Lecture:
Scale Selection, Local Descriptors, SIFT

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

What we will learn today?

- Scale invariant keypoint detection
  - Automatic scale selection
  - Difference-of-Gaussian (DoG) detector
- SIFT: an image region descriptor

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

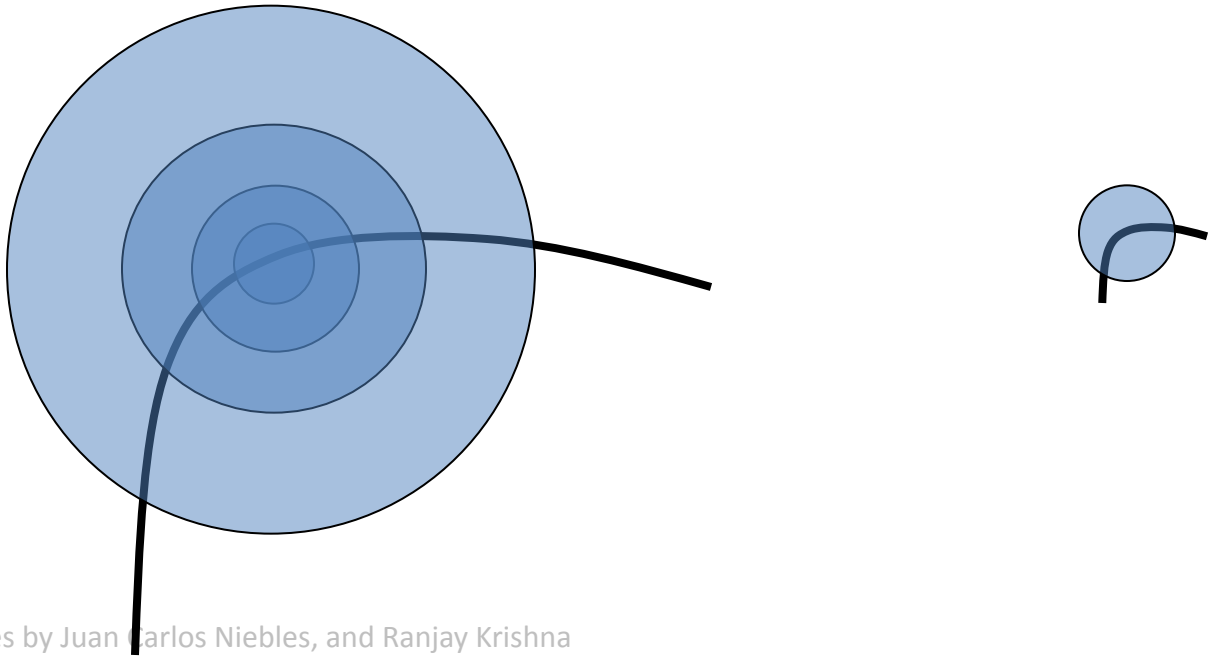
Does scale matter?

- When detecting corners, the `scale` of the window you use can change the corners you detect.

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Scale Invariant Detection

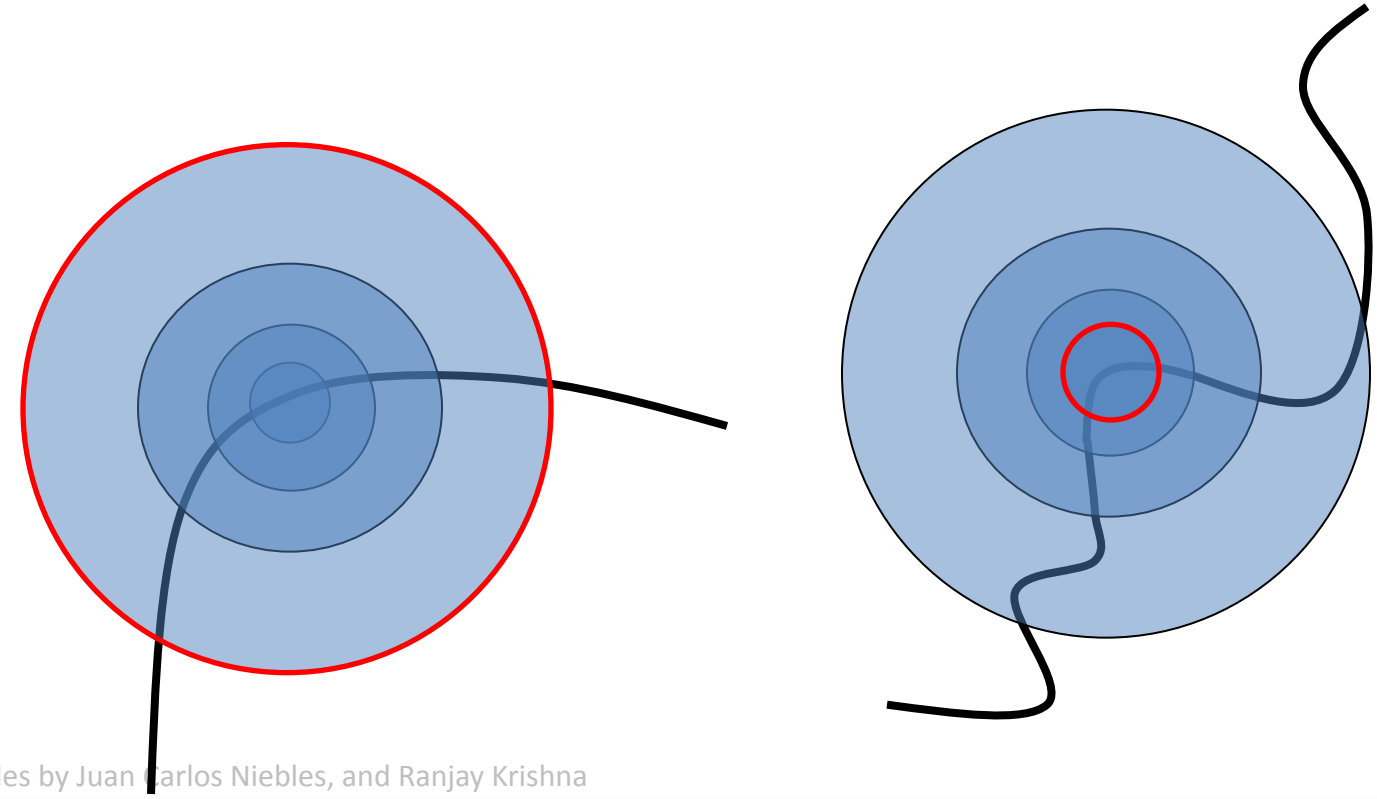
- Consider regions (e.g. circles) of different sizes around a point
- Find regions of corresponding sizes that will look the same in both images?



Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Scale Invariant Detection

- The problem: how do we choose corresponding circles ***independently*** in each image?



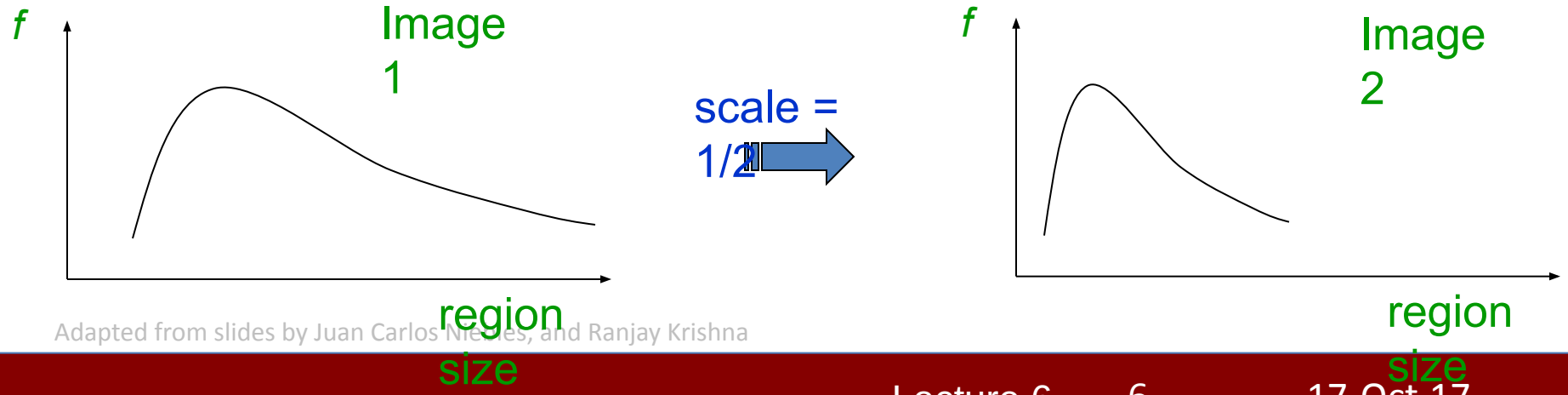
Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Scale Invariant Detection

- Solution:
  - Design a function on the region (circle), which is "scale invariant" in the following sense: the same for corresponding regions, even if they are at different scales)

Example: average intensity. For corresponding regions (even if regions have # of pixels), it will be the same.

