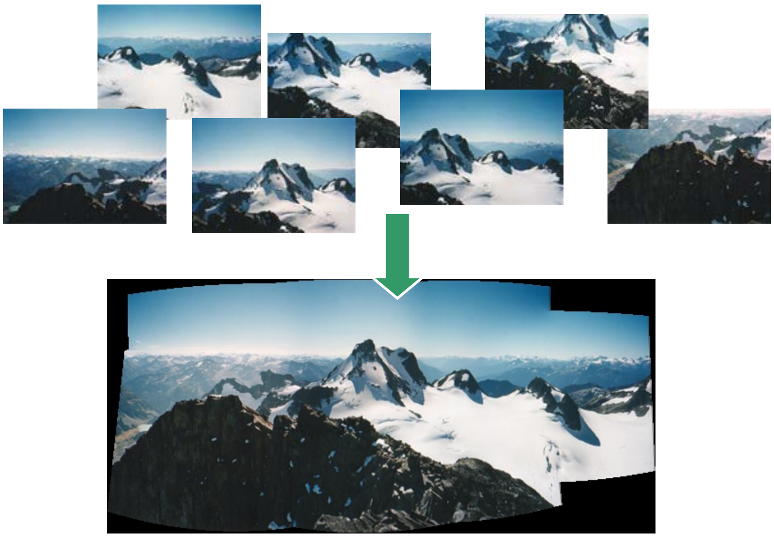
DIGITAL IMAGE PROCESSING

Lecture 9
Descriptors
Tammy Riklin Raviv
Electrical and Computer Engineering
Ben-Gurion University of the Negev

Feature detection and matching – Why?

Image stitching

Image Stitching



[Brown, Szeliski, Winder CVPR 2005]

Feature detection and matching – Why?


3D Reconstruction and Alignment

Feature detection and matching – Why?

Object detection and classification

http://inthefray.org/2015/07/strays-street-people-and-their-dogs/

Feature detection and matching – Why?

Object detection and classification

Feature detectors and descriptors

Point-like interest operators (Brown, Szeliski, and Winder 2005)

Feature detectors and descriptors

region-like interest operators (Matas, Chum, Urban et al. 2004)

Feature detectors and descriptors

Edges (Elder and Goldberg 2001)

Feature detectors and descriptors

Straight lines (Sinha, Steedly, Szeliski et al. 2008)

Finding feature points and their correspondences

- Two main approaches:
  - Find features in one image that can be accurately tracked using a local search technique, such as correlation or least squares

  Nearby viewpoints

- Independently detect features in all the images under consideration and then match features based on their local appearance

  Large distance, appearance change

Feature detection and matching

- Feature detection (extraction)
- Feature description
- Feature matching
- Feature tracking

Feature detection and matching

- Feature detection (extraction)
  - each image is searched for locations that are likely to match well in other images.
- Feature description
- Feature matching
- Feature tracking

Feature detection and matching

- Feature detection (extraction)
- Feature description
  - each region around detected keypoint locations is converted into a more compact and stable (invariant) descriptor that can be matched against other descriptors.
- Feature matching
- Feature tracking

Feature detection and matching

- Feature detection (extraction)
- Feature description
- Feature matching
  - efficiently searches for likely matching candidates in other images.
- Feature tracking

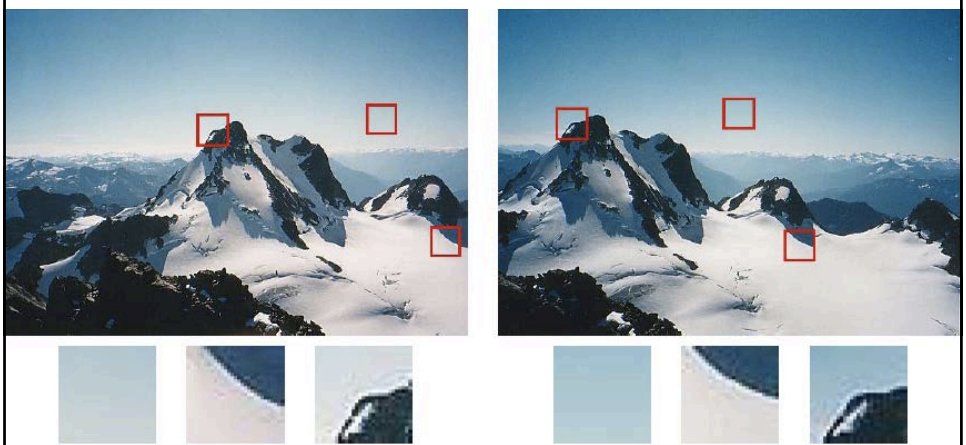
Feature detection and matching

- Feature detection (extraction)
- Feature description
- Feature matching
- Feature tracking
  - alternative to the third stage that only searches a small neighborhood around each detected feature and is therefore more suitable for video processing.

