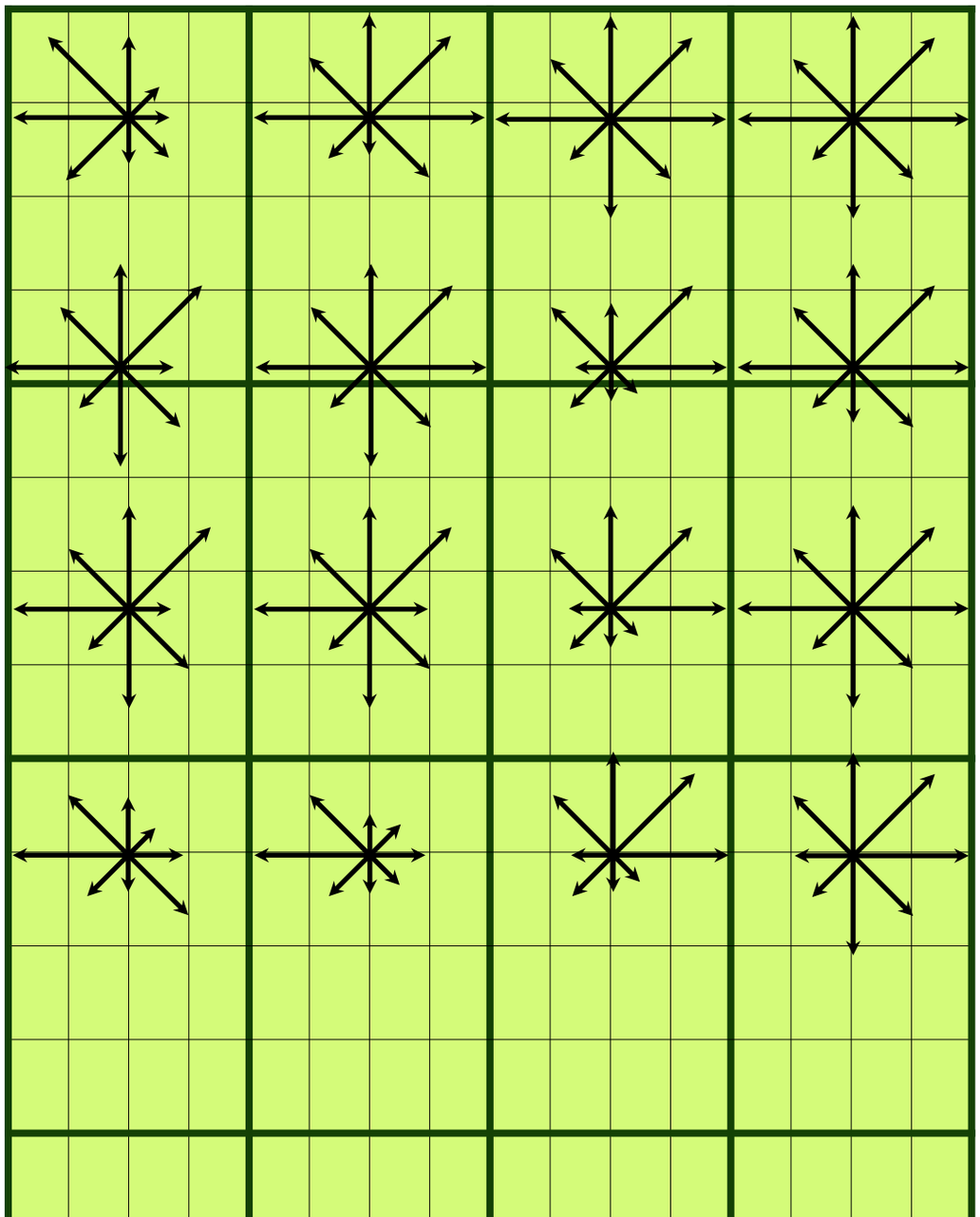
Image Gradients

(4 x 4 pixel per cell, 4 x 4 cells)



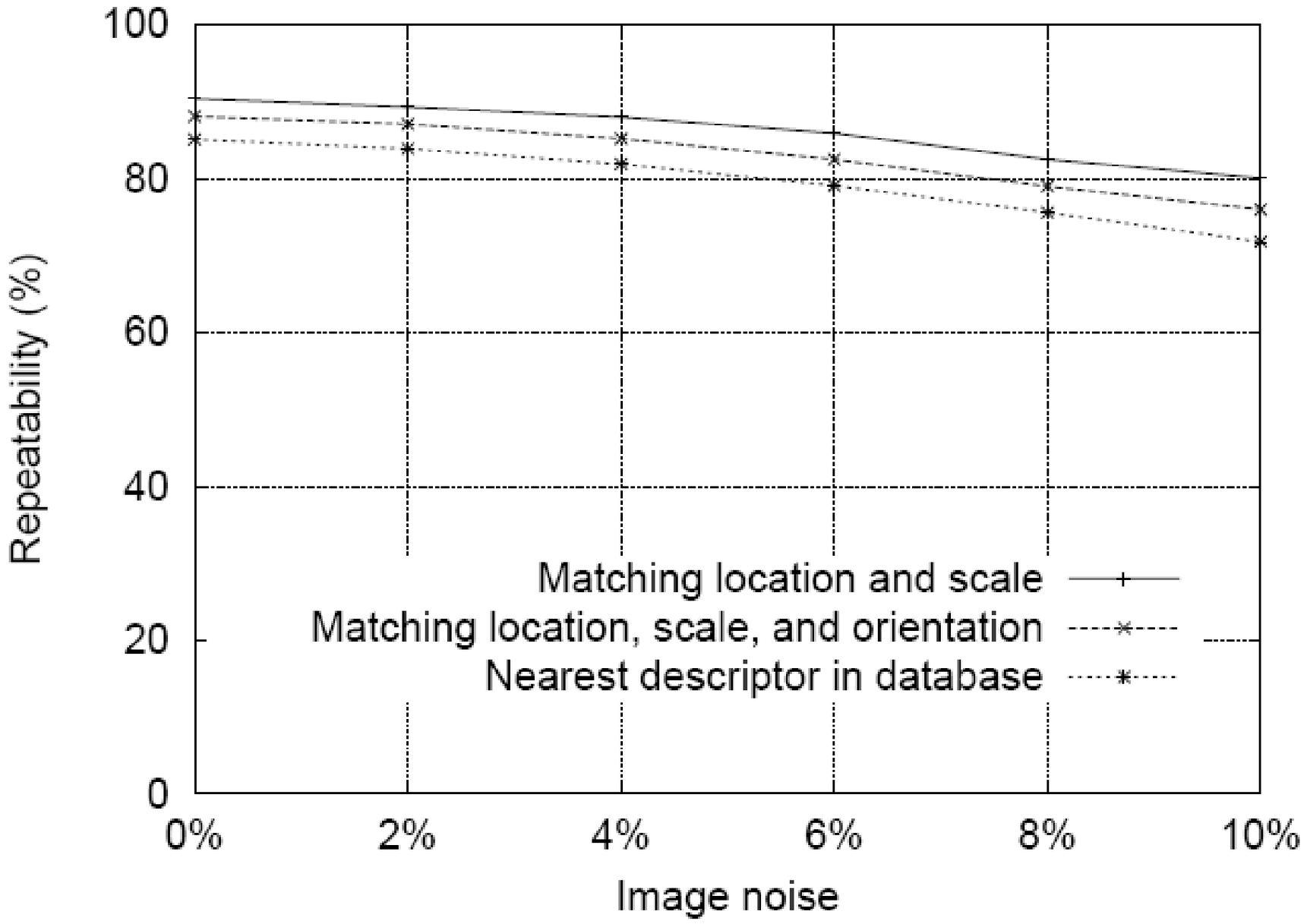
SIFT descriptor

(16 cells x 8 directions = 128 dims)



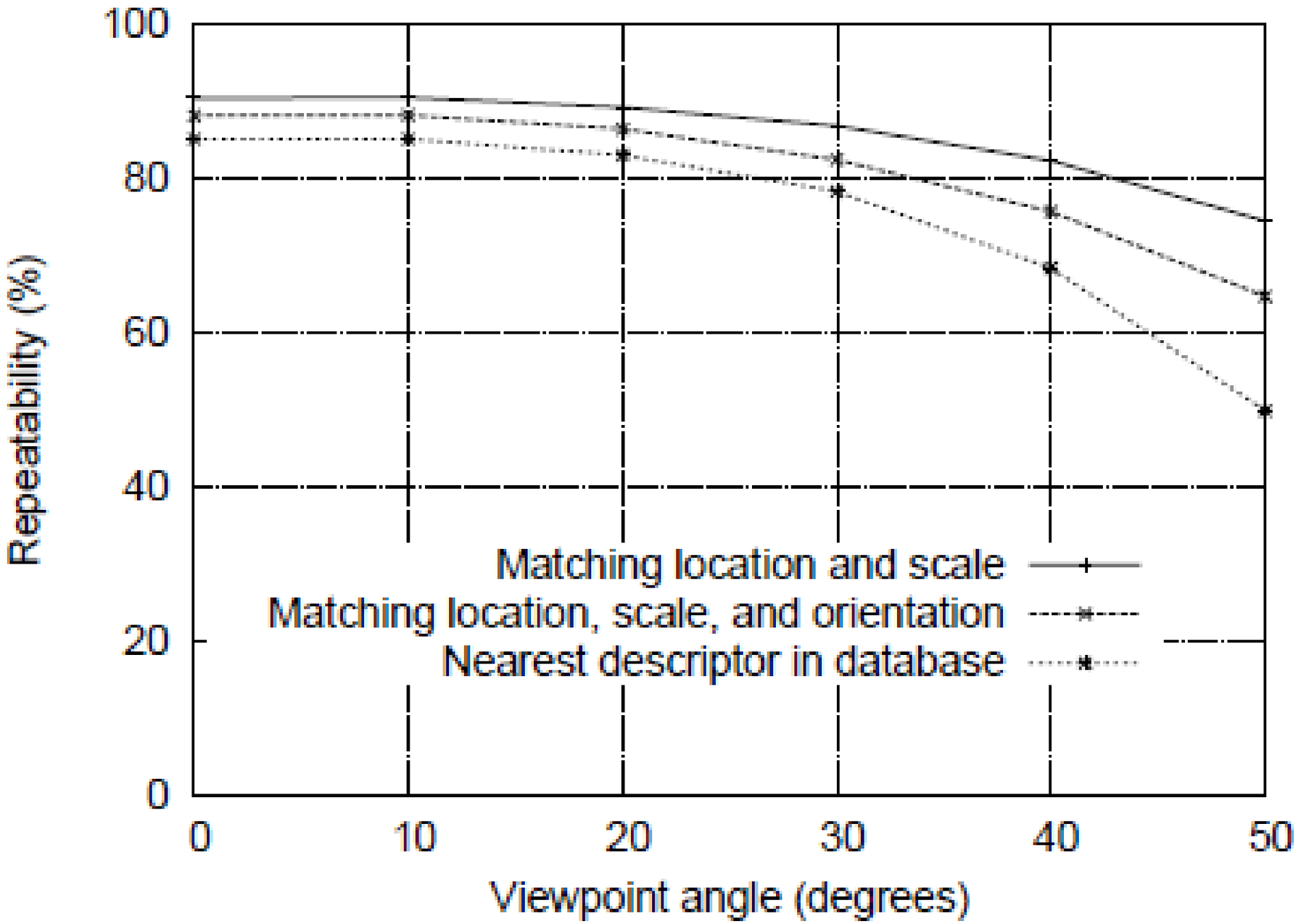
HOW GOOD IS SIFT?