  - For a point in one image, we can consider it as a function of region size (circle radius)

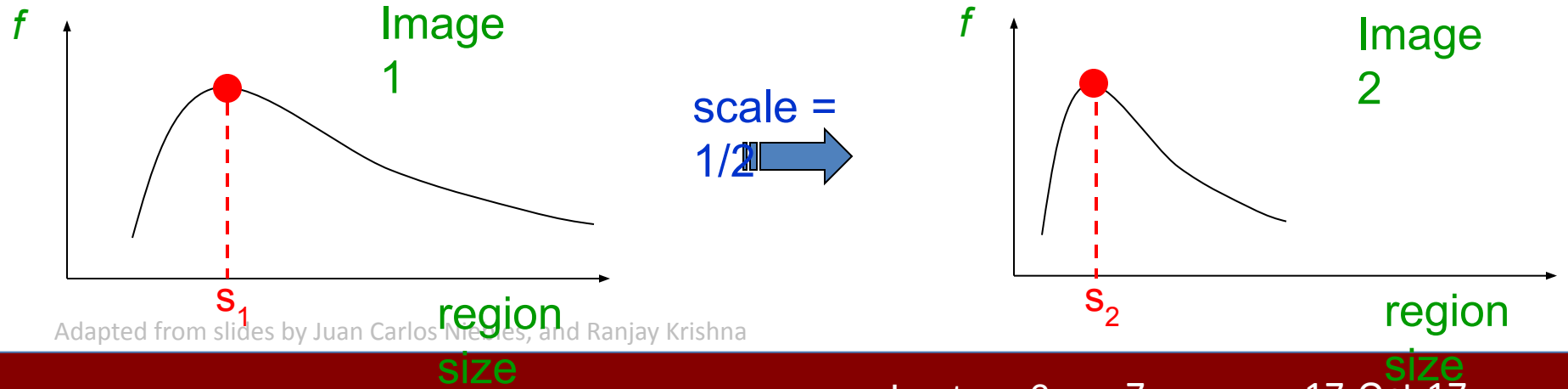


Scale Invariant Detection

- Common approach:
  Take a local maximum of this function

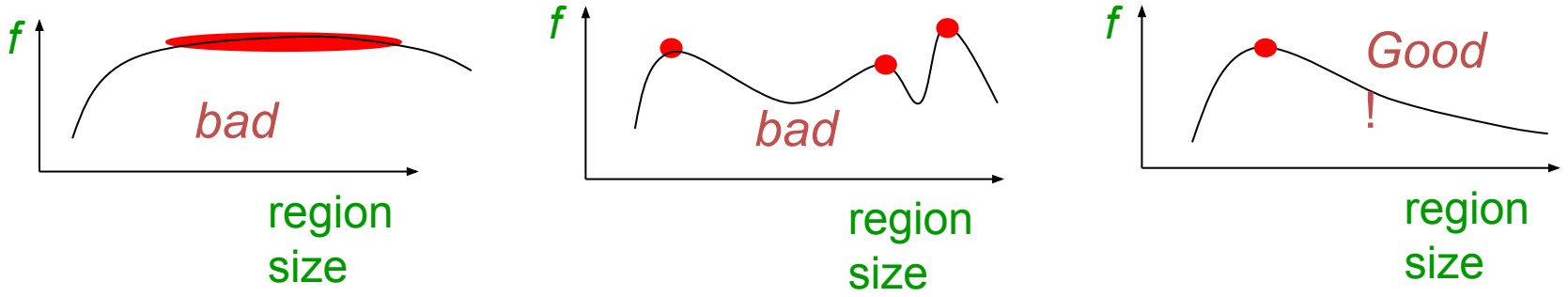
- Observation: region size, for which the maximum is achieved, should be *co-variant* with image scale.

Important: this scale invariant region size is found in each image independently!



Scale Invariant Detection

- A "good" function for scale detection:
  has one stable sharp peak



- For usual images: a good function would be one which responds to contrast (sharp local intensity change)

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

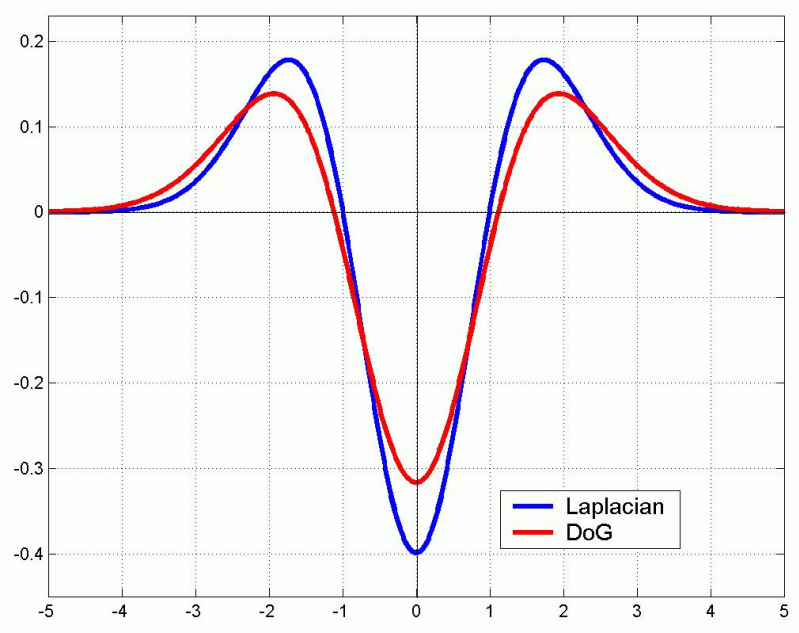
Scale Invariant Detection

- Functions for determining scale $f = \text{Kernel} * \text{Image}$

Kernels:

$$L = \sigma^2 \left(G_{xx}(x, y, \sigma) + G_{yy}(x, y, \sigma) \right)$$

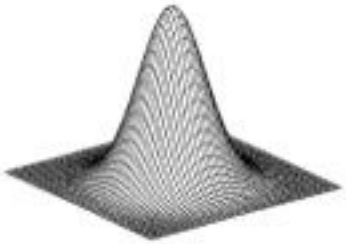
(Laplacian)



Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

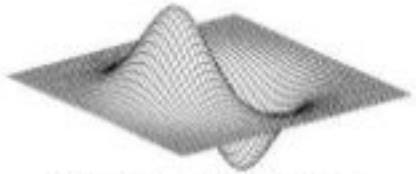
Derivative-of-Gaussian versus Laplacian-of-Gaussian

2D edge detection filters



Gaussian

$$h_\sigma(u,v) = \frac{1}{2\pi\sigma^2} e^{-\frac{u^2+v^2}{2\sigma^2}}$$



derivative of Gaussian

$$\frac{\partial}{\partial x} h_\sigma(u,v)$$

Laplacian of Gaussian

$$\nabla^2 h_\sigma(u,v)$$

∇^2 is the **Laplacian** operator:

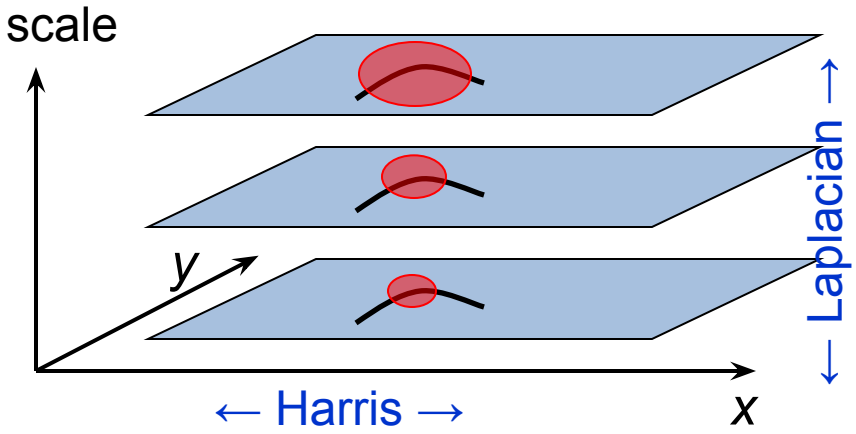
$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

24

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Scale Invariant Detectors

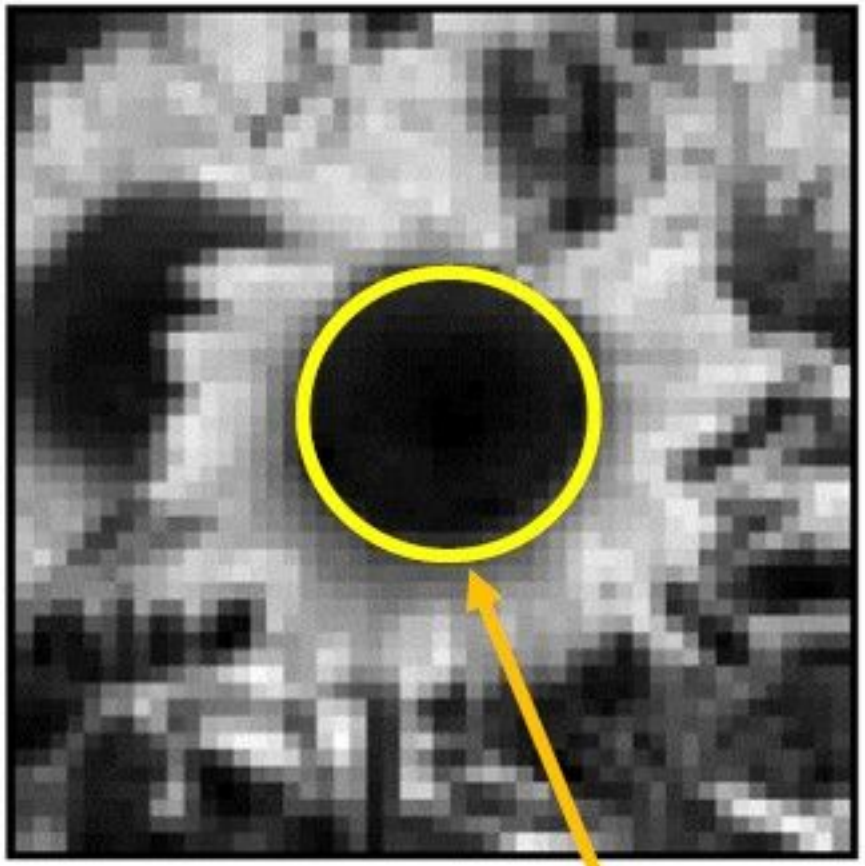
- Harris-Laplacian[1]
  Find local maximum of:
  - Harris corner detector in **space** (image coordinates)
  - Laplacian in **scale**