Feature detection and matching

- Feature detection (extraction)
  - each image is searched for locations that are likely to match well in other images.
- Feature description
  - each region around detected keypoint locations is converted into a more compact and stable (invariant) descriptor that can be matched against other descriptors.
- Feature matching
  - efficiently searches for likely matching candidates in other images.
- Feature tracking
  - alternative to the third stage that only searches a small neighborhood around each detected feature and is therefore more suitable for video processing.

What are good key-points (patches)?



Comparing two image patches

$$E_{\mathrm{WSSD}}(\boldsymbol{u}) = \sum_i w(\boldsymbol{x}_i)[I_1(\boldsymbol{x}_i + \boldsymbol{u}) - I_0(\boldsymbol{x}_i)]^2$$

Weighted Sum Square Differences (WSSD)

I_0, I_1 two images being compared

$\mathbf{u} = (u, v)$ displacement vector

$w(x_i)$ Spatially varying weighting function

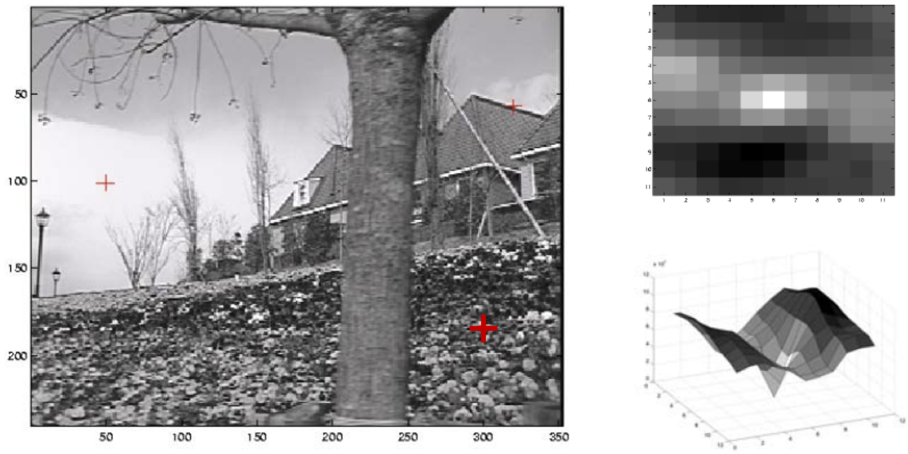
Comparing an image patch against itself

$$E_{\mathrm{AC}}(\Delta\boldsymbol{u}) = \sum_i w(\boldsymbol{x}_i)[I_0(\boldsymbol{x}_i + \Delta\boldsymbol{u}) - I_0(\boldsymbol{x}_i)]^2$$

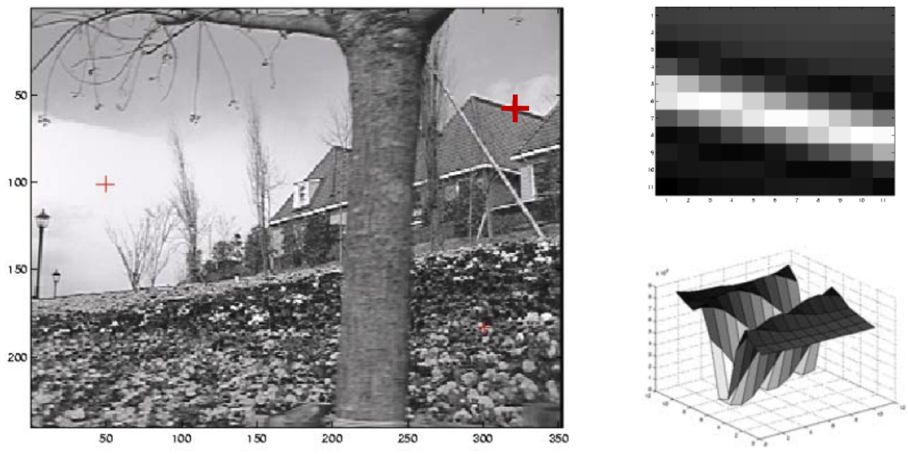
an auto-correlation function or surface

Measure how stable this metric with respect to small variations in positions $\Delta\mathbf{u}$