SIFT Repeatability



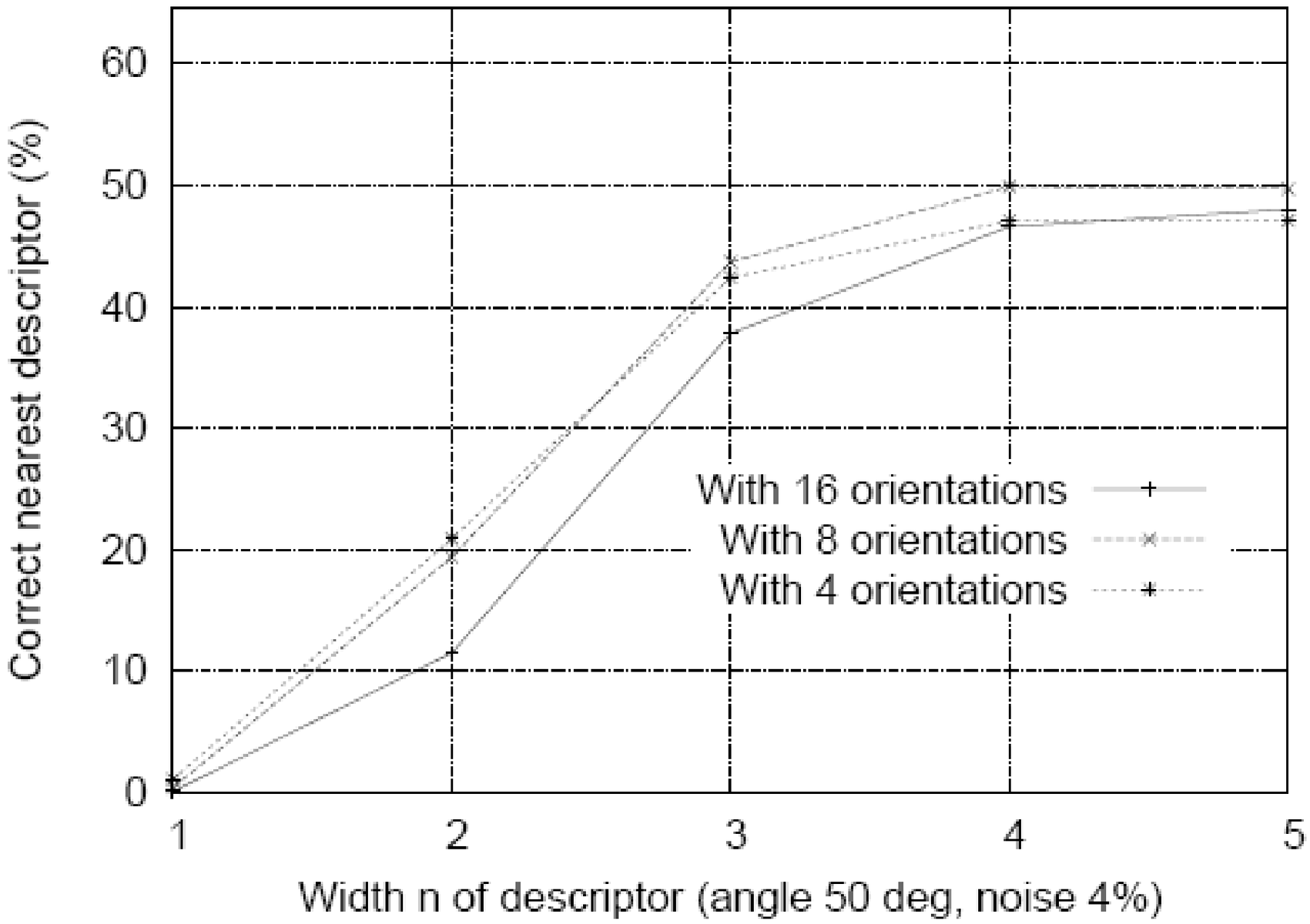
Lowe IJCV 2004

SIFT Repeatability



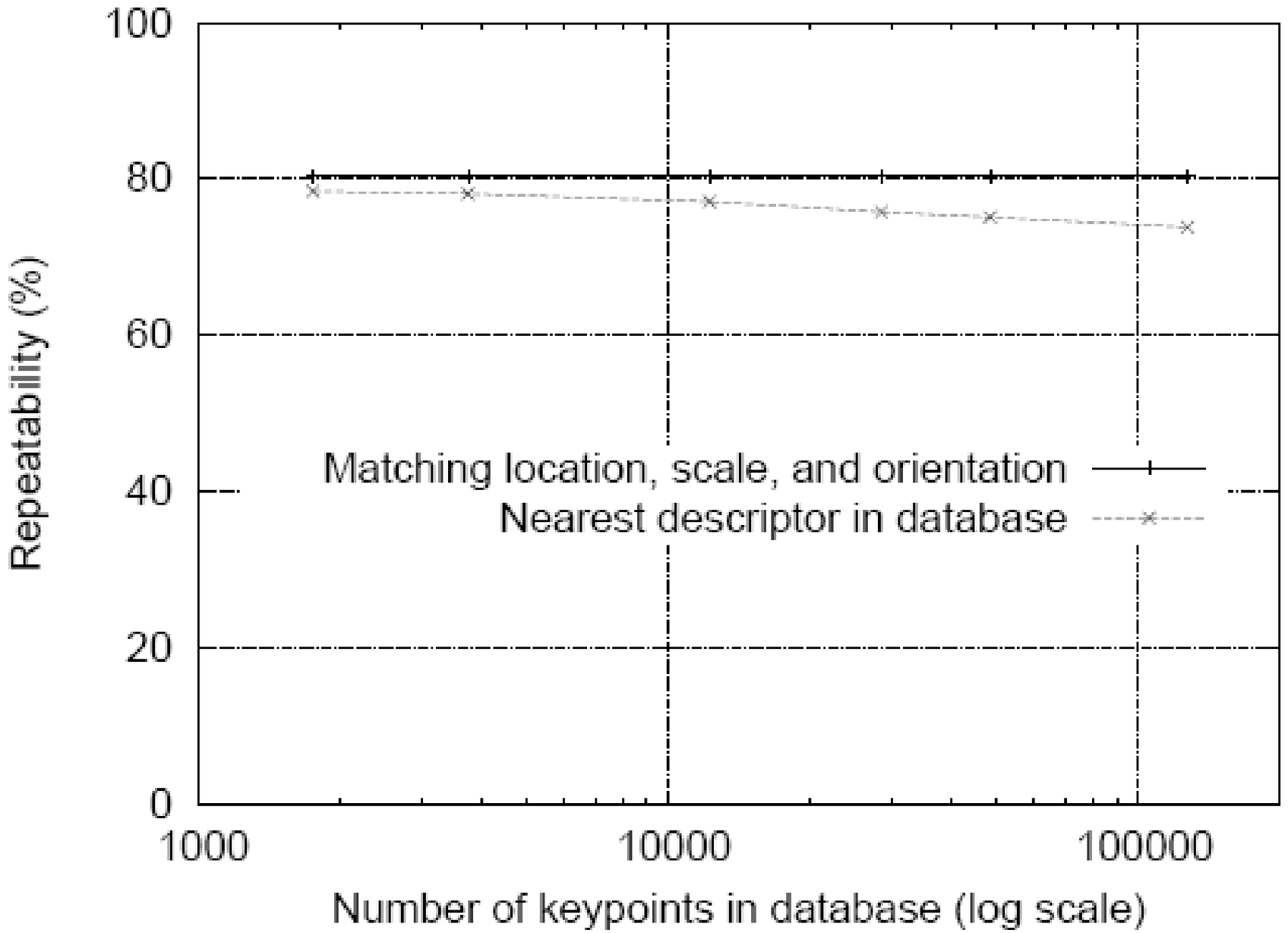
Lowe IJCV 2004

SIFT Repeatability



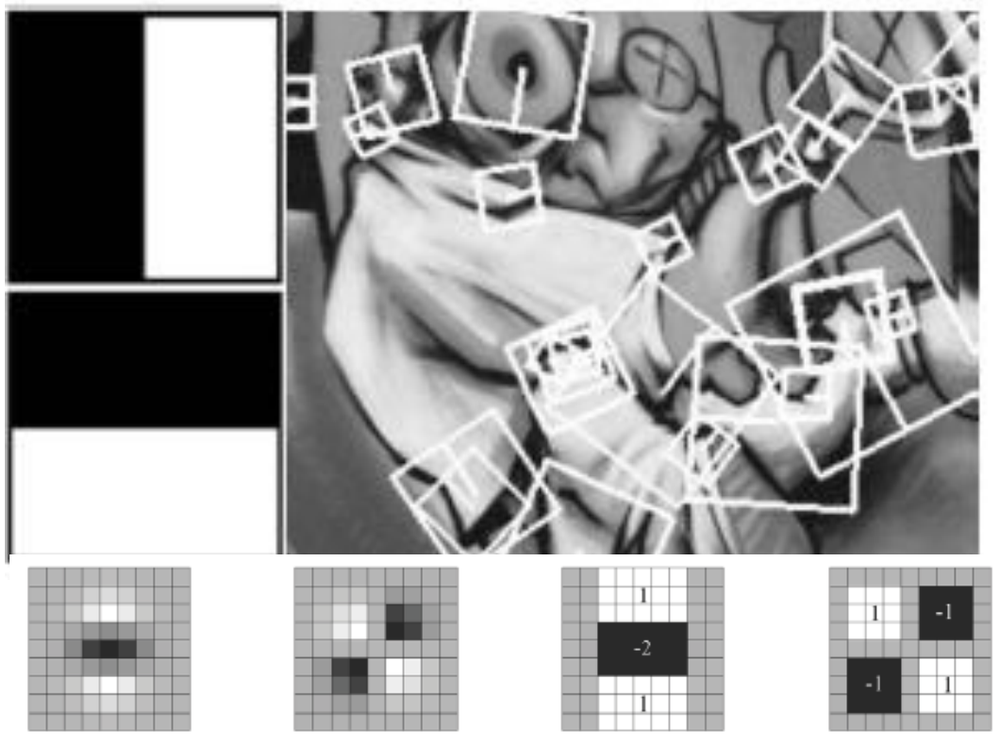
Lowe IJCV 2004

SIFT Repeatability



Lowe IJCV 2004

Local Descriptors: SURF



Fast approximation of SIFT idea

- **Efficient computation by 2D box filters & integral images**
  ⇒ 6 times faster than SIFT
- **Equivalent quality for object identification**

GPU implementation available

- **Feature extraction @ 200Hz (detector + descriptor, 640×480 img)**
- http://www.vision.ee.ethz.ch/~surf

[Bay, ECCV'06], [Cornelis, CVGPU'08]

SURF: Speeded Up Robust Features

Speed-up computations by fast approximation of (i) Hessian matrix and (ii) descriptor using "integral images".

$$\mathcal{H}(\mathbf{x}, \sigma) = \begin{bmatrix} L_{xx}(\mathbf{x}, \sigma) & L_{xy}(\mathbf{x}, \sigma) \\ L_{xy}(\mathbf{x}, \sigma) & L_{yy}(\mathbf{x}, \sigma) \end{bmatrix},$$

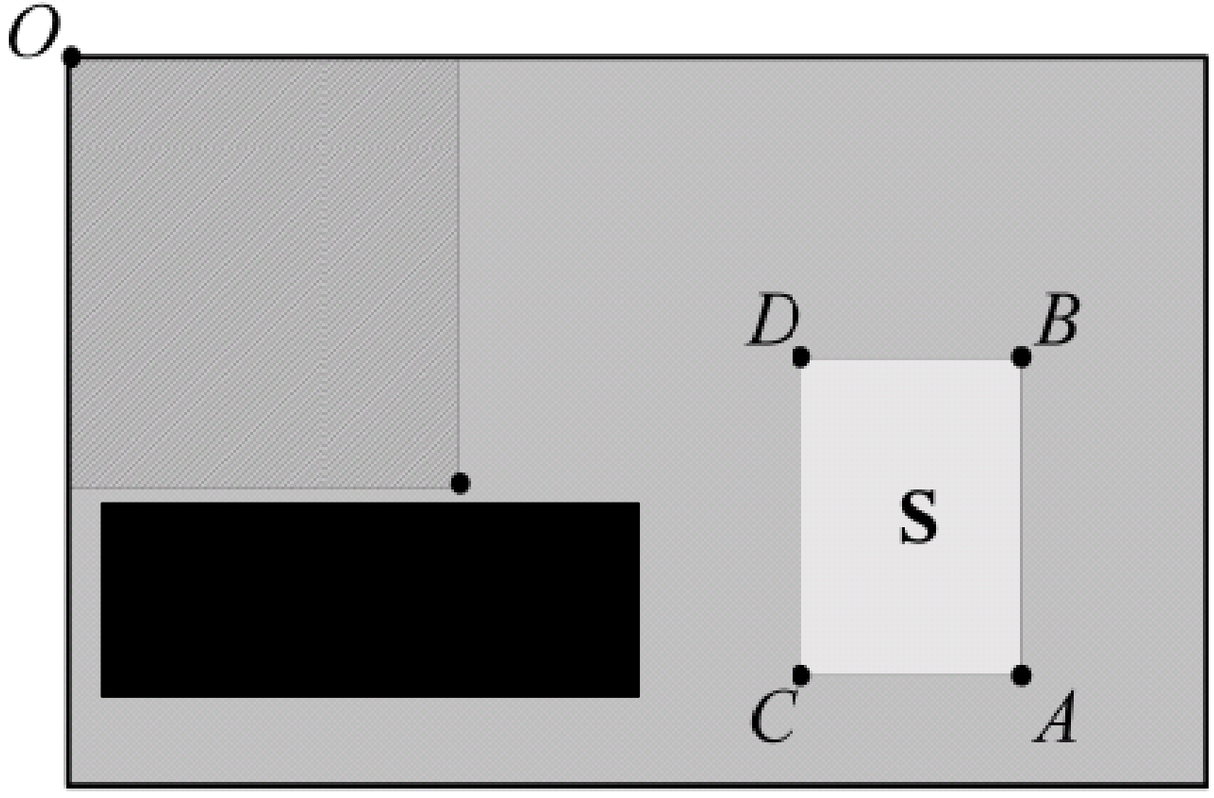
What is an "integral image"?