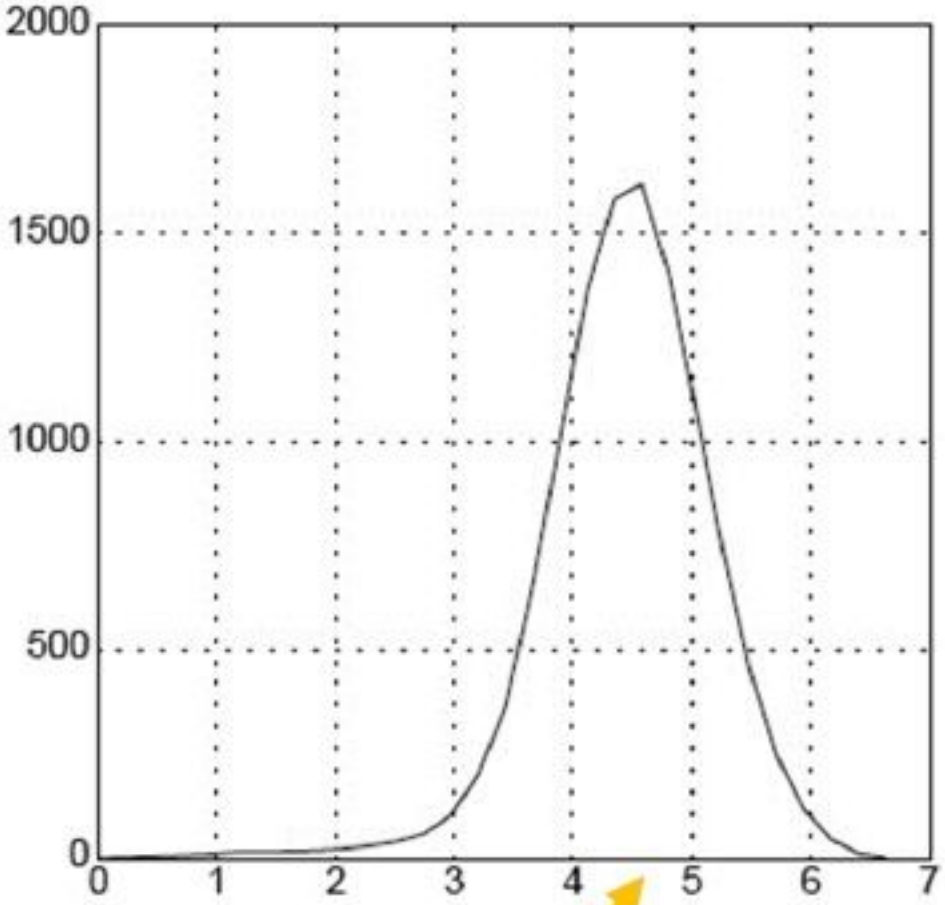


[1] K.Mikolajczyk, C.Schmid. "Indexing Based on Scale Invariant Interest Points". ICCV 2001
[2] D.Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". IJCV 2004

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Laplacian



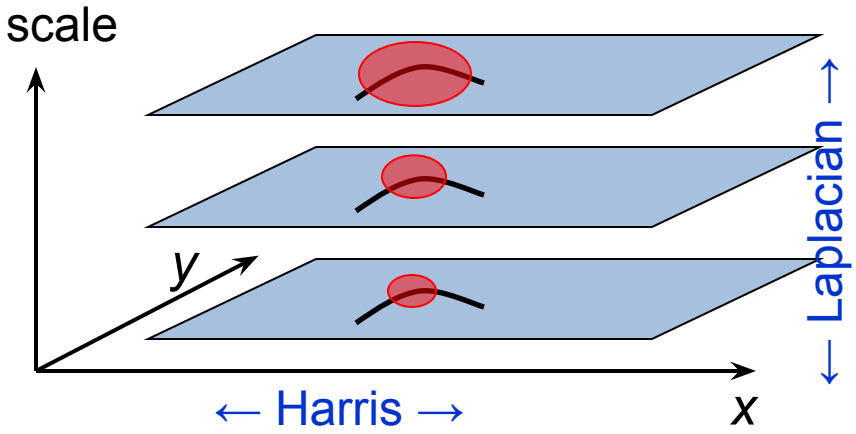


Characteristic scale

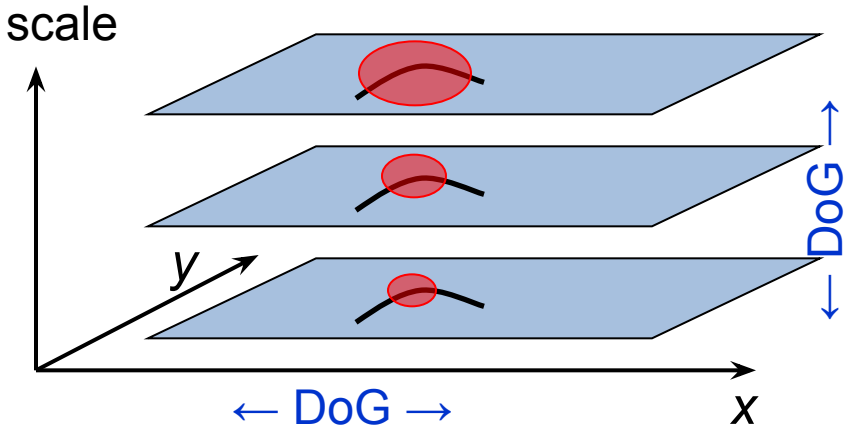
Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Scale Invariant Detectors

- Harris-Laplacian[1]
  Find local maximum of:
  - Harris corner detector in **space** (image coordinates)
  - Laplacian in **scale**



- SIFT (Lowe)[2]
  Find local maximum of:
  - Difference of Gaussians **in space and scale**



[1] K.Mikolajczyk, C.Schmid. "Indexing Based on Scale Invariant Interest Points". ICCV 2001
[2] D.Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". IJCV 2004

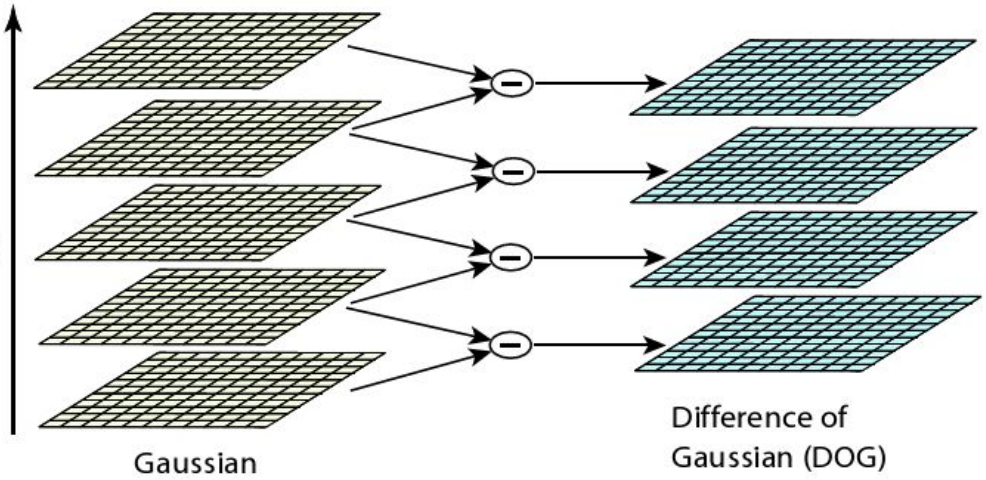
Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Difference-of-Gaussians

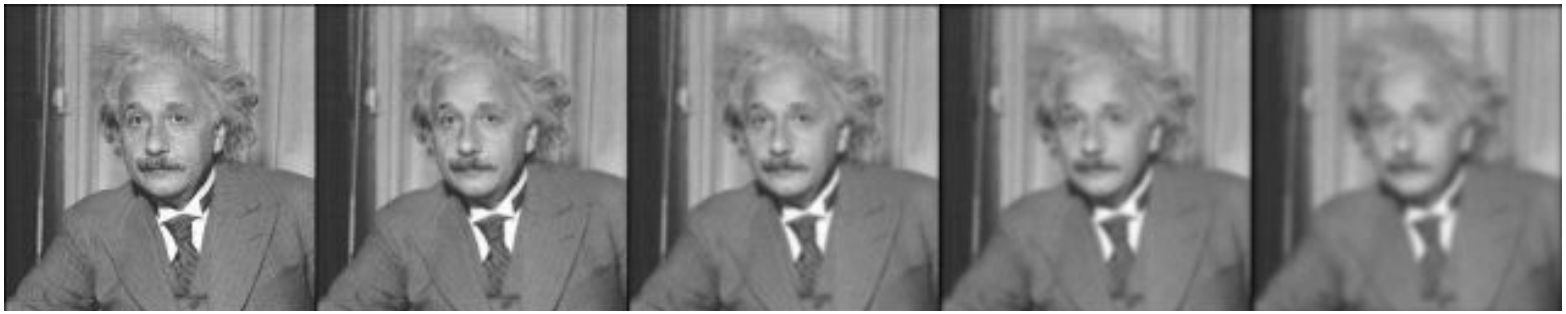
$G(k^2\sigma) * I$

$G(k\sigma) * I$

$G(\sigma) * I$



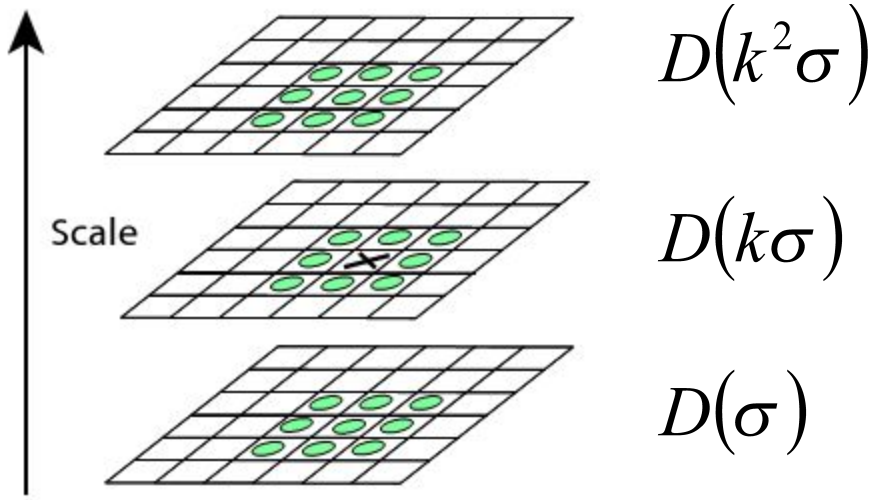
$D(\sigma) \equiv (G(k\sigma) - G(\sigma)) * I$



Adapted from

Scale-Space Extrema

- Choose all extrema within 3x3x3 neighborhood.