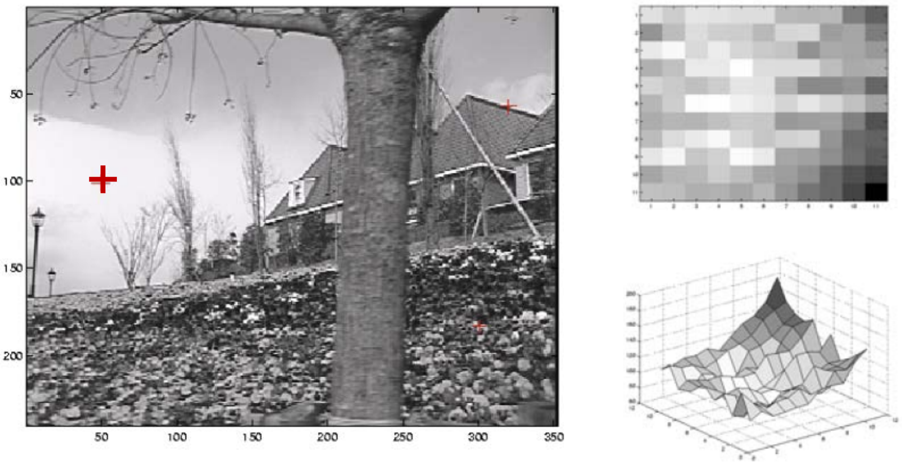
Auto-correlation surfaces



Auto-correlation surfaces



Auto-correlation surfaces



Auto-correlation surfaces

Using a Taylor Series expansion of the image function

$$I_0(\boldsymbol{x}_i + \Delta\boldsymbol{u}) \approx I_0(\boldsymbol{x}_i) + \nabla I_0(\boldsymbol{x}_i) \cdot \Delta\boldsymbol{u}$$

we can approximate the auto-correlation surface as

$$\begin{aligned} E_{\mathrm{AC}}(\Delta\boldsymbol{u}) &= \sum_i w(\boldsymbol{x}_i)[I_0(\boldsymbol{x}_i + \Delta\boldsymbol{u}) - I_0(\boldsymbol{x}_i)]^2 \\ &\approx \sum_i w(\boldsymbol{x}_i)[I_0(\boldsymbol{x}_i) + \nabla I_0(\boldsymbol{x}_i) \cdot \Delta\boldsymbol{u} - I_0(\boldsymbol{x}_i)]^2 \\ &= \sum_i w(\boldsymbol{x}_i)[\nabla I_0(\boldsymbol{x}_i) \cdot \Delta\boldsymbol{u}]^2 \\ &= \Delta\boldsymbol{u}^T \boldsymbol{A} \Delta\boldsymbol{u}, \end{aligned}$$

where, $\nabla I_0(\boldsymbol{x}_i) = (\frac{\partial I_0}{\partial x}, \frac{\partial I_0}{\partial y})(\boldsymbol{x}_i)$ is the image gradient at x_i .

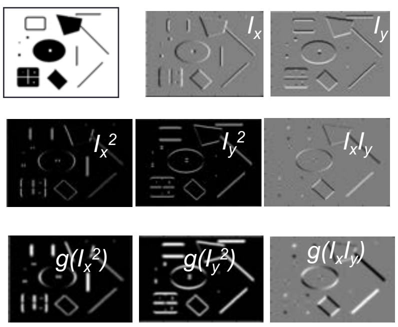
Auto-correlation surfaces

The auto-correlation matrix A can be written as

$$\boldsymbol{A} = w * \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} := \begin{bmatrix} \sum_{(x,y)\in W} I_x^2 & \sum_{(x,y)\in W} I_x I_y \\ \sum_{(x,y)\in W} I_x I_y & \sum_{(x,y)\in W} I_y^2 \end{bmatrix}$$

As first shown by Anandan (1984; 1989) that the inverse of the matrix A provides a lower bound on the uncertainty in the location of a matching patch.
It is therefore a useful indicator of which patches can be reliably matched. See examples

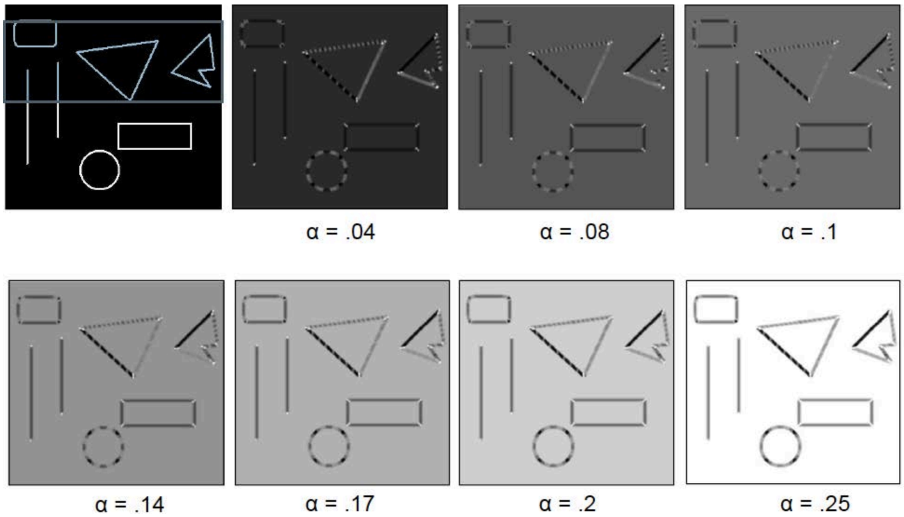
Harris Feature detector (Harris 88)