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "SURF: Speeded Up Robust Features", **European Computer Vision Conference (ECCV),** 2006.

Integral Image

- The integral image $I_{\Sigma}(x,y)$ of an image I(x, y) represents the sum of all pixels in I(x,y) of a rectangular region formed by (0,0) and (x,y).

- .



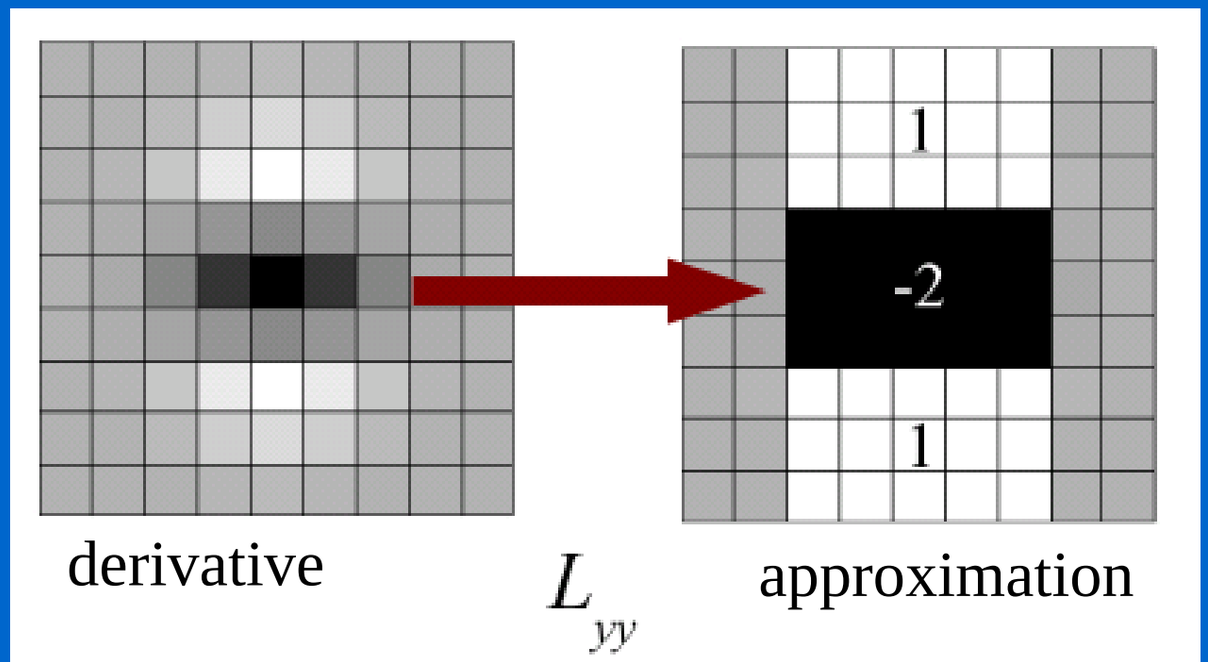
Using integral images, it takes only four array references to calculate the sum of pixels over a rectangular region of any size.

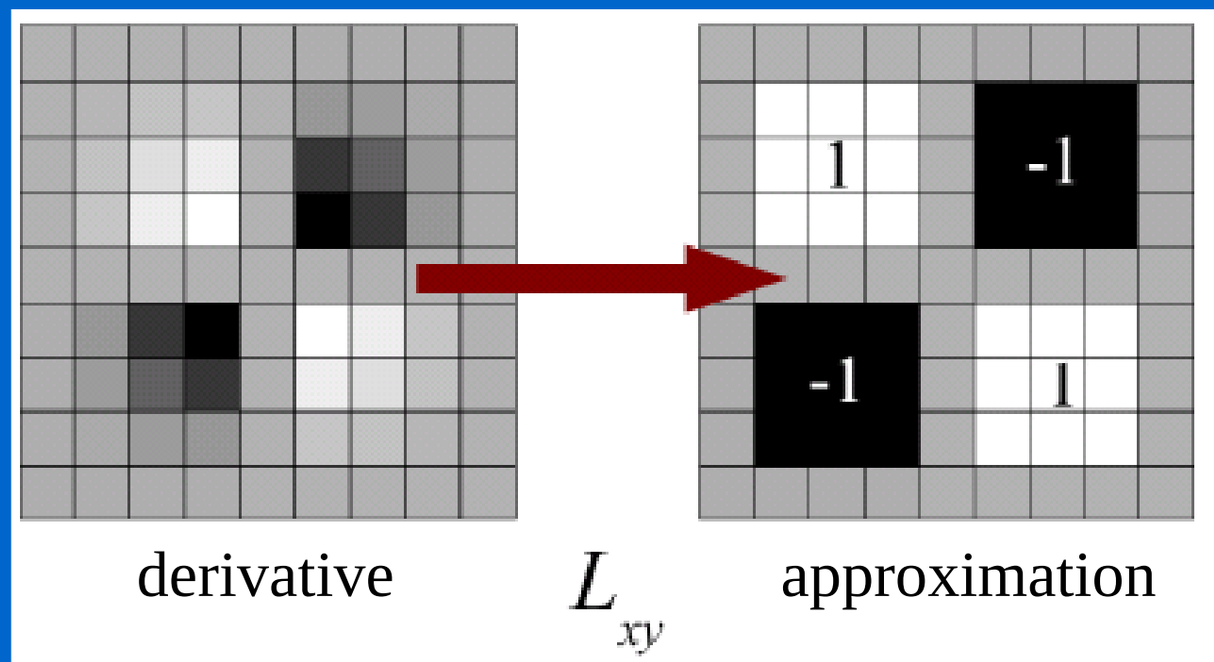
$$\mathbf{S}=A-B-C+D$$

SURF: Speeded Up Robust Features (cont'd)

- Approximate L_{xx}, L_{yy}, and L_{xy} using box filters.

(box filters shown are 9 x 9 – good approximations for a Gaussian with σ=1.2)

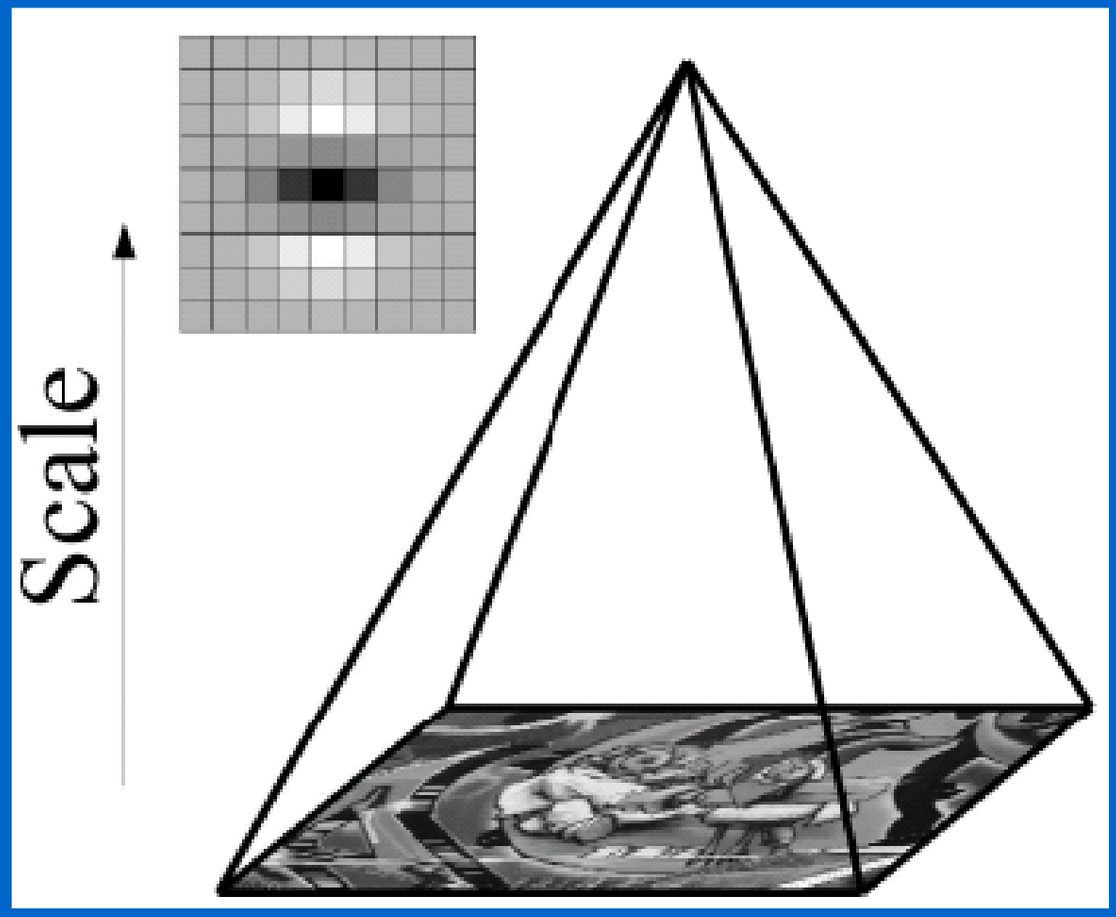




- Can be computed very fast using integral images!

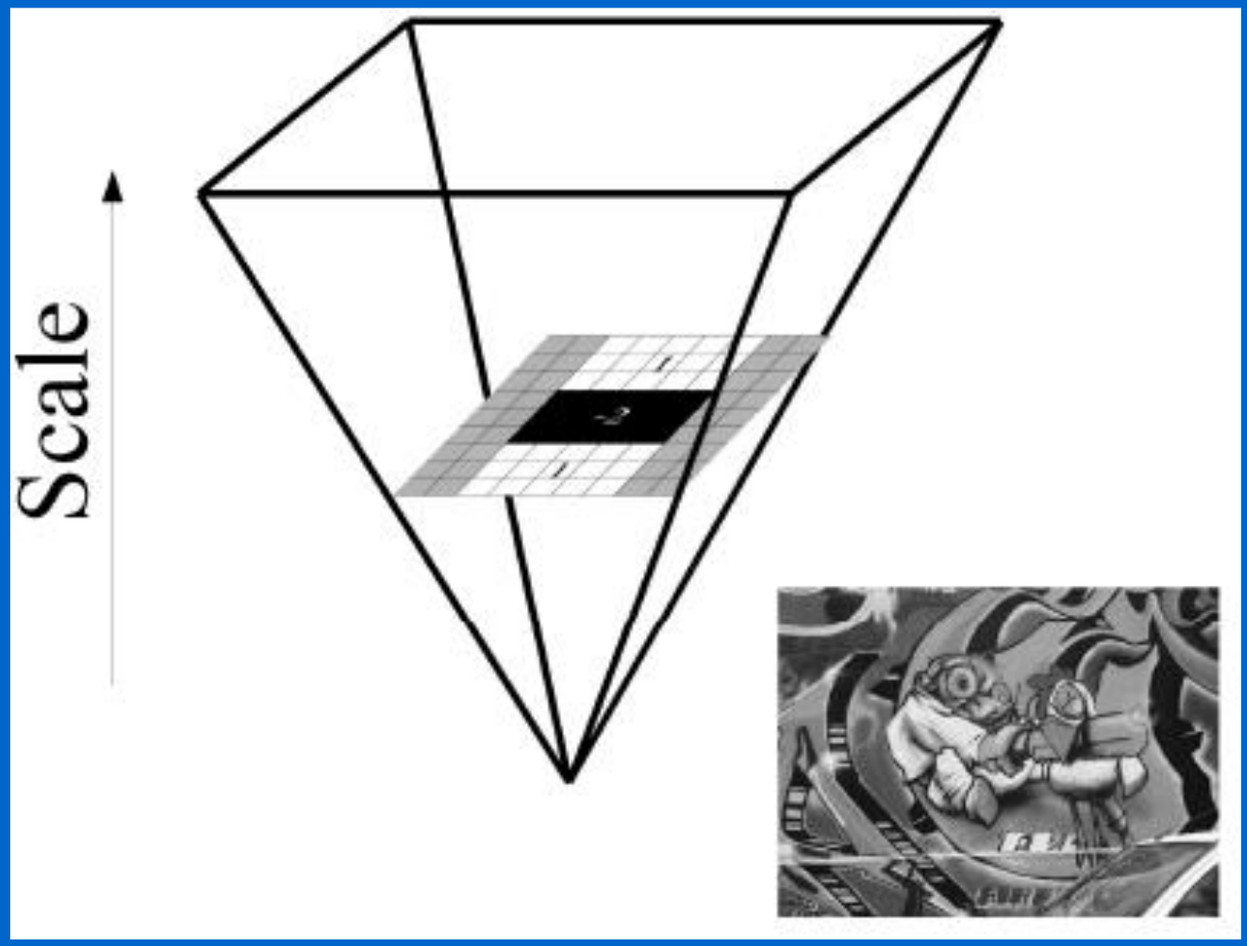
SURF: Speeded Up Robust Features (cont'd)

- In SIFT, images are repeatedly smoothed with a Gaussian and subsequently sub-sampled in order to achieve a higher level of the pyramid.