X is selected if it is larger or smaller than all 26 neighbors

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

DoG approximates Laplacian

- Functions for determining scale $f = \text{Kernel} * \text{Image}$

Kernels:

$$L = \sigma^2 \left(G_{xx}(x, y, \sigma) + G_{yy}(x, y, \sigma) \right)$$

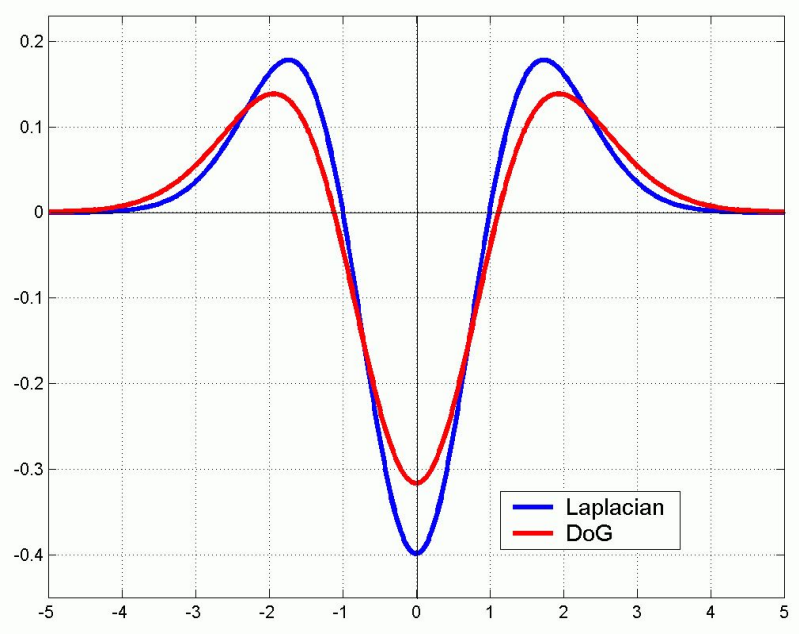
(Laplacian)

$$DoG = G(x, y, k\sigma) - G(x, y, \sigma)$$

(Difference of Gaussians)

where Gaussian

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



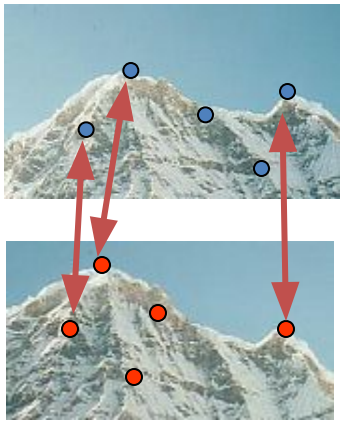
Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

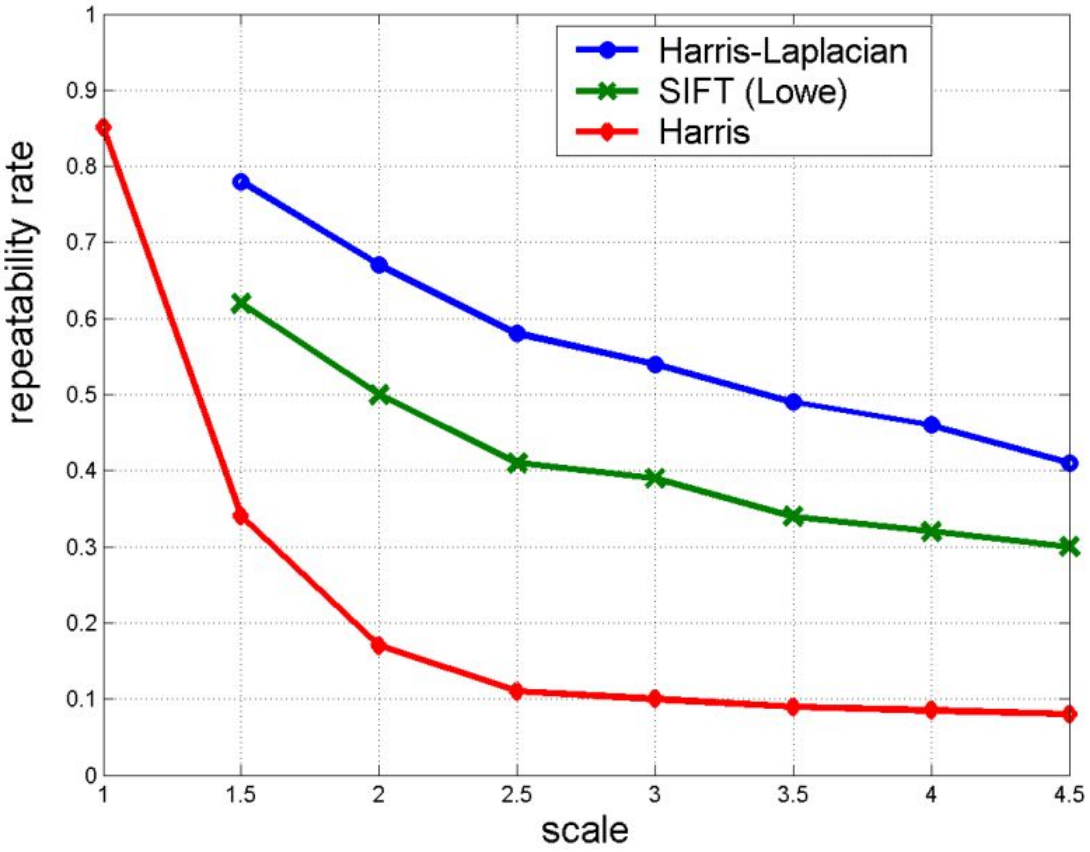
Scale Invariant Detectors

- Experimental evaluation of detectors w.r.t. scale change

Repeatability rate:

$$\frac{\#\ \text{correspondences}}{\#\ \text{possible correspondences}}$$





K.Mikolajczyk, C.Schmid. "Indexing Based on Scale Invariant Interest Points". ICCV 2001

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Scale Invariant Detection: Summary

- Given: two images of the same scene with a large *scale difference* between them
- Goal: find *the same* interest points *independently* in each image
- Solution: search for *maxima* of suitable functions in *scale* and in *space* (over the image)

Methods:

1. Harris-Laplacian [Mikolajczyk, Schmid]: maximize Laplacian over scale, Harris' measure of corner response over the image
2. SIFT [Lowe]: maximize Difference of Gaussians over scale and space

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

What's next?

So we have can detect keypoints at varying scales. But what can we do with those keypoints?

Things we would like to do:

- Search:
  - We would need to find similar key points in other images
- Panorama
  - Match keypoints from one image to another.
- Etc…

For all such applications, we need a way of `describing` the keypoints.

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

What we will learn today?

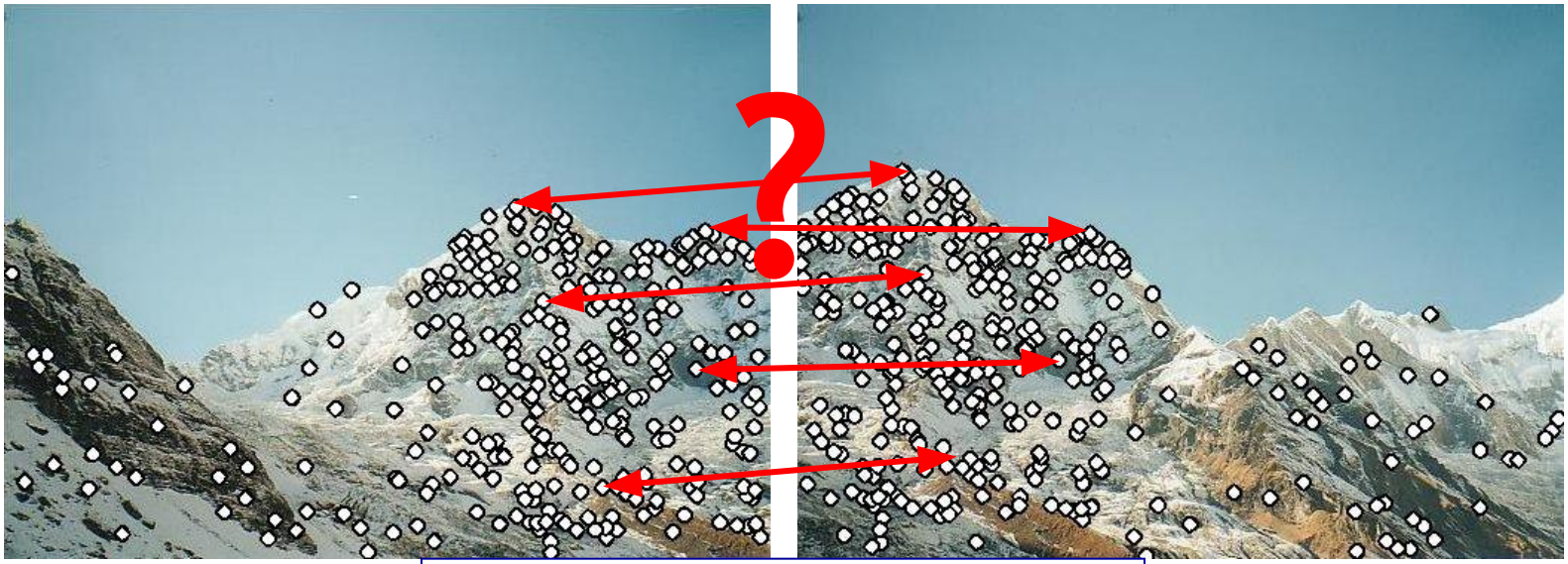
- Scale invariant region selection
  - Automatic scale selection
  - Difference-of-Gaussian (DoG) detector
- SIFT: an image region descriptor

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Local Descriptors

- We know how to detect points
- Next question:

How to *describe* them for matching?