$$\boldsymbol{A} = w * \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}$$

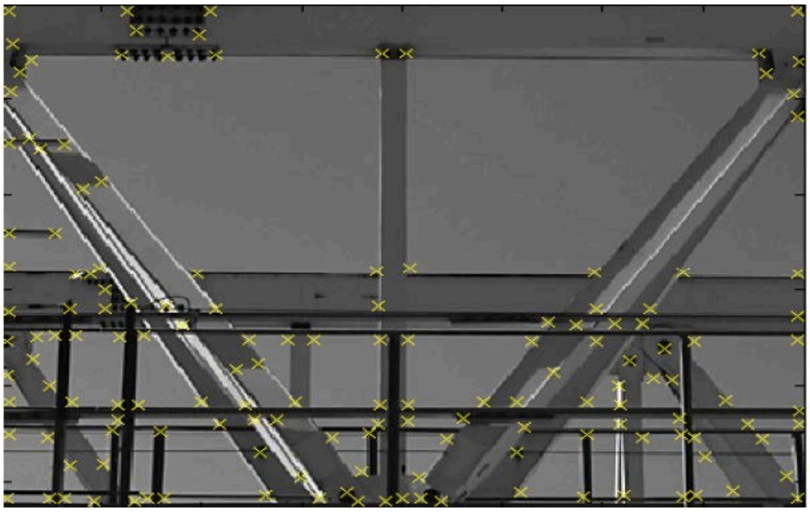
Harr= $\det(\boldsymbol{A}) - \alpha \operatorname{trace}(\boldsymbol{A})^2 = \lambda_0\lambda_1 - \alpha(\lambda_0 + \lambda_1)^2$

Cornerness – Harris Corner

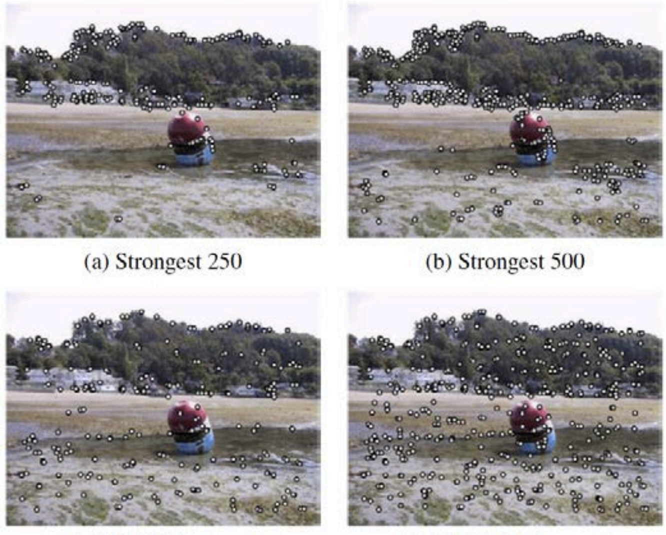


Fei-Fei Li

Example: Harris Corner



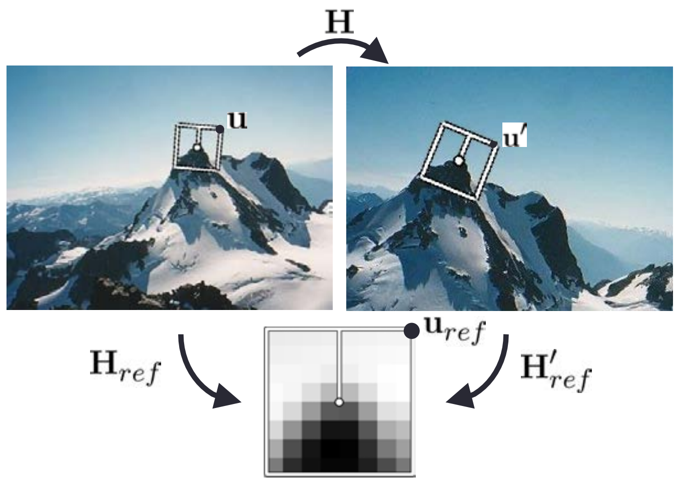
Adaptive non-maximal suppression (ANMS, Brown, Szeliski, and Winder 2005)