SURF: Speeded Up Robust Features (cont'd)

- Alternatively, we can use filters of larger size on the original image.



- Due to using integral images, filters of any size can be applied at exactly the same speed!

(see Tuytelaars' paper for details)

SURF: Speeded Up Robust Features (cont'd)

- Approximation of H:

Using DoG

$$\mathcal{H}(\mathbf{x}, \sigma) = \begin{bmatrix} L_{xx}(\mathbf{x}, \sigma) & L_{xy}(\mathbf{x}, \sigma) \\ L_{xy}(\mathbf{x}, \sigma) & L_{yy}(\mathbf{x}, \sigma) \end{bmatrix},$$

$$SIFT: \; H_{approx}^{SIFT} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{bmatrix}$$

Using box filters

$$SURF: \; H_{approx}^{SURF} = \begin{bmatrix} \hat{L}_{xx} & \hat{L}_{xy} \\ \hat{L}_{yx} & \hat{L}_{yy} \end{bmatrix}$$

SURF: Speeded Up Robust Features (cont'd)

- Instead of using a different measure for selecting the location and scale of interest points (e.g., Hessian and DOG in SIFT), SURF uses the determinant of H_{approx}^{SURF} to find both.

- Determinant elements must be weighted to obtain a good approximation:

$$\det(H_{approx}^{SURF}) = \hat{L}_{xx}\hat{L}_{yy} - (0.9\hat{L}_{xy})^2$$

SURF: Speeded Up Robust Features (cont'd)

- Once interest points have been localized both in space and scale, the next steps are:

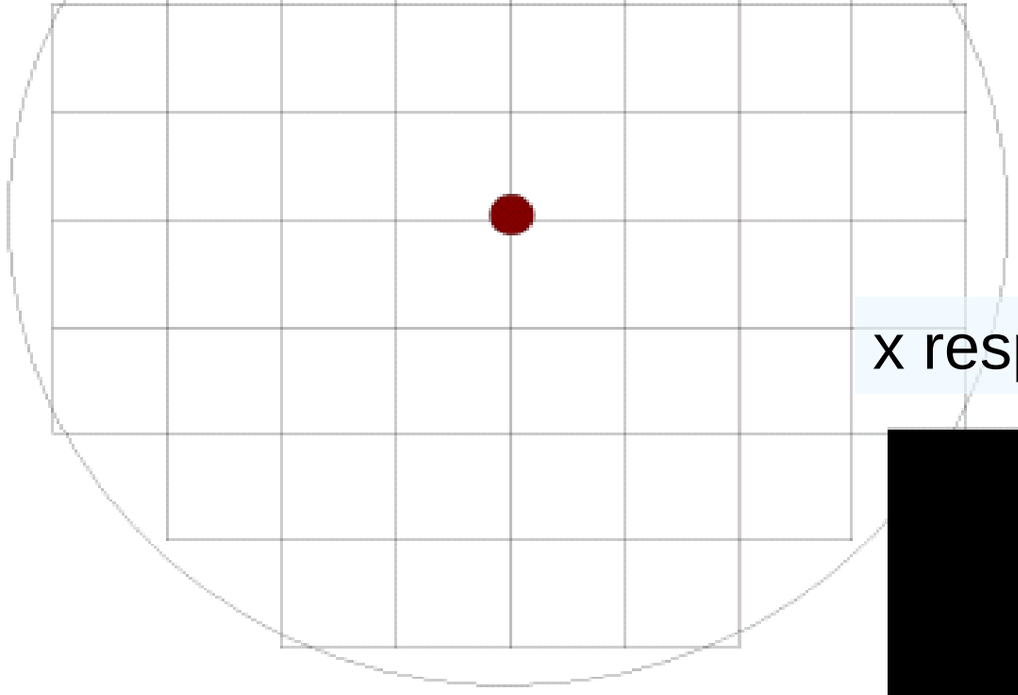
  (1) Orientation assignment

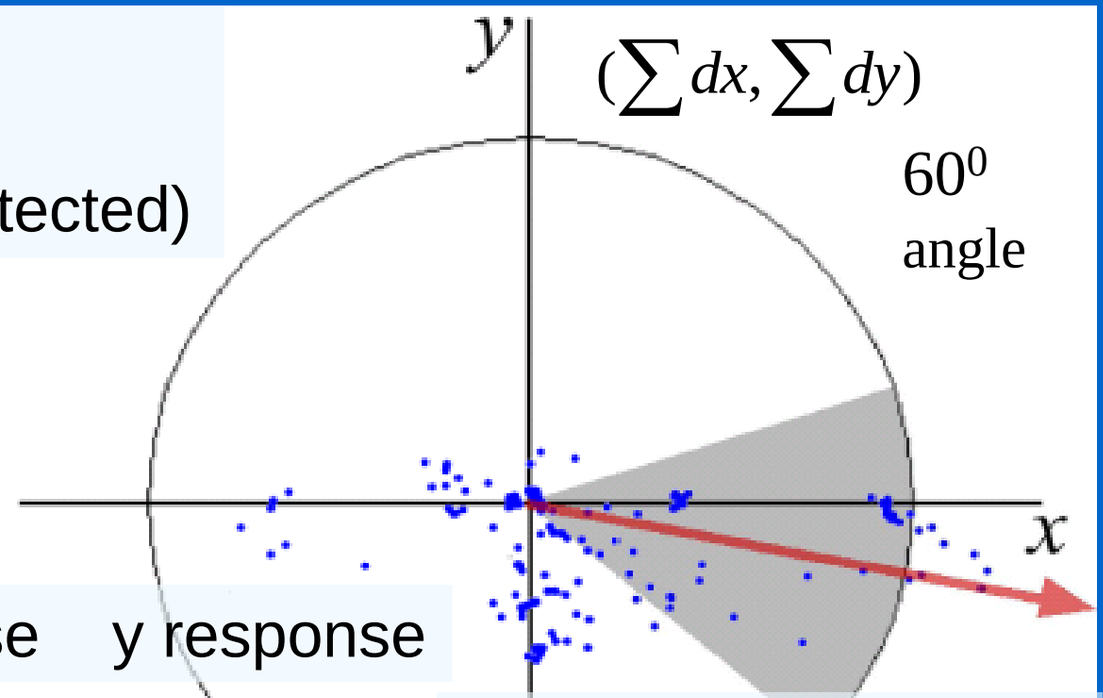
  (2) Keypoint descriptor

SURF: Speeded Up Robust Features (cont'd)

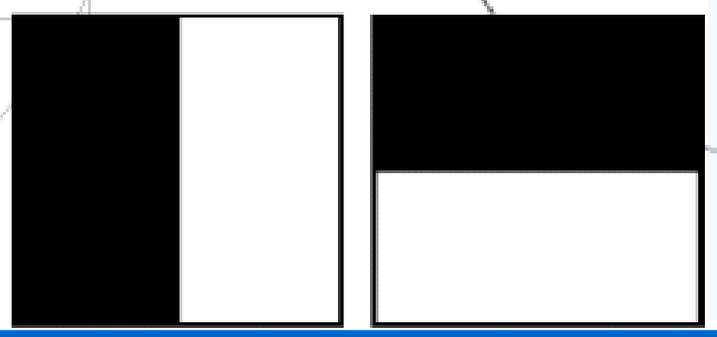
- Orientation assignment

Circular neighborhood of
radius **6σ** around the interest point
(**σ** = the scale at which the point was detected)





x response y response

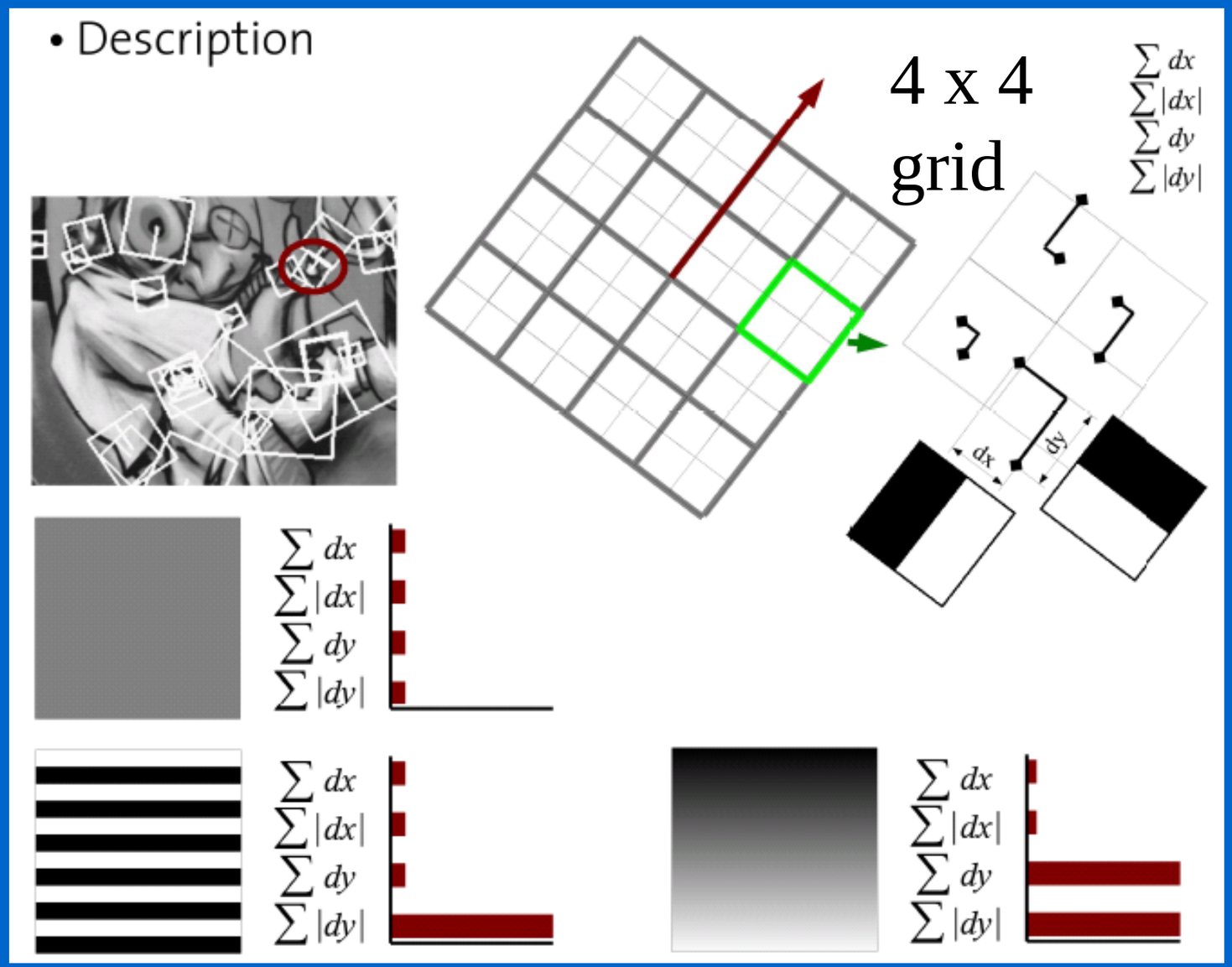


Haar wavelets
(responses weighted
with Gaussian)
Side length = **4σ**

Can be computed very fast using integral images!