Point descriptor should be:

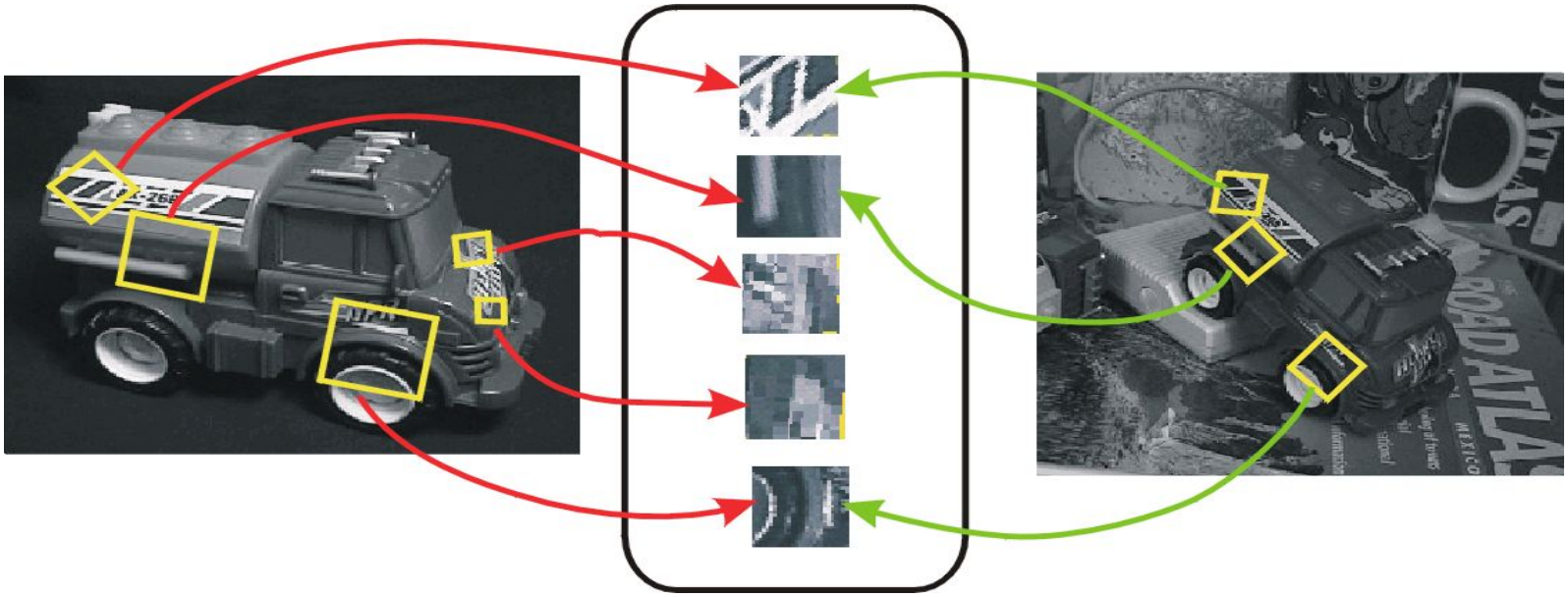
1. **Invariant**
2. **Distinctive**

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Slide credit: Kristen Grauman

Invariant Local Features

- Image content is transformed into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters



Following slides credit: CVPR 2003 Tutorial on **Recognition and Matching Based on Local Invariant Features** David Lowe

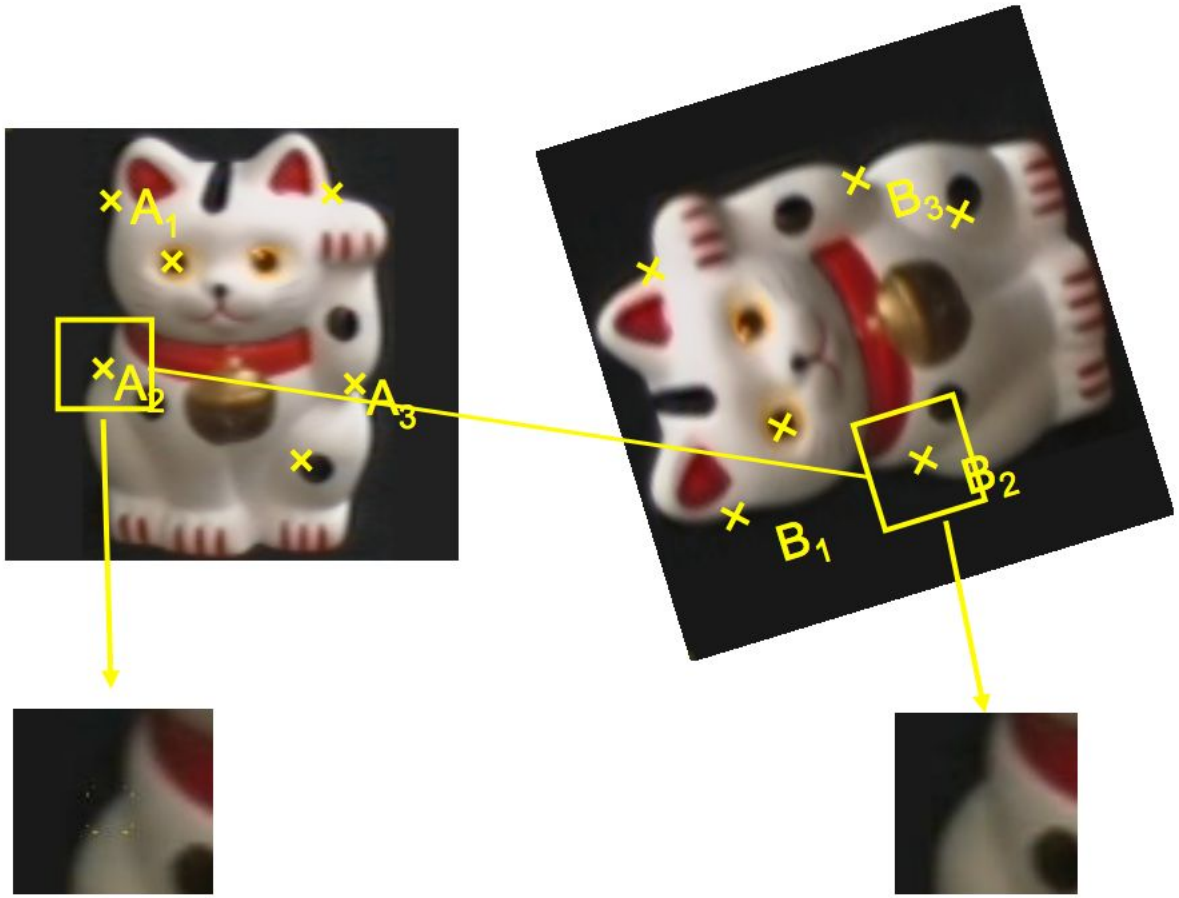
Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Advantages of invariant local features

- **Locality:** features are local, so robust to occlusion and clutter (no prior segmentation)
- **Distinctiveness:** individual features can be matched to a large database of objects
- **Quantity:** many features can be generated for even small objects
- **Efficiency:** close to real-time performance
- **Extensibility:** can easily be extended to wide range of differing feature types, with each adding robustness

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

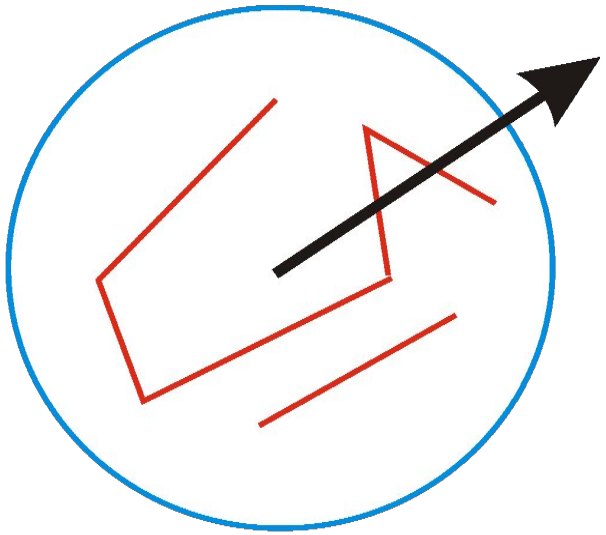
Becoming rotation invariant



Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Becoming rotation invariant

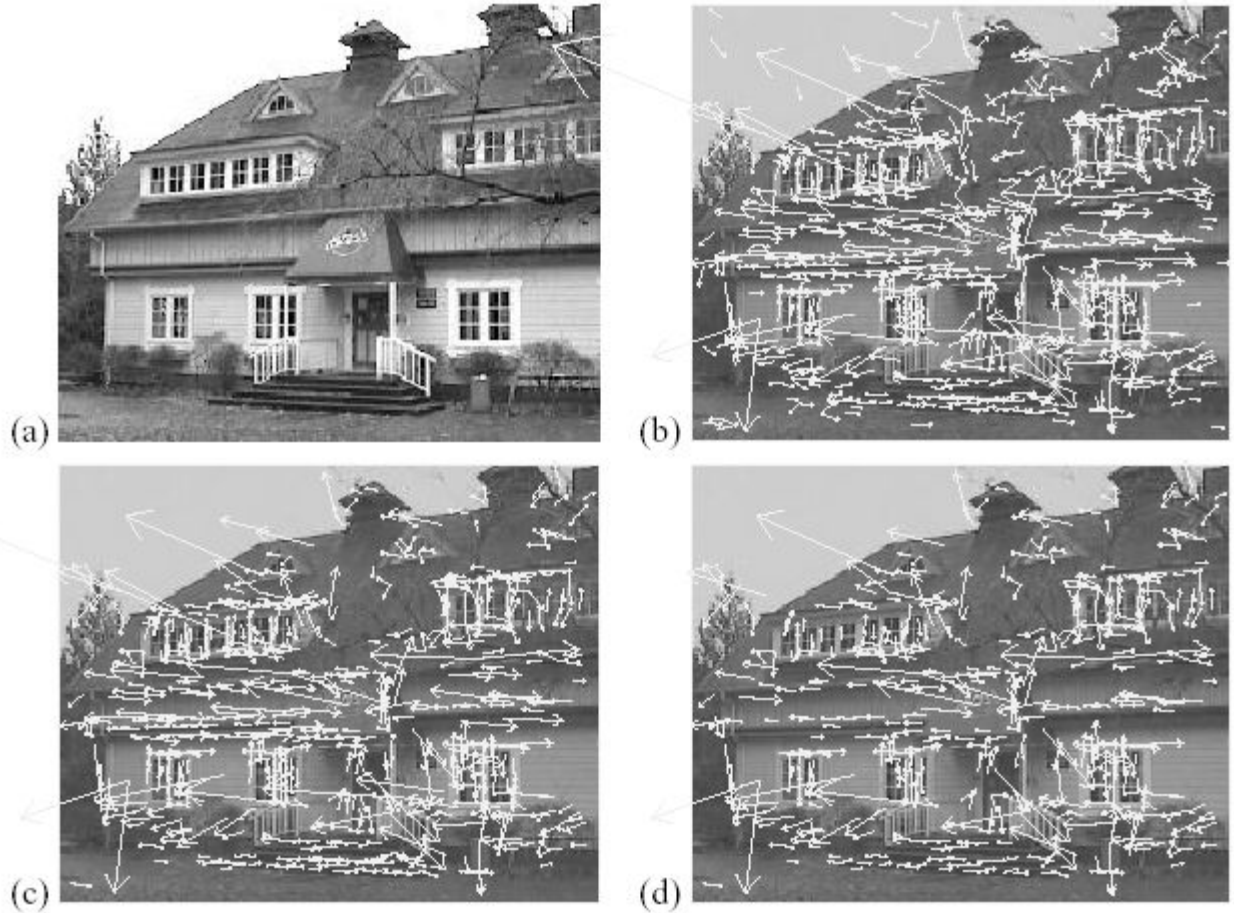
- We are given a keypoint and its scale from DoG
- We will select a characteristic orientation for the keypoint (based on the most prominent gradient there; discussed next slide)
- We will describe all features **relative** to this orientation
- Causes features to be rotation invariant!
  - If the keypoint appears rotated in another image, the features will be the same, because they're **relative** to the characteristic orientation



Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Example of keypoint detection

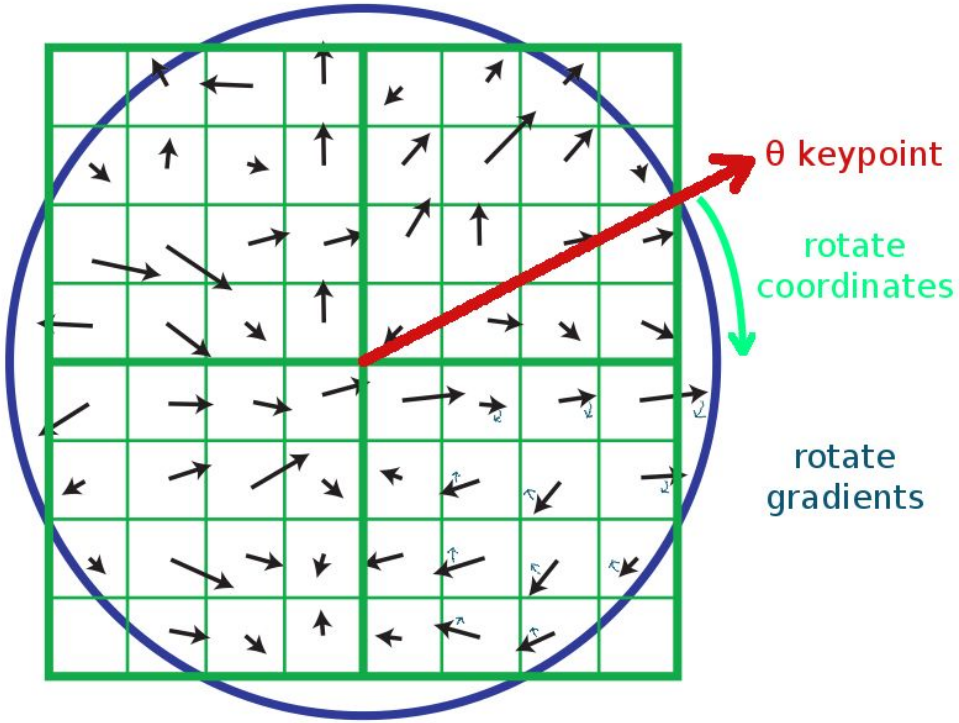
Threshold on value at DOG peak and on ratio of principle curvatures (Harris approach)



(a) 233x189 image
(b) 832 DOG extrema
(c) 729 left after peak value threshold
(d) 536 left after testing ratio of principle curvatures (Harris-like measure to suppress edges)

Vectors indicate scale, orientation and location.

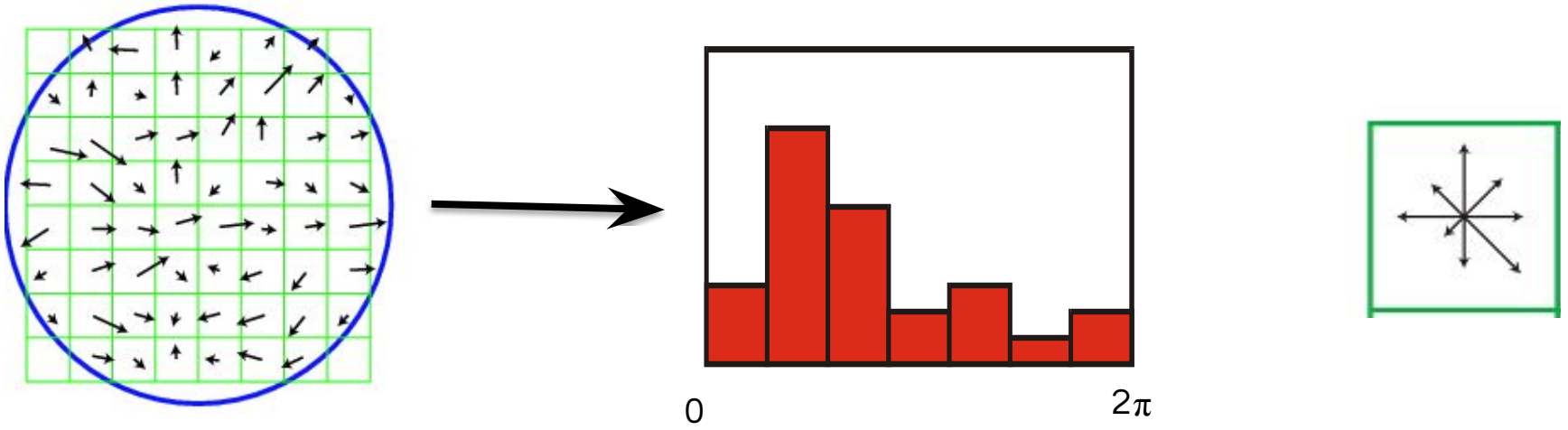
SIFT descriptor formation



- Use the blurred image associated with the keypoint's scale
- Take image gradients over the keypoint neighborhood.
- To become rotation invariant, rotate the gradient directions AND locations by (-keypoint orientation)
  - Now we've cancelled out rotation and have gradients expressed at locations **relative** to keypoint orientation θ
  - We could also have just rotated the whole image by -θ, but that would be slower.

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

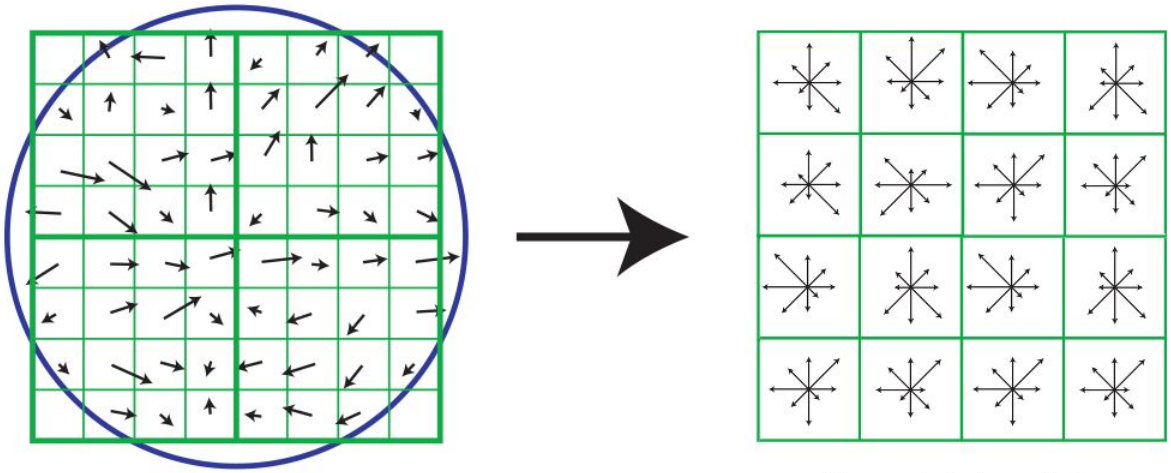
SIFT descriptor formation



- Using precise gradient locations is fragile. We'd like to allow some "slop" in the image, and still produce a very similar descriptor ⇒ HISTOGRAM!
- Create array of orientation histograms (1 histogram is shown)
- Put the rotated gradients into their local orientation histograms

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

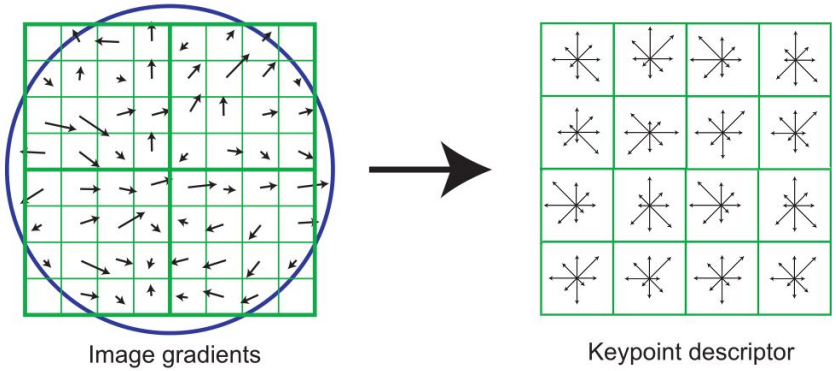
SIFT descriptor formation



- Example with 4x4 spatial grid is shown
- Put the rotated gradients into their local orientation histograms
  - A gradients's contribution is divided among the nearby histograms based on distance. If it's halfway between two histogram locations, it gives a half contribution to both.
  - Also, scale down gradient contributions for gradients far from the center
- The SIFT authors found that best results were with 8 orientation bins per histogram, and a 4x4 histogram array.

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

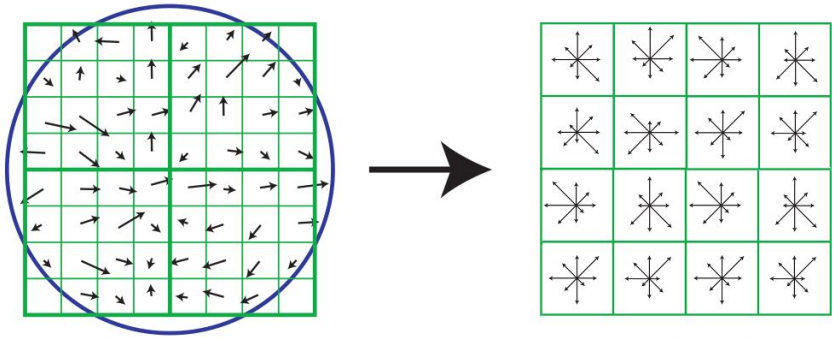
SIFT descriptor formation



- 8 orientation bins per histogram, and a 4x4 histogram array, yields 8 x 4x4 = 128 numbers.
- So a SIFT descriptor is a length 128 vector, which is invariant to rotation (because we rotated the descriptor) and scale (because we worked with the scaled image from DoG)
- We can compare each vector from image A to each vector from image B to find matching keypoints!
  - Euclidean “distance” between descriptor vectors gives a good measure of keypoint similarity

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

SIFT descriptor formation



- Adding robustness to illumination changes:
- Remember that the descriptor is made of gradients (differences between pixels), so it's already invariant to changes in brightness (e.g. adding 10 to all image pixels yields the exact same descriptor)
- A higher-contrast photo will increase the magnitude of gradients linearly. So, to correct for contrast changes, **normalize the vector** (scale to length 1.0)
- Very large image gradients are usually from unreliable 3D illumination effects (glare, etc). So, to reduce their effect, **clamp all values in the vector to be ≤ 0.2 (an experimentally tuned value). Then normalize the vector again.**
- Result is a vector which is fairly invariant to illumination changes.