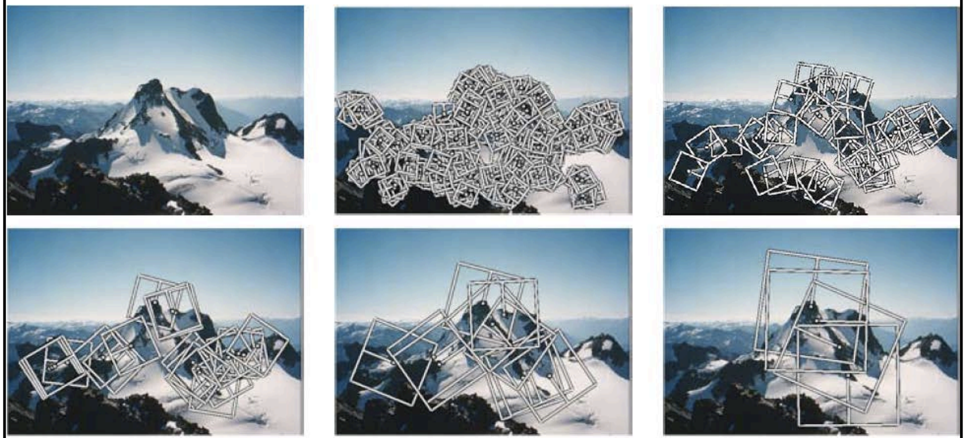
(a) Strongest 250 (b) Strongest 500

(c) ANMS 250, $r = 24$ (d) ANMS 500, $r = 16$

Rotation Invariance (Brown et al)



Scale Invariance



Multi-scale oriented patches (MOPS) extracted at five pyramid levels (Brown, Szeliski, and Winder 2005). The boxes show the feature orientation and the region from which the descriptor vectors are sampled.

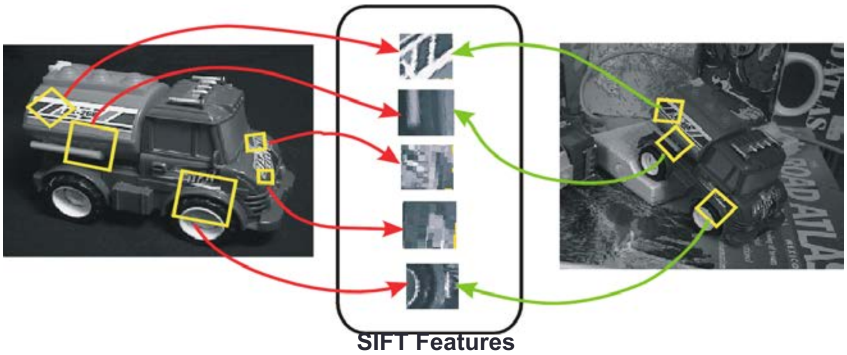
SIFT: Motivation

- The Harris operator is not invariant to scale and correlation is not invariant to rotation[1].
- For better image matching, Lowe's goal was to develop an interest operator that is invariant to scale and rotation.
- Also, Lowe aimed to create a descriptor that was robust to the variations corresponding to typical viewing conditions. The descriptor is the most-used part of SIFT.

[1]But Schmid and Mohr developed a rotation invariant descriptor for it in 1997.

Idea of SIFT

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



39

Claimed Advantages of SIFT

- **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

Overall Procedure at a High Level

1. Scale-space extrema detection

   Search over multiple scales and image locations.

2. Keypoint localization

   Fit a model to determine location and scale.
   Select key points based on a measure of stability.

3. Orientation assignment

   Compute best orientation(s) for each key point region.

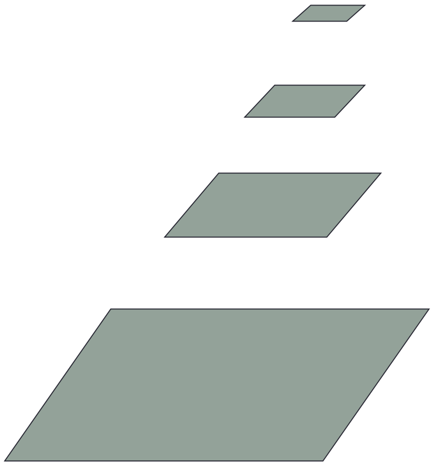
4. Key point description

   Use local image gradients at selected scale and rotation to describe each key point region.

1. Scale-space extrema detection

- Goal: Identify locations and scales that can be repeatably assigned under different views of the same scene or object.
- Method: search for stable features across multiple scales using a continuous function of scale.
- Prior work has shown that under a variety of assumptions, the best function is a Gaussian function.
- The scale space of an image is a function $L(x,y,\sigma)$ that is produced from the convolution of a Gaussian kernel (at different scales) with the input image.

Aside: Image Pyramids



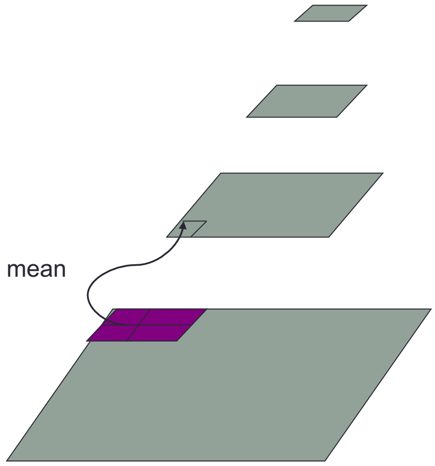
And so on.