SURF: Speeded Up Robust Features (cont'd)

- Keypoint descriptor (square region of size 20σ)

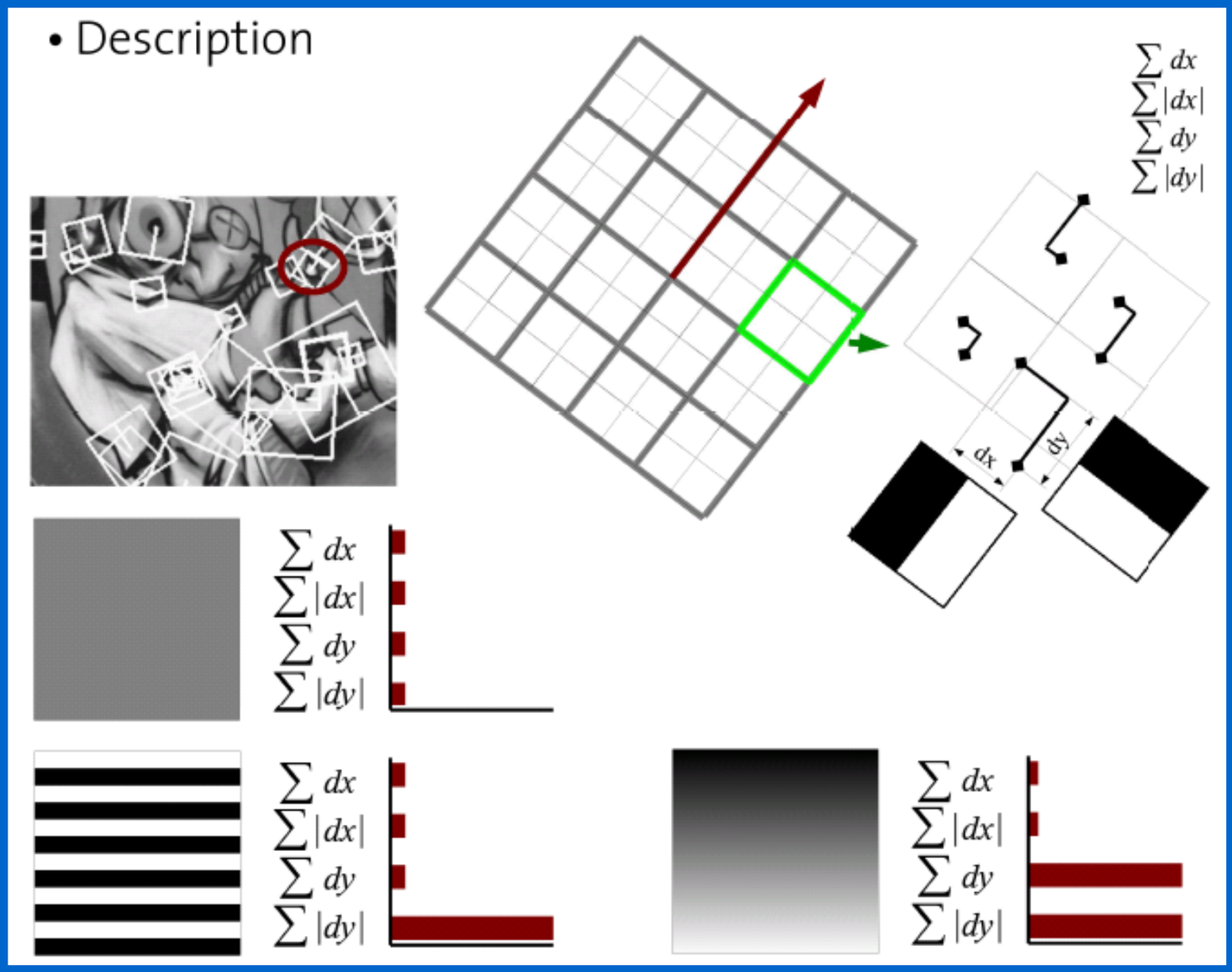


- Sum the response over each sub-region for d_x and d_y separately.
- To bring in information about the polarity of the intensity changes, extract the sum of absolute value of the responses too.

Feature vector size:

4 x 16 = 64

SURF: Speeded Up Robust Features (cont'd)



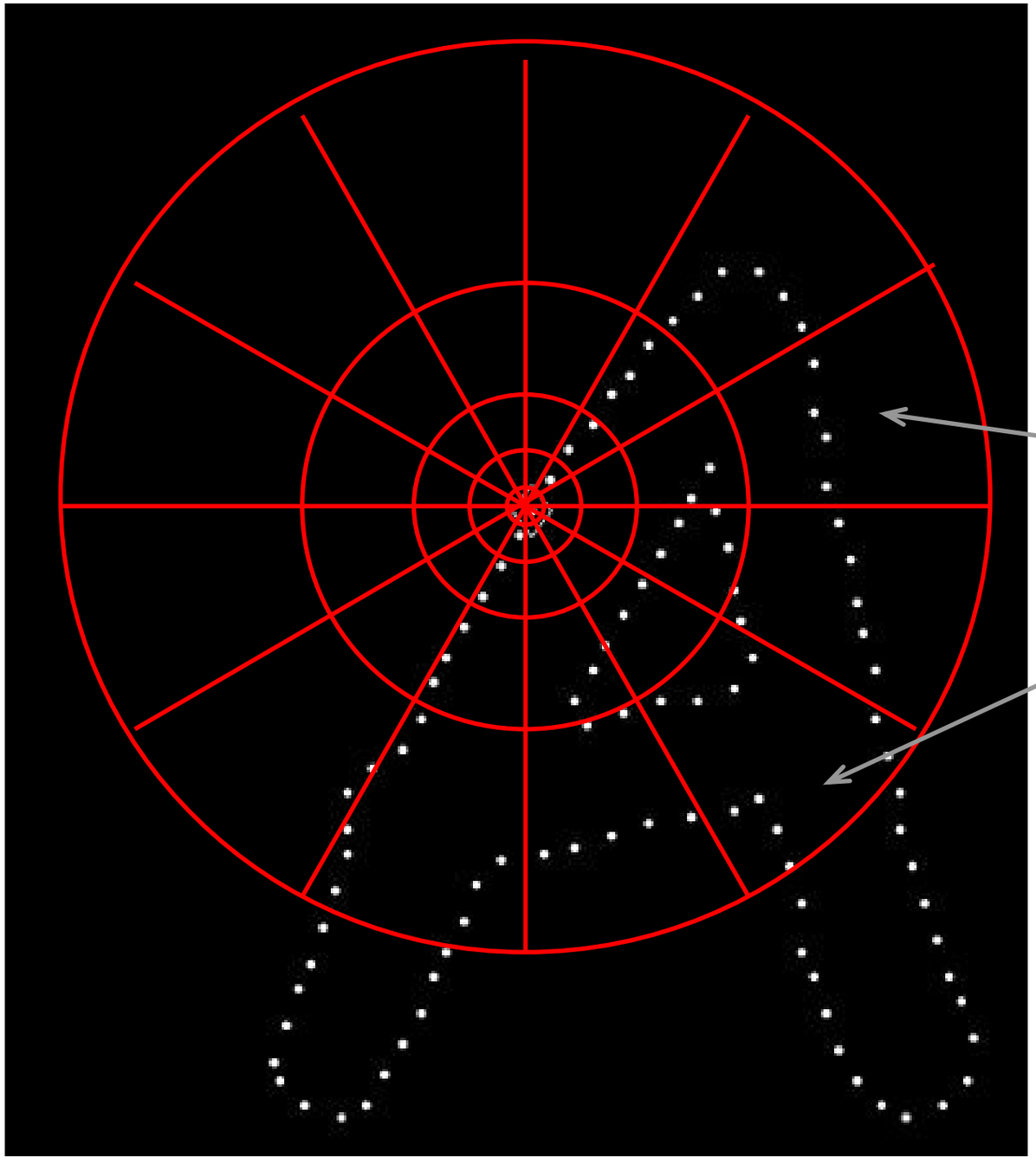
- SURF-128
  - The sum of d_x and $|d_x|$ are computed separately for points where $d_y < 0$ and $d_y > 0$
  - Similarly for the sum of d_y and $|d_y|$
  - **More discriminatory!**

SURF: Speeded Up Robust Features

- Has been reported to be 3 times faster than SIFT.
- Less robust to illumination and viewpoint changes compared to SIFT.

K. Mikolajczyk and C. Schmid,"A Performance Evaluation of Local Descriptors", **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 27, no. 10, pp. 1615-1630, 2005.