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Sensitivity to number of histogram orientations

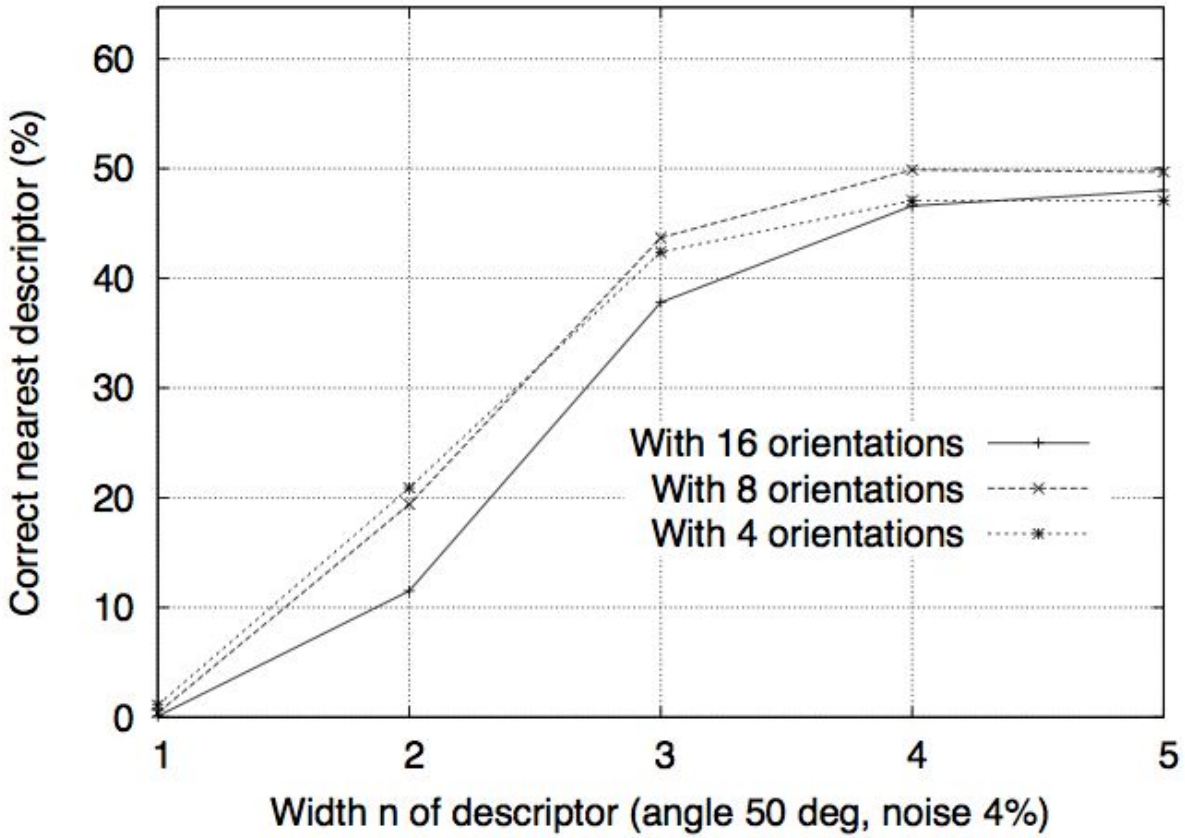


Figure 8: This graph shows the percent of keypoints giving the correct match to a database of 40,000 keypoints as a function of width of the $n \times n$ keypoint descriptor and the number of orientations in each histogram. The graph is computed for images with affine viewpoint change of 50 degrees and addition of 4% noise.

David G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110

Adapted from slides by Juan Carlos Niebles and Ranjay Krishna

Ratio of distances can be used reliably for matching

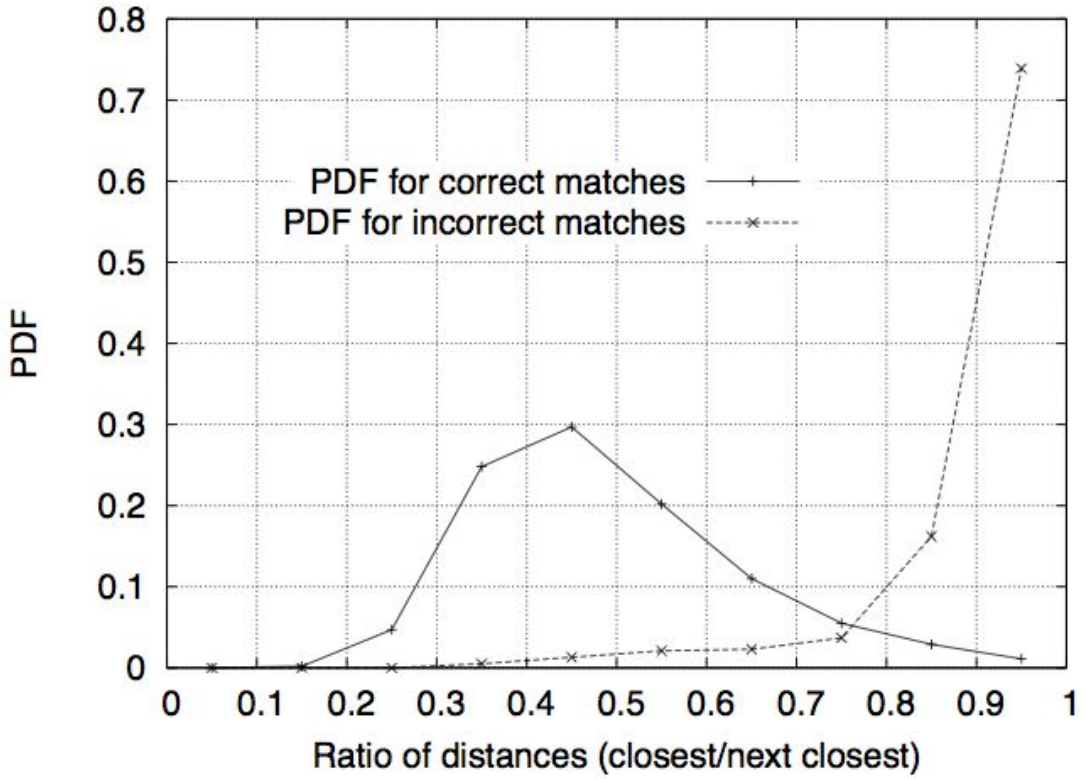


Figure 11: The probability that a match is correct can be determined by taking the ratio of distance from the closest neighbor to the distance of the second closest. Using a database of 40,000 keypoints, the solid line shows the PDF of this ratio for correct matches, while the dotted line is for matches that were incorrect.

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

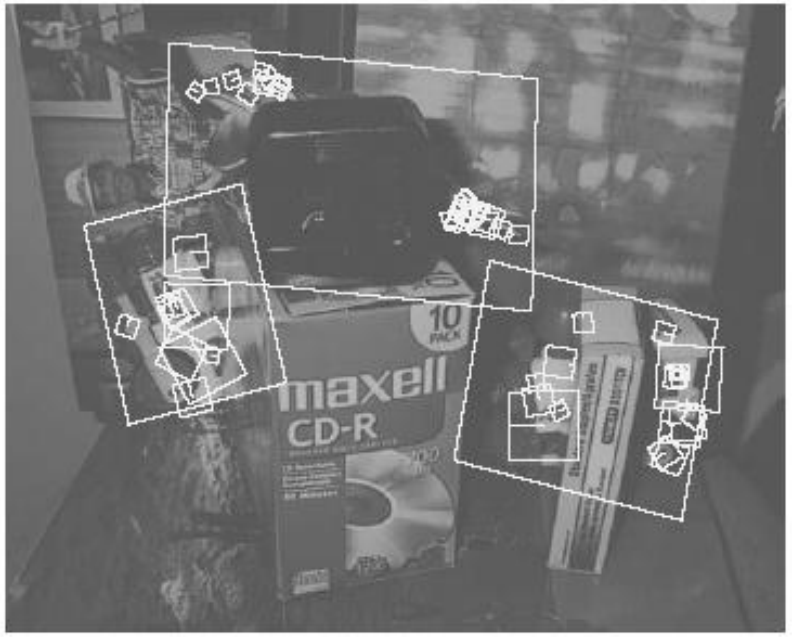


Figure 12: The training images for two objects are shown on the left. These can be recognized in a cluttered image with extensive occlusion, shown in the middle. The results of recognition are shown on the right. A parallelogram is drawn around each recognized object showing the boundaries of the original training image under the affi ne transformation solved for during recognition. Smaller squares indicate the keypoints that were used for recognition.