3^{rd} level is derived from the 2^{nd} level according to the same funtion

2^{nd} level is derived from the original image according to some function

Bottom level is the original image.

Aside: Mean Pyramid



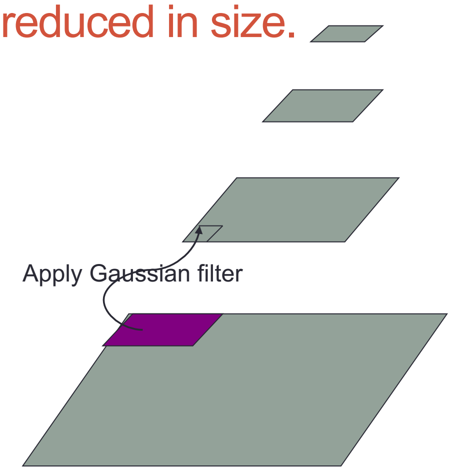
And so on.

At 3^{rd} level, each pixel is the mean of 4 pixels in the 2^{nd} level.

At 2^{nd} level, each pixel is the mean of 4 pixels in the original image.

Bottom level is the original image.

Aside: Gaussian Pyramid
At each level, image is smoothed and reduced in size.

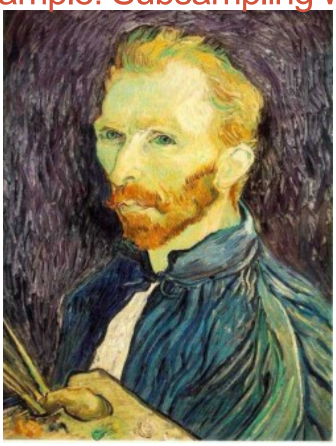


And so on.

At 2nd level, each pixel is the result of applying a Gaussian mask to the first level and then subsampling to reduce the size.

Bottom level is the original image.

Example: Subsampling with Gaussian pre-filtering



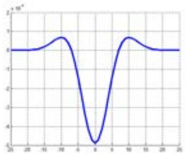
Gaussian 1/2

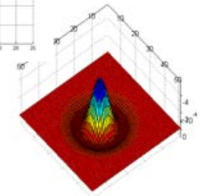
G 1/4

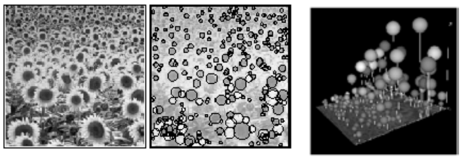
G 1/8

Lowe's Scale-space Interest Points

- **Laplacian of Gaussian** kernel
  - Scale normalized (x by scale^2)
  - Proposed by Lindeberg
- Scale-space detection
  - Find local maxima across scale/space
  - A good "blob" detector





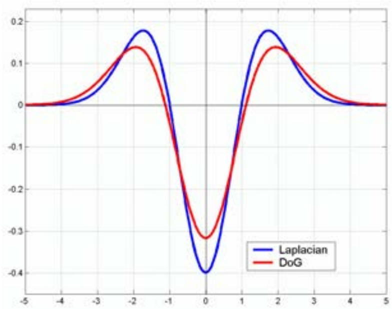


$$G(x, y, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{x^2+y^2}{\sigma^2}}$$

$$\nabla^2 G(x, y, \sigma) = \frac{\partial^2 G}{\partial x^2} + \frac{\partial^2 G}{\partial y^2}$$

[T. Lindeberg IJCV 1998]

Lowe's Scale-space Interest Points: Difference of Gaussians



- Gaussian is an ad hoc solution of heat diffusion equation

$$\frac{\partial G}{\partial \sigma} = \sigma \nabla^2 G.$$

- Hence

$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k-1)\sigma^2 \nabla^2 G.$$

- k is not necessarily very small in practice

Lowe's Pyramid Scheme

- Scale space is separated into octaves:
  - Octave 1 uses scale σ
  - Octave 2 uses scale 2σ
  - etc.
- In each octave, the initial image is repeatedly convolved with Gaussians to produce a set of scale space images.
- Adjacent Gaussians are subtracted to produce the DOG
- After each octave, the Gaussian image is down-sampled by a factor of 2 to produce an image ¼ the size to start the next level.

Lowe's Pyramid Scheme

s+2 filters

$\sigma_{s+1}=2^{(s+1)/s}\sigma_0$

.

.

$\sigma_i=2^{i/s}\sigma_0$

.