Local Descriptors: Shape Context



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

Count = 4

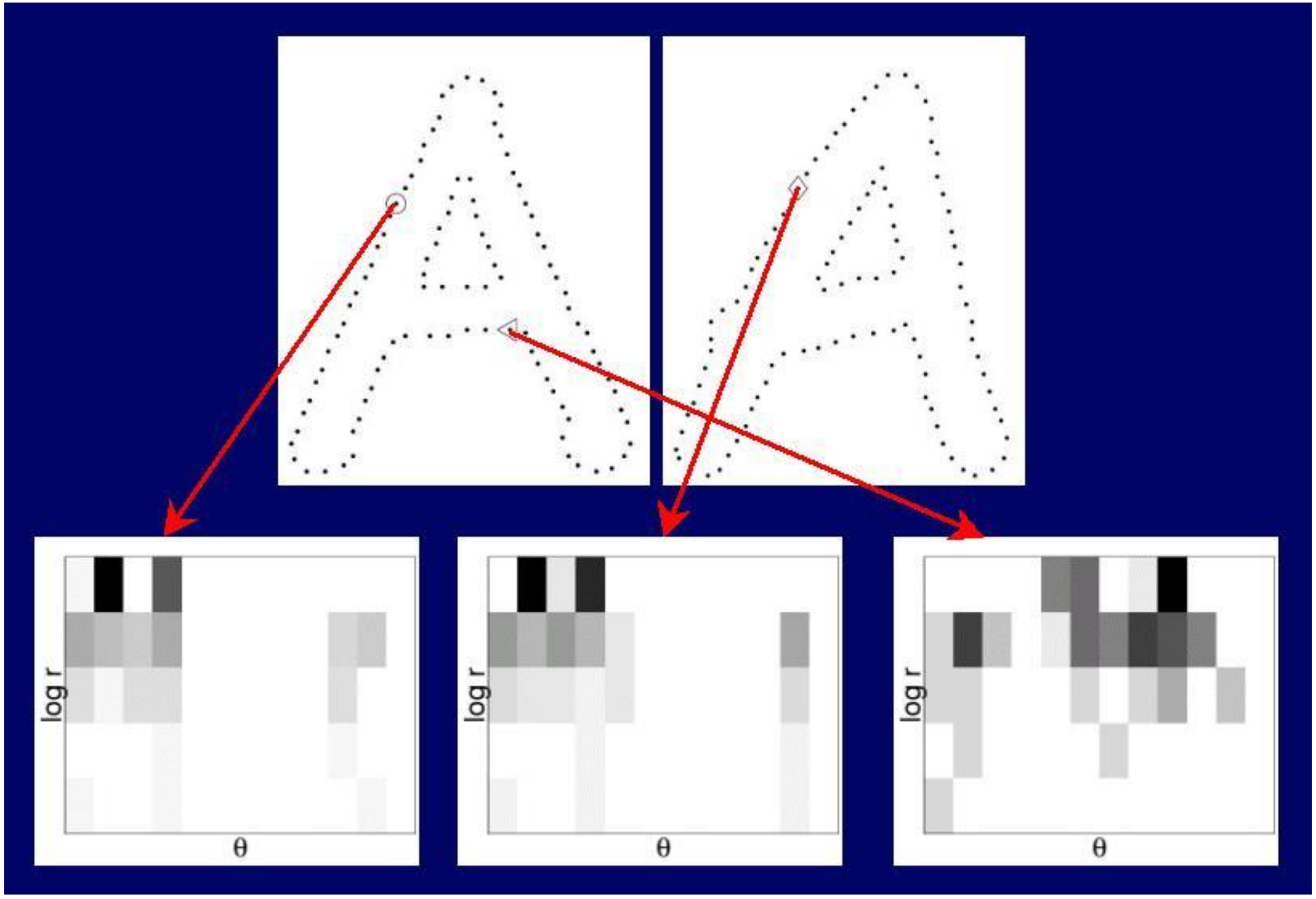
⋮

Count = 10

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

Belongie & Malik, ICCV 2001

Shape Context Descriptor

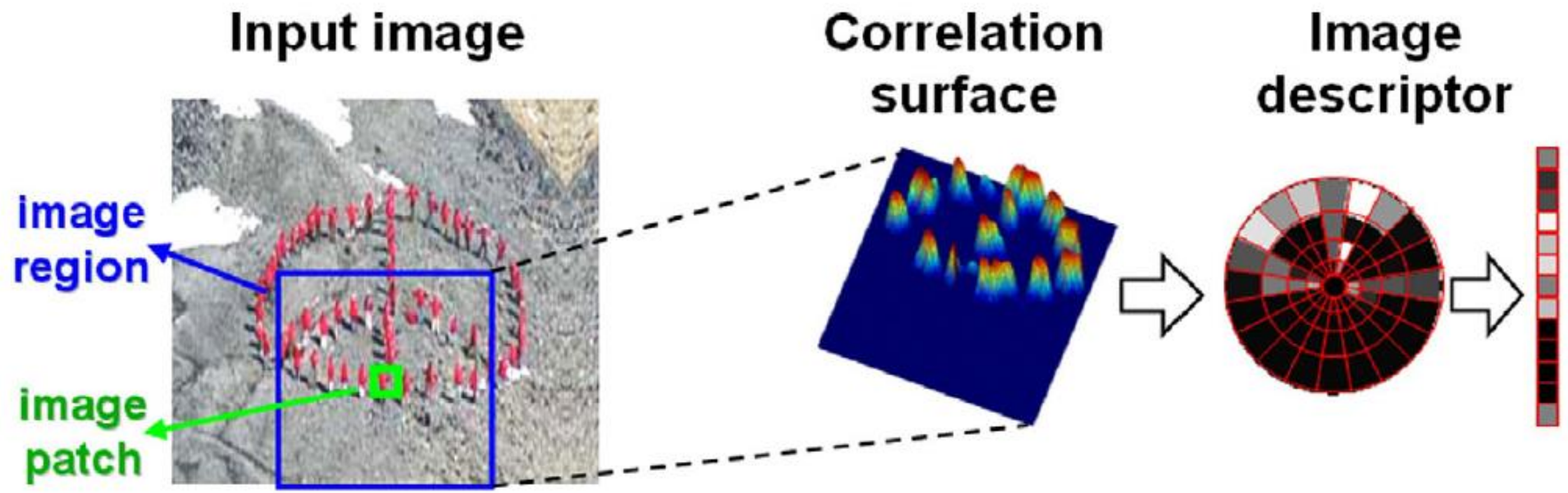


Self-similarity Descriptor

Figure 1. *These images of the same object (a heart) do NOT share common image properties (colors, textures, edges), but DO share a similar geometric layout of local internal self-similarities.*

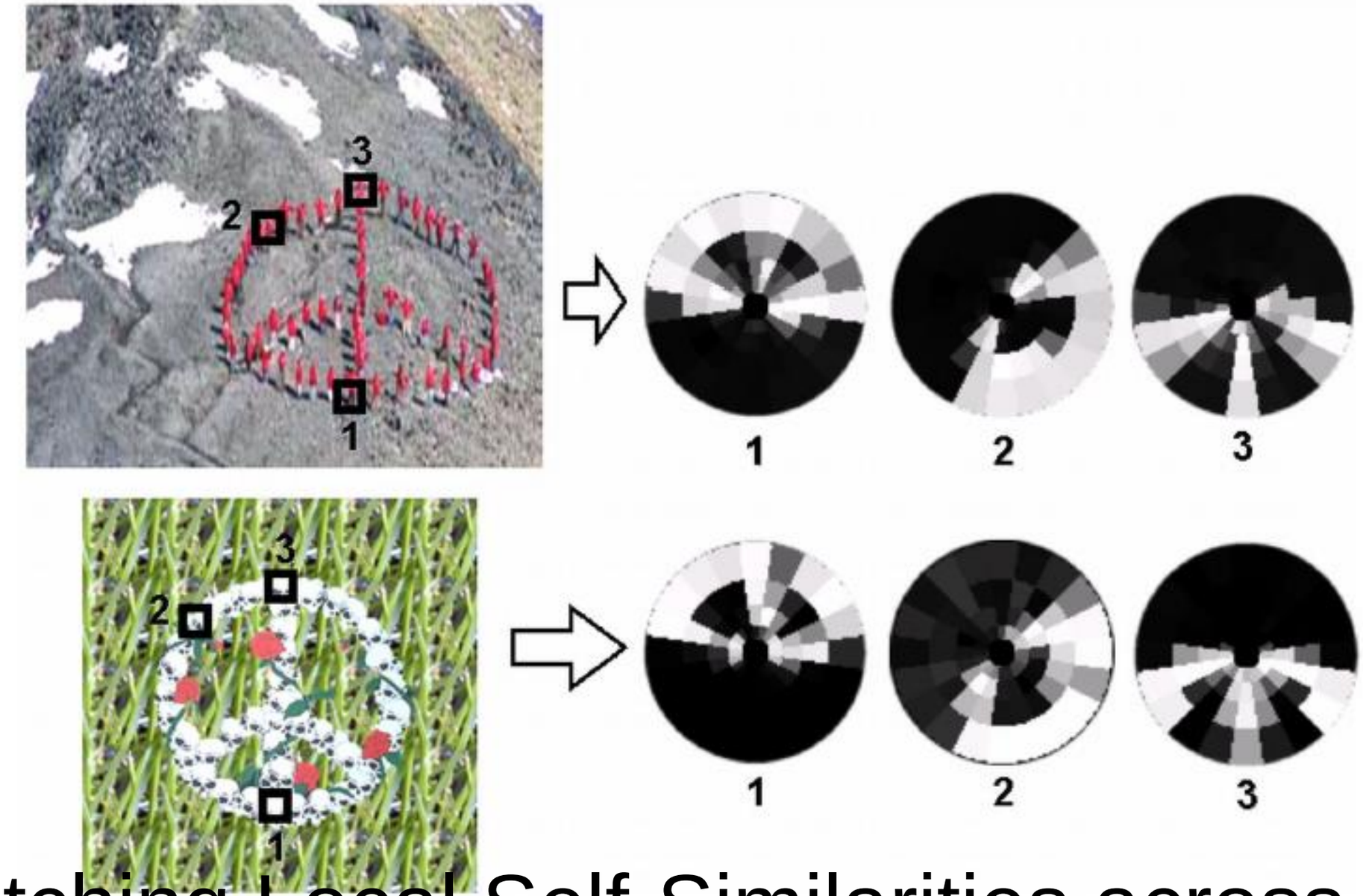
Matching Local Self-Similarities across Images and Videos, Shechtman and Irani, 2007

Self-similarity Descriptor



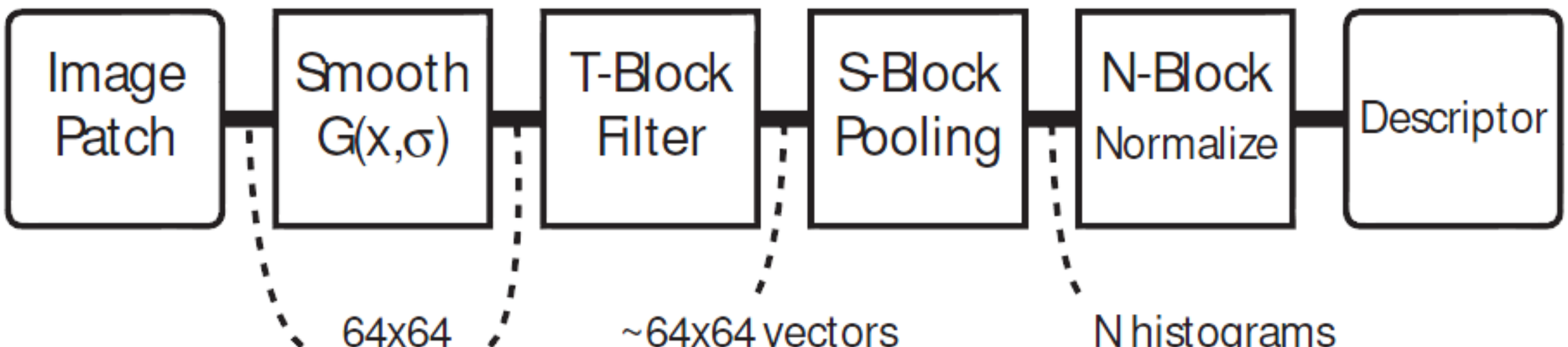
Matching Local Self-Similarities across Images and Videos, Shechtman and Irani, 2007

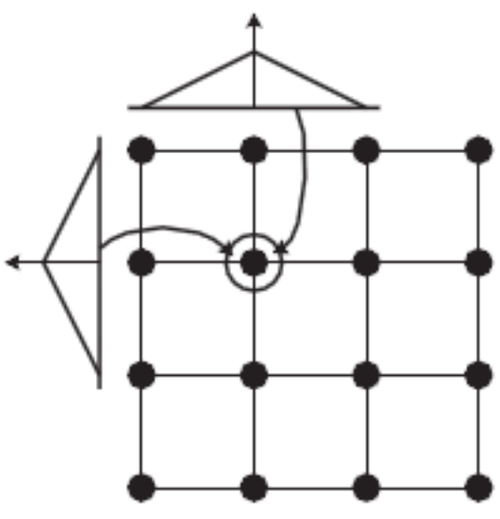
Self-similarity Descriptor



Matching Local Self-Similarities across Images and Videos, Shechtman and Irani, 2007

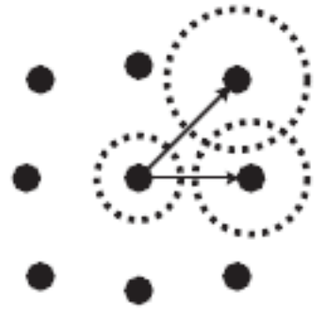
Learning Local Image Descriptors
Winder and Brown, 2007