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

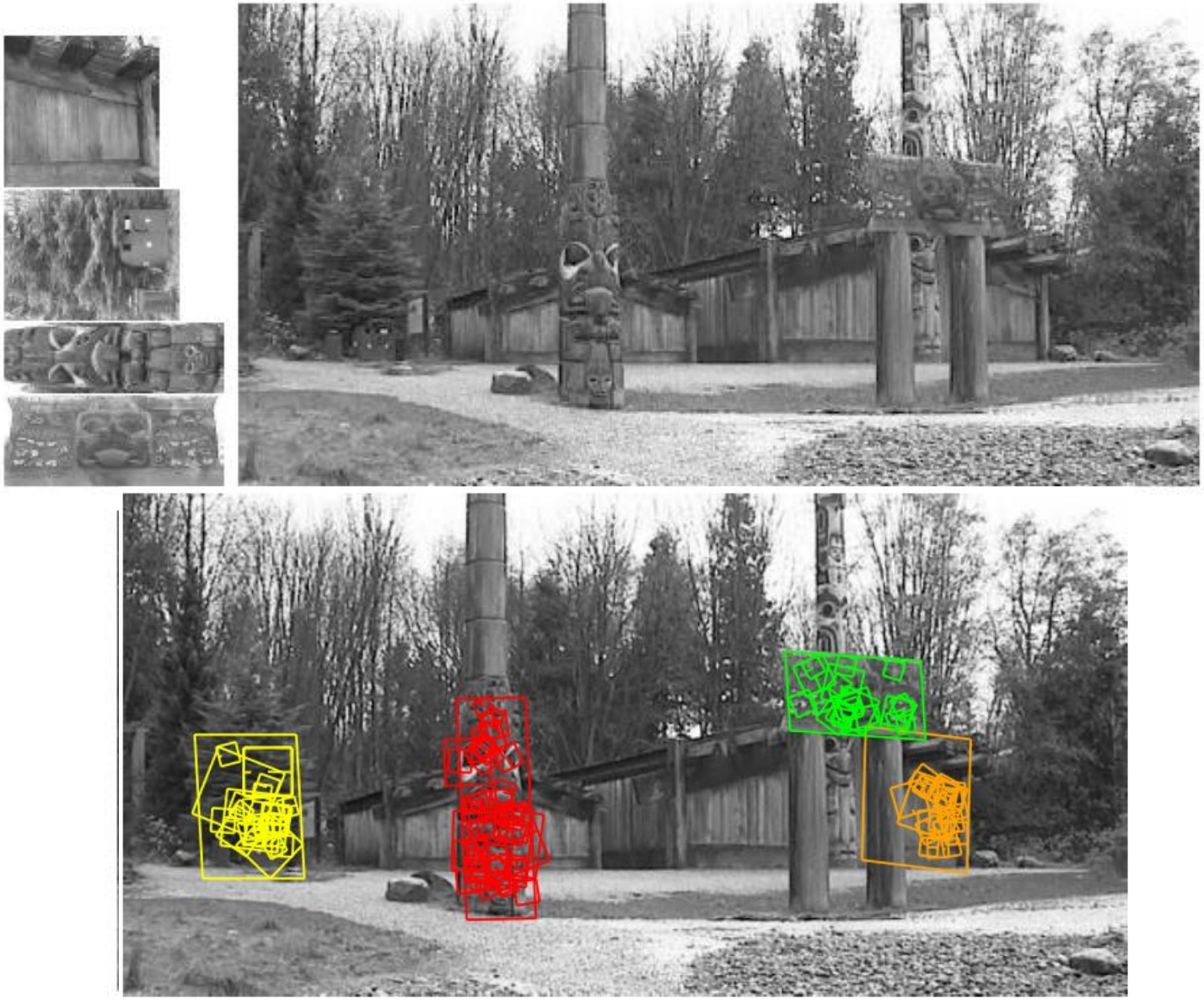


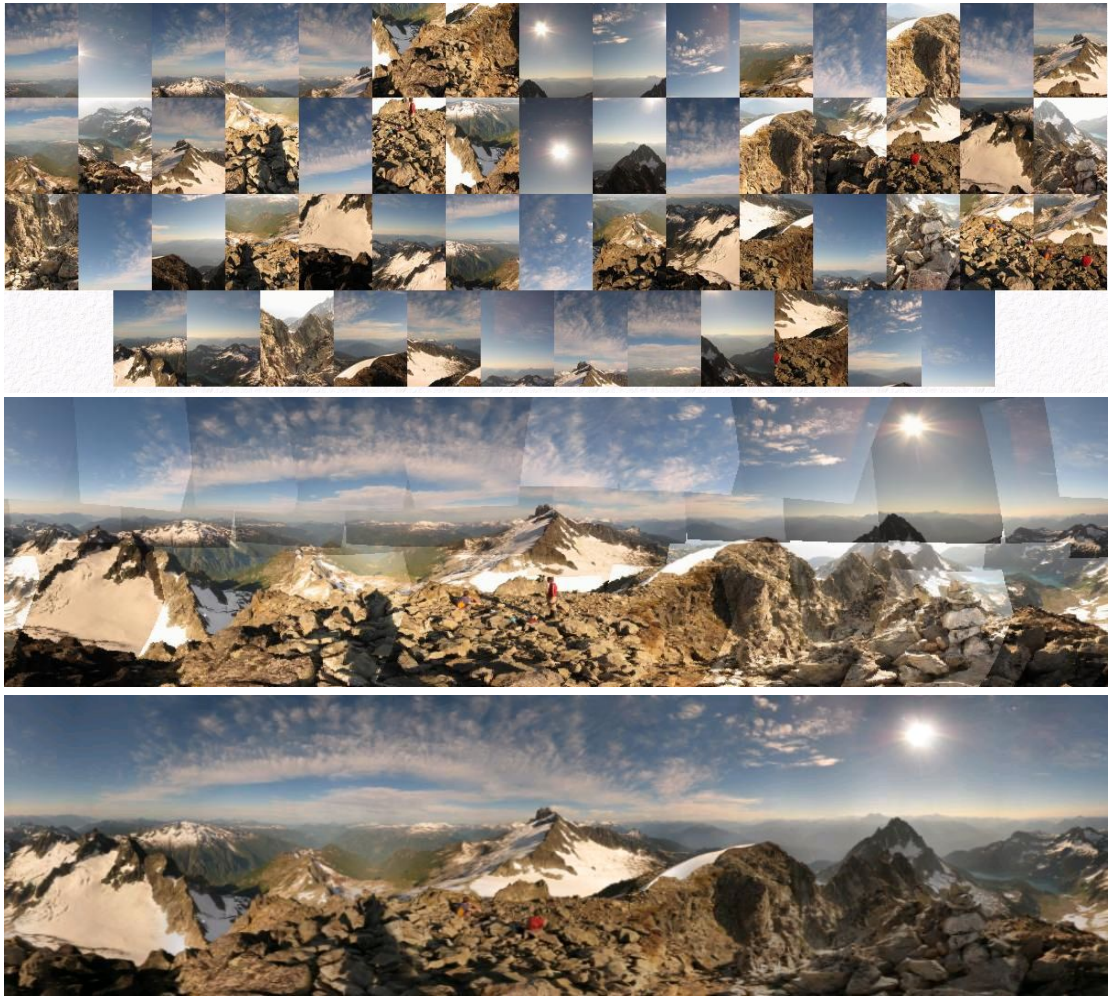
Figure 13: This example shows location recognition within a complex scene. The training images for locations are shown at the upper left and the 640x315 pixel test image taken from a different viewpoint is on the upper right. The recognized regions are shown on the lower image, with keypoints shown as squares and an outer parallelogram showing the boundaries of the training images under the affi ne transform used for recognition.

Required SIFT reading

- Wikipedia - well written, required reading): http://en.wikipedia.org/wiki/Scale-invariant_feature_transform

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Automatic mosaicing



http://www.cs.ubc.ca/~mbrown/autostitch/autostitch.html

Adapted from slides by Juan Carlos Niebles and Ranjay Krishna

Wide baseline stereo

[Image from T. Tuytelaars ECCV 2006 tutorial]

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Recognition of specific objects, scenes

Schmid and Mohr 1997

Sivic and Zisserman, 2003

Rothganger et al. 2003

Lowe 2002

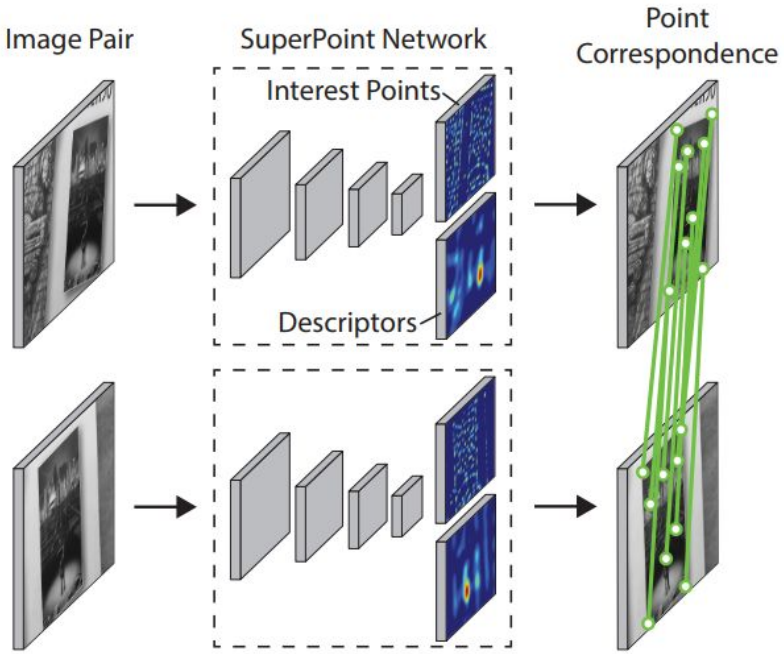
Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

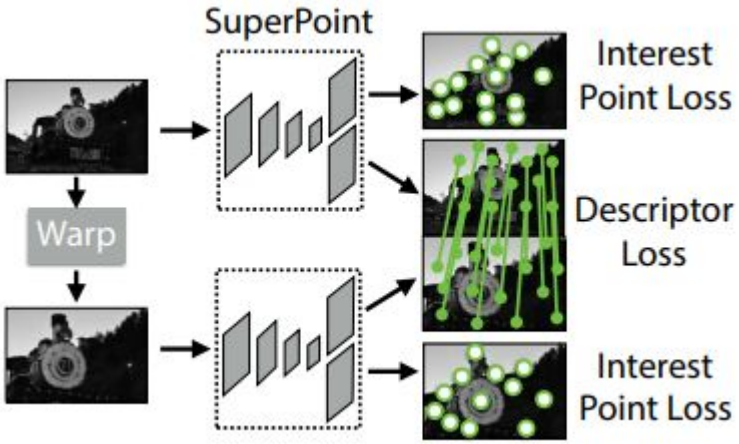
Applications of local invariant features

- Wide baseline stereo
- Motion tracking
- **Panoramas**
- Mobile robot navigation
- 3D reconstruction
- Recognition
- …

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

Deep learning based interest point detection & description





SuperPoint: Self-Supervised Interest Point Detection and Description (CVPRW'18)

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna

What we have learned

- Scale invariant region selection
  - Automatic scale selection
  - Difference-of-Gaussian (DoG) detector
- SIFT: an image region descriptor

Required reading:

- https://en.m.wikipedia.org/wiki/Scale-invariant_feature_transform

Some background reading: R. Szeliski, Ch 4.1.1; David Lowe, IJCV 2004

Adapted from slides by Juan Carlos Niebles, and Ranjay Krishna