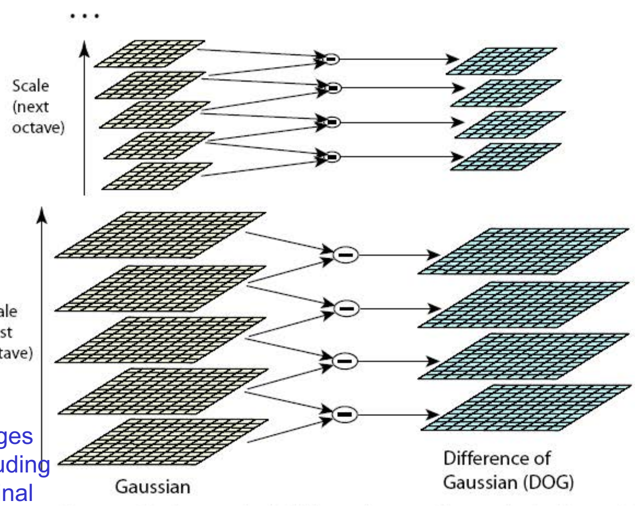
.

$\sigma_2=2^{2/s}\sigma_0$

$\sigma_1=2^{1/s}\sigma_0$

σ_0

s+3 images including original



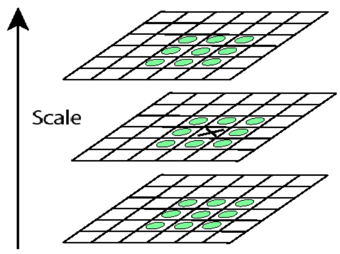
s+2 differ-ence images

The parameter **s** determines the number of images per octave.

Key point localization

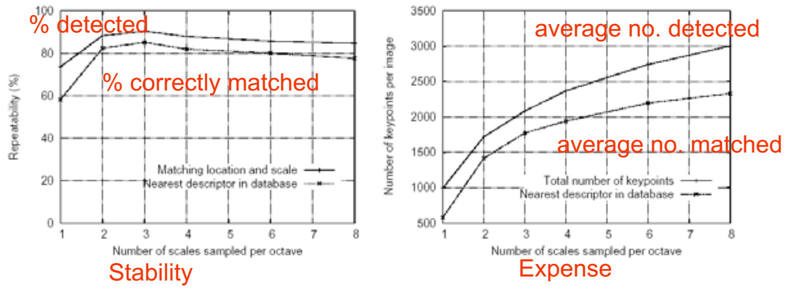
s+2 difference images.
top and bottom ignored.
s planes searched.

- Detect maxima and minima of difference-of-Gaussian in scale space
- Each point is compared to its 8 neighbors in the current image and 9 neighbors each in the scales above and below



For each max or min found, output is the **location** and the **scale**.

Scale-space extrema detection: experimental results over 32 images that were synthetically transformed and noise added.



- Sampling in scale for efficiency
  - How many scales should be used per octave? S=?
    - More scales evaluated, more keypoints found
    - S < 3, stable keypoints increased too
    - S > 3, stable keypoints decreased
    - S = 3, maximum stable keypoints found

2. Keypoint localization

- Once a keypoint candidate is found, perform a detailed fit to nearby data to determine
  - location, scale, and ratio of principal curvatures
- In initial work keypoints were found at location and scale of a central sample point.
- In newer work, they fit a 3D quadratic function to improve interpolation accuracy.
- The Hessian matrix was used to eliminate edge responses.

Eliminating the Edge Response

- Reject flats:
  - $|D(\hat{\mathbf{x}})|$ < 0.03
- Reject edges:

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}$$

Let α be the eigenvalue with larger magnitude and β the smaller.

$$\mathrm{Tr}(\mathbf{H}) = D_{xx} + D_{yy} = \alpha + \beta,$$

$$\mathrm{Det}(\mathbf{H}) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta.$$

Let r = α/β.
So α = rβ

$$\frac{\mathrm{Tr}(\mathbf{H})^2}{\mathrm{Det}(\mathbf{H})} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r},$$

$(r+1)^2/r$ is at a min when the 2 eigenvalues are equal.

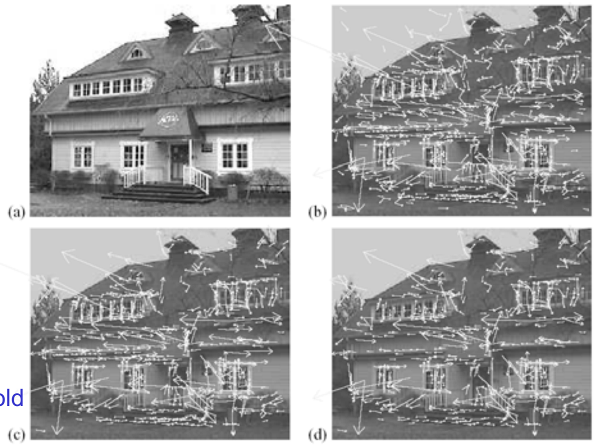
  - r < 10
- What does this look like?