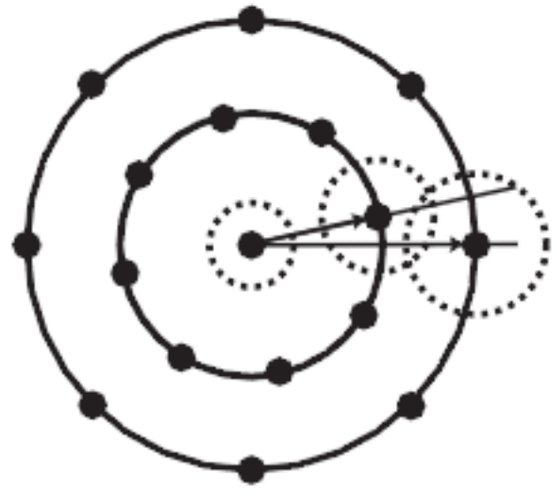


S1: SIFT grid with bilinear weights

S2: GLOH polar grid with bilinear radial and angular weights

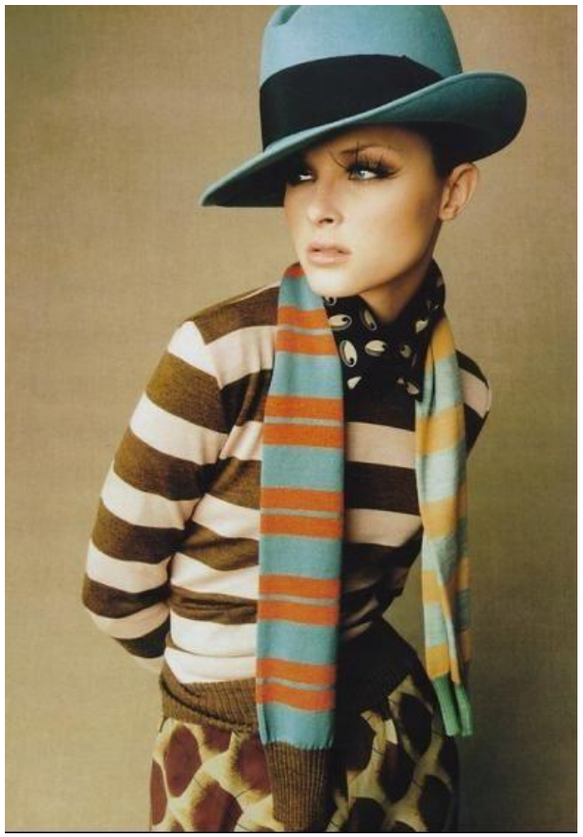


S3: 3x3 grid with Gaussian weights

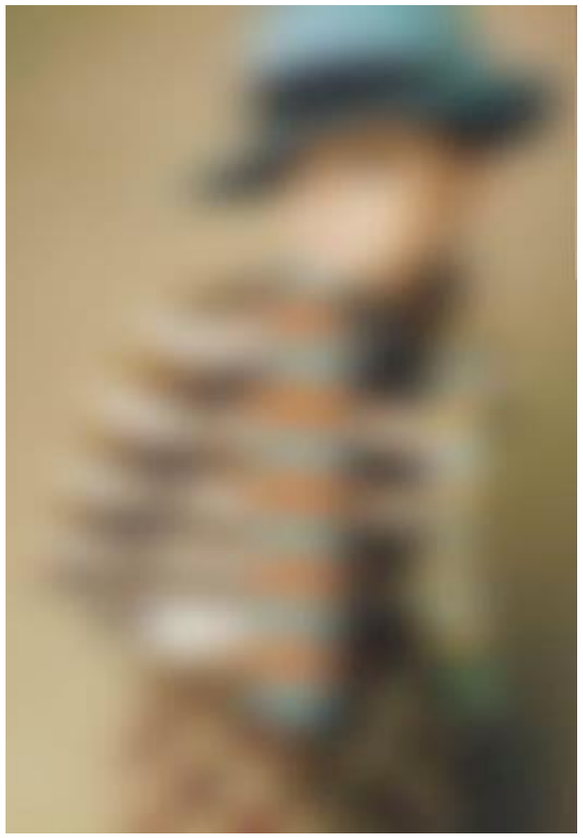


S4: 17 polar samples with Gaussian weights

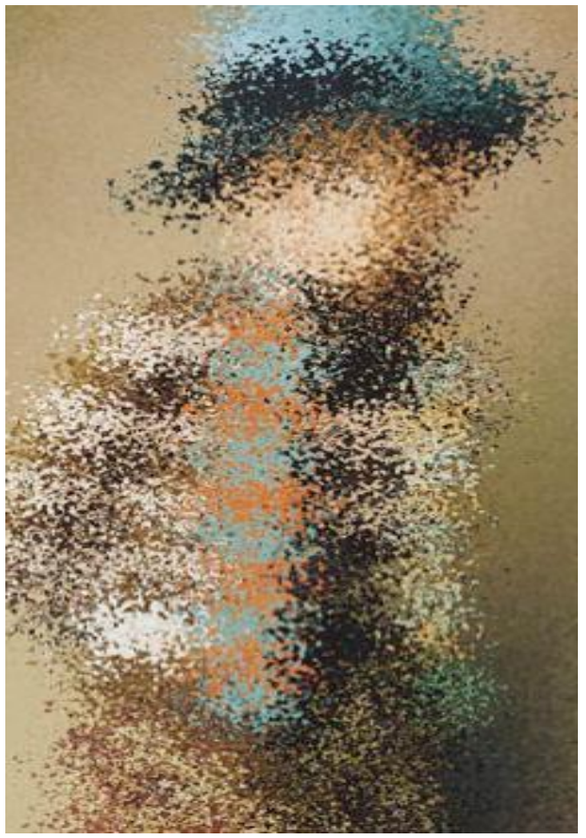
Discriminative power



Raw pixels



Sampled



Locally orderless

Global histogram

Generalization power

Feature Matching

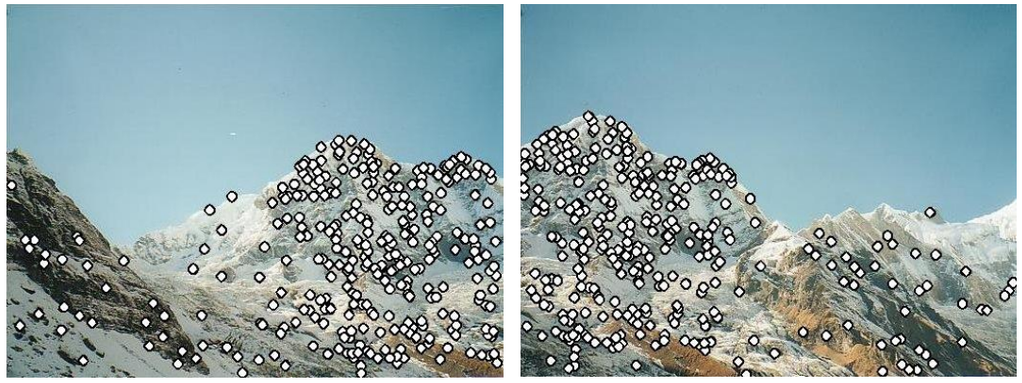
Read Szeliski 4.1

Many slides from James Hays, Derek Hoiem, and Grauman&Leibe 2008 AAAI Tutorial

Local features: main components

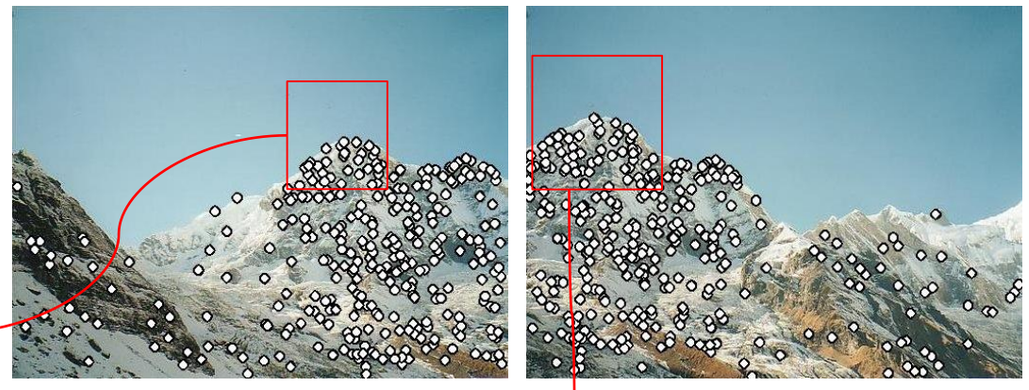
1) Detection:

Find a set of distinctive key points.



2) Description:

Extract feature descriptor around each interest point as vector.

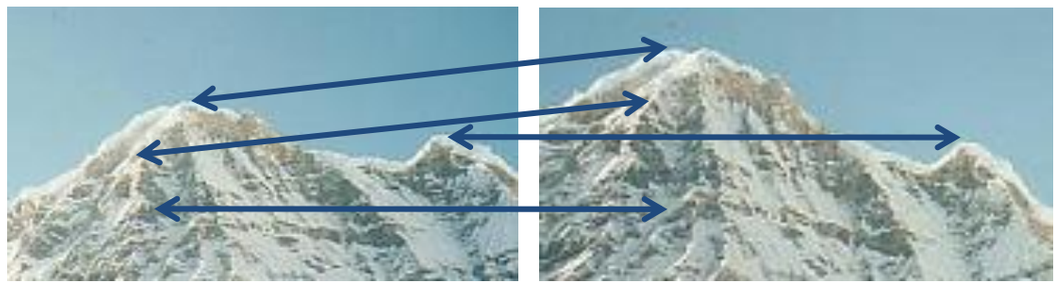


$$\mathbf{x}_1 = [x_1^{(1)}, \ldots, x_d^{(1)}]$$

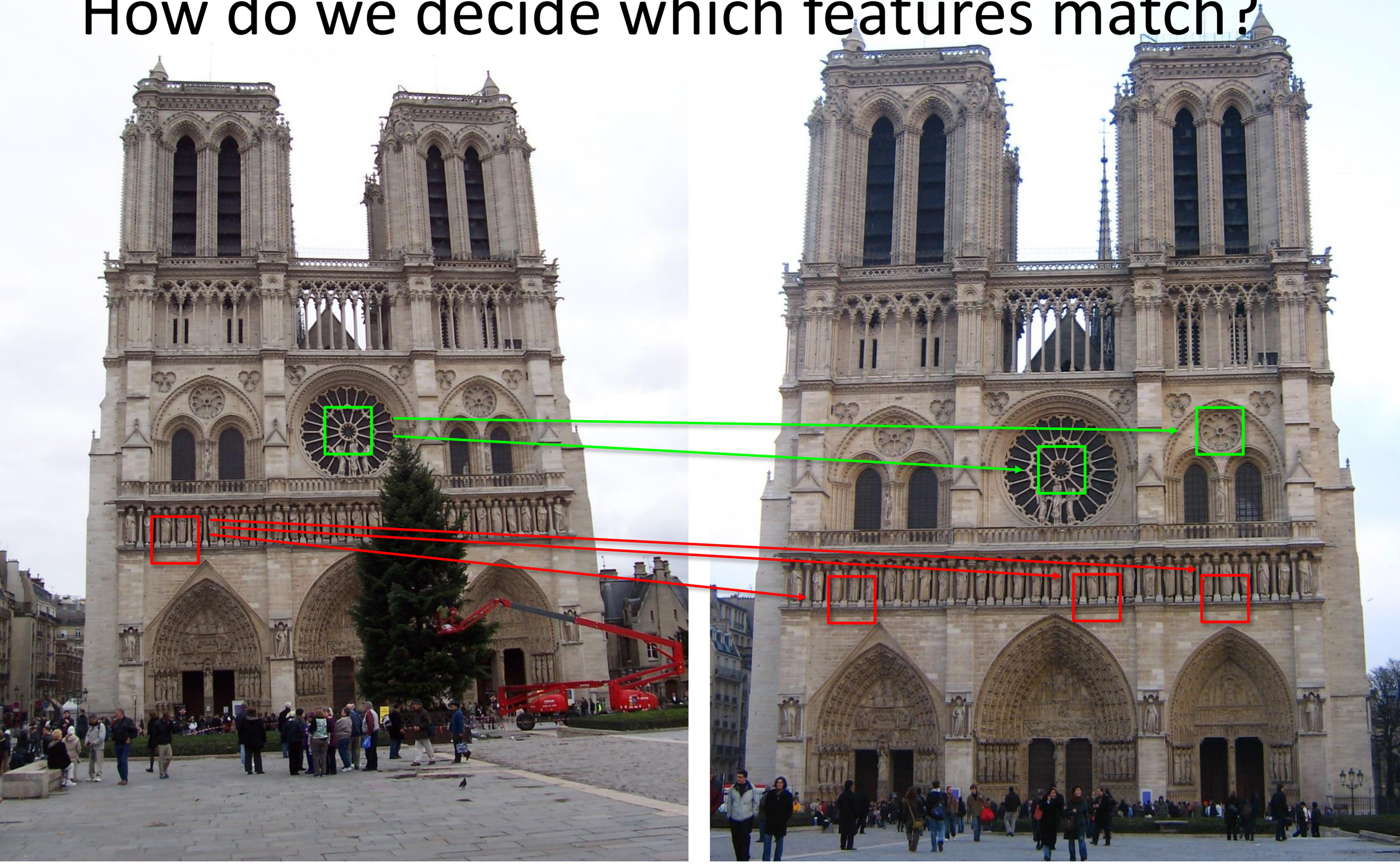
$$\mathbf{x}_2 = [x_1^{(2)}, \ldots, x_d^{(2)}]$$

3) Matching:

Compute distance between feature vectors to find correspondence.



How do we decide which features match?

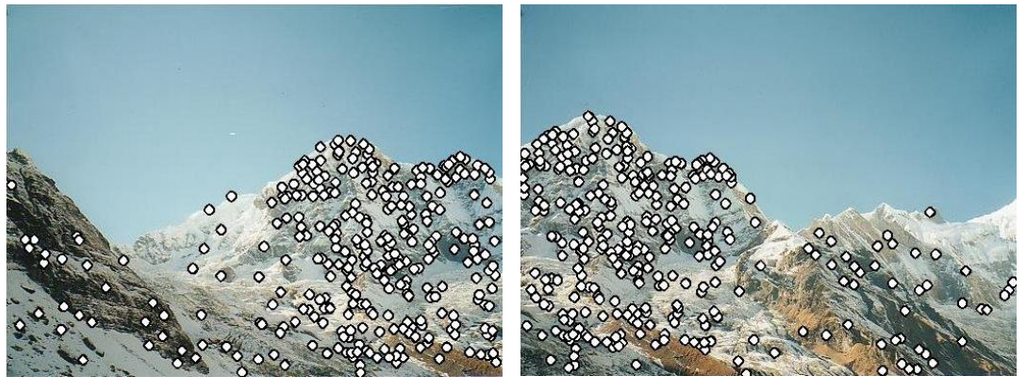


Distance: 0.34, 0.30, 0.40

Distance: 0.61, 1.22

Think-Pair-Share

- *Design a feature point matching scheme.*
- Two images, I_1 and I_2



- Two sets X_1 and X_2 of feature points
  - Each feature point $\mathbf{x_1}$ has a descriptor $\mathbf{x}_1 = [x_1^{(1)}, \ldots, x_d^{(1)}]$

- Distance, bijective/injective/surjective, noise, confidence, computational complexity, generality…

Euclidean distance vs. Cosine Similarity

- Euclidean distance:

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

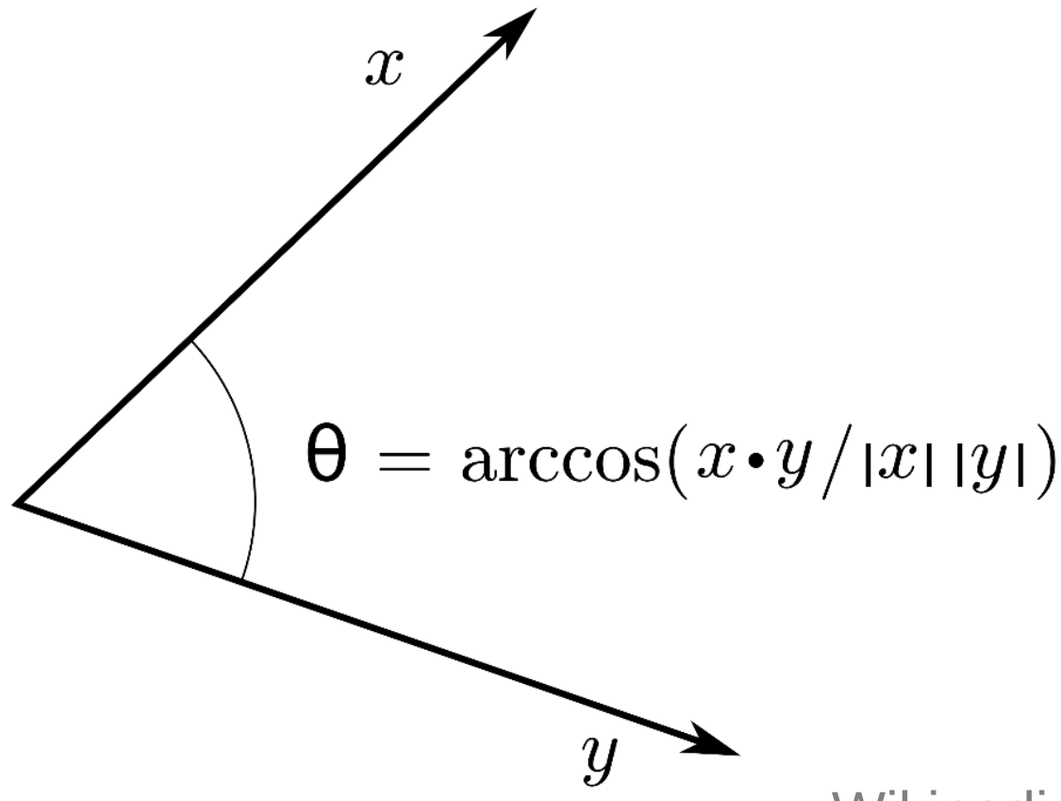
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

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

- 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 descriptors

- Simplest descriptor: vector of raw intensity values
- How to compare two such vectors?
  - Sum of squared differences (SSD)

$$\mathrm{SSD}(u,v) = \sum_i \left(u_i - v_i\right)^2$$

    - Not invariant to intensity change

- Normalized correlation

$$\rho(u,v) = \frac{\sum_i (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\left(\sum_j (u_j - \bar{u})^2\right)\left(\sum_j (v_j - \bar{v})^2\right)}}$$

    - Invariant to affine intensity change

Feature Matching

- Criteria 1:
  - Compute distance in feature space, e.g., Euclidean distance between 128-dim SIFT descriptors
  - Match point to lowest distance (nearest neighbor)

- Problems:
  - Does everything have a match?

Feature Matching

- Criteria 2:
  - 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 (no match!)

- 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.

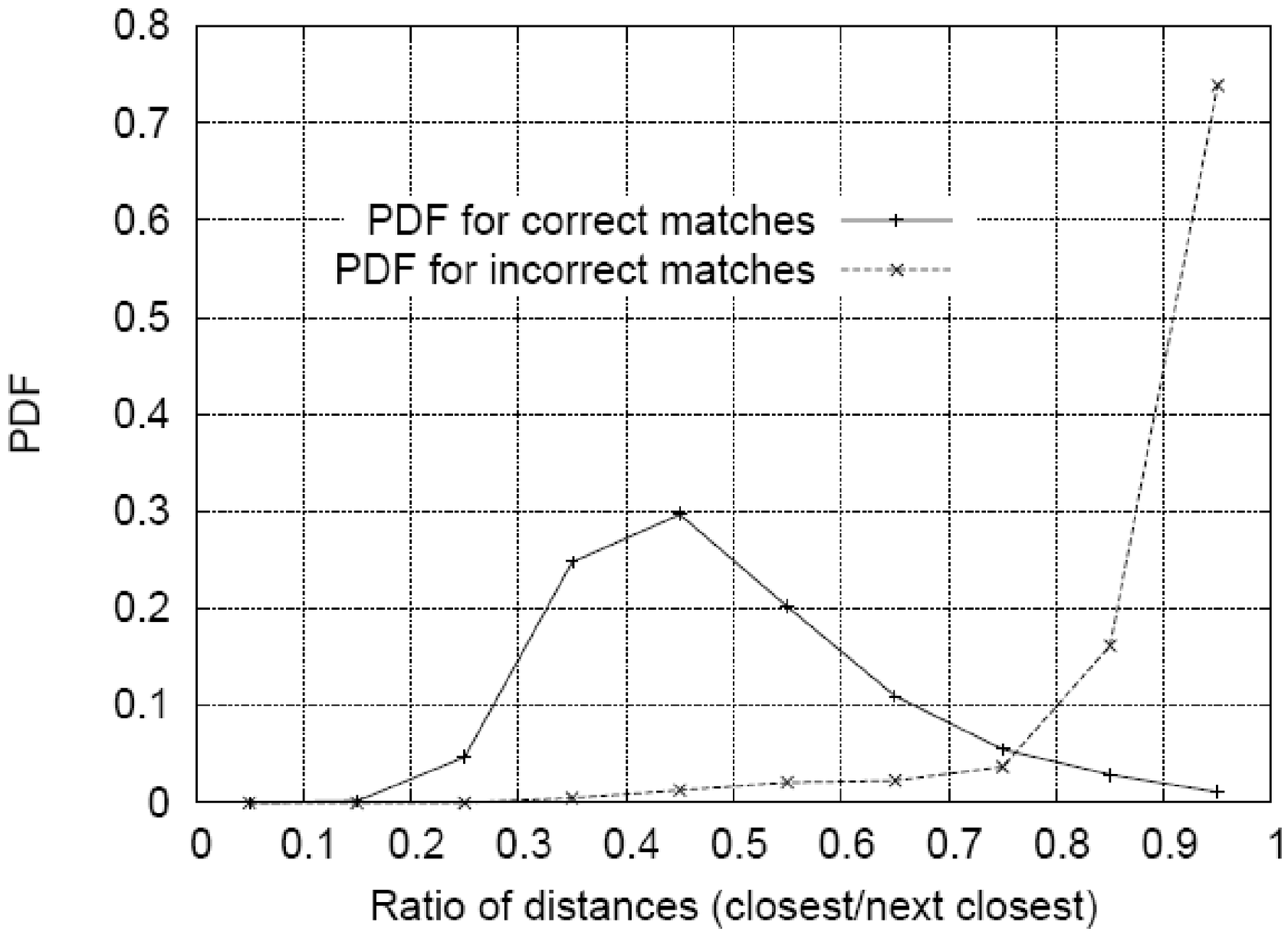
- If NN1 ≈ NN2, ratio $\frac{NN1}{NN2}$ will be ≈ 1 -> matches too close.
- As NN1 << NN2, ratio $\frac{NN1}{NN2}$ tends to 0.

Sorting by this ratio puts matches in order of confidence.

Threshold ratio – but how to choose?

Nearest Neighbor Distance Ratio

- Lowe computed a probability distribution functions of ratios
- 40,000 keypoints with hand-labeled ground truth



Ratio threshold depends on your application's view on the trade-off between the number of false positives and true positives!

Lowe IJCV 2004

Efficient compute cost

- Naïve looping: Expensive

- Operate on matrices of descriptors
- E.g., for row vectors,

`features_image1 * features_image2`^{T}

produces matrix of dot product results for all pairs of features (cosine similarity). What can we do for Euclidean distance?

Live SIFT Demo

Actually a similar alternative, called SURF.

Speeded-Up Robust Features
Bay et al., ECCV 2006

(Both are patented; SURF has non-commercial license.)

Right features are application specific

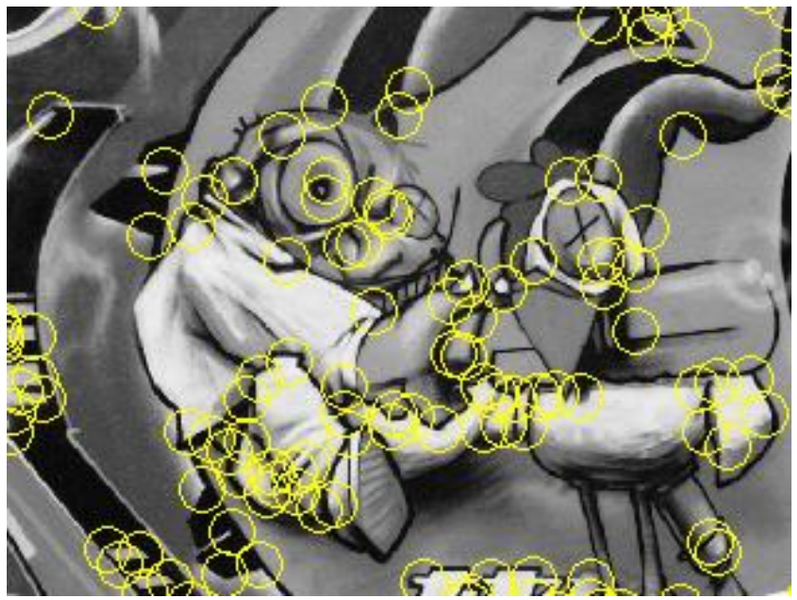
- Shape: scene-scale, object-scale, detail-scale
  - 2D form, shading, shadows, texture, linear perspective
- Material properties: albedo, feel, hardness, …
  - Color, texture
- Motion
  - Optical flow, tracked points
- Distance
  - Stereo, position, occlusion, scene shape
  - If known object: size, other objects

Available at a web site near you…

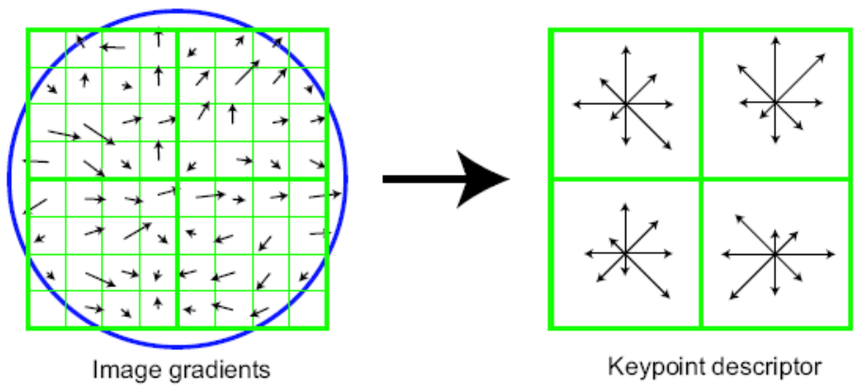
- Many local feature detectors have executables available online:
  - http://www.robots.ox.ac.uk/~vgg/research/affine
  - http://www.cs.ubc.ca/~lowe/keypoints/
  - http://www.vision.ee.ethz.ch/~surf

Review: Interest points

- Keypoint detection: repeatable and distinctive
  - Corners, blobs, stable regions
  - Harris, DoG



- Descriptors: robust and selective
  - Spatial histograms of orientation
  - SIFT



References

Basic reading:

- Szeliski textbook, Sections 4.1.2, 14.1.2.