Keypoint localization with orientation

233x189

832

initial keypoints



729

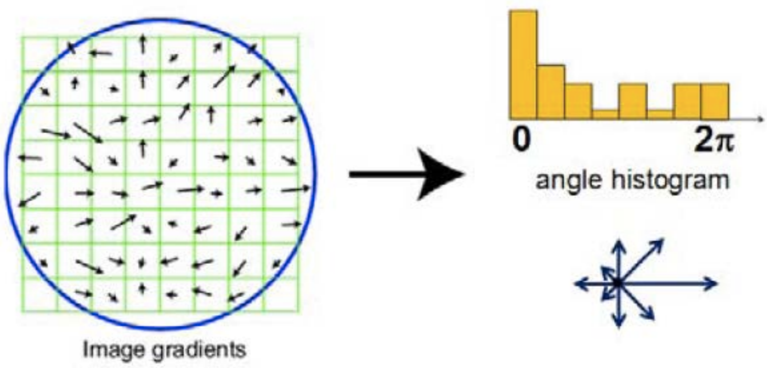
keypoints after gradient threshold

536

keypoints after ratio threshold

3.Orientation estimation

If 2 major orientations, use both.

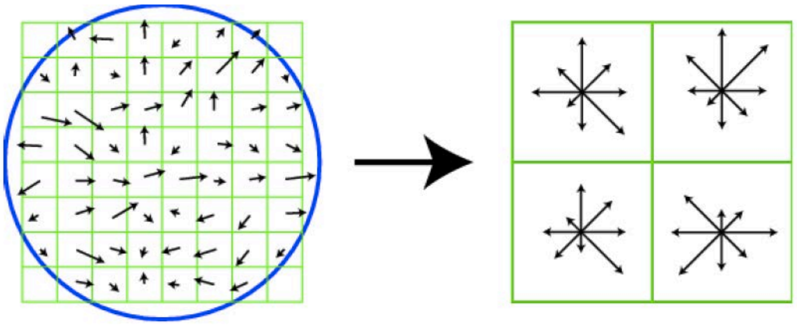


Create histogram of local gradient directions at selected scale
Assign canonical orientation at peak of smoothed histogram
Each key specifies stable 2D coordinates (x, y, scale,orientation)

4. Keypoint Descriptors

- At this point, each keypoint has
  - location
  - scale
  - orientation
- Next is to compute a descriptor for the local image region about each keypoint that is
  - highly distinctive
  - invariant as possible to variations such as changes in viewpoint and illumination

SIFT



(a) image gradients (b) keypoint descriptor

A schematic representation of Lowe's (2004) scale invariant feature transform (SIFT): (a) Gradient orientations and magnitudes are computed at each pixel and weighted by a Gaussian fall-off function (blue circle). (b) A weighted gradient orientation histogram is then computed in each sub-region, using trilinear interpolation. While this figure shows an 8 X8 pixel patch and a 2X2 descriptor array, Lowe's actual implementation uses 16X16 patches and a 4 4 array of eight-bin histograms.

SIFT Keypoint Descriptor

- use the normalized region about the keypoint
- compute gradient magnitude and orientation at each point in the region
- weight them by a Gaussian window overlaid on the circle
- create an orientation histogram over the 4 X 4 subregions of the window
- 4 X 4 descriptors over 16 X 16 sample array were used in practice. 4 X 4 times 8 directions gives a vector of 128 values.

SIFT Results

Panorama

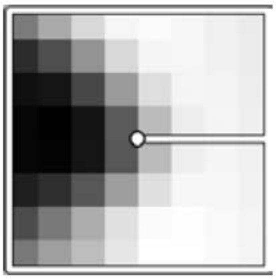
SIFT Results: Matching "Objects"

SIFT Results: Recognizing objects in clutter scenes

Feature Descriptors (other than SIFT)

- Multiscale Oriented Patches (MOPs).
- Scale invariant feature transform (MSERs)
- PCA-SIFT
- Gradient location-orientation histogram (GLOH).
- Histograms of Oriented Gradients (HOGs)
- Speeded Up Robust Features (SURF)
- and many others …
- (e.g. BRISK)

MOPs Descriptors



MOPS descriptors are formed using an 8 x8 sampling of bias and gain normalized intensity values, with a sample spacing of five pixels relative to the detection scale. This low frequency sampling gives the features some robustness to interest point location error and is achieved by sampling at a higher pyramid level than the detection scale.

Maximally stable extremal regions (MSERs)

MSER

Binary regions are computed by thresholding the image at all possible gray levels

- This operation can be performed efficiently by
- first sorting all pixels by gray value and then incrementally adding pixels to each connected component

As the threshold is changed, the area of each component (region) is monitored; regions whose rate of change of area with respect to the threshold is minimal are defined as maximally stable.

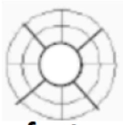
Matal et al, 2004

Gradient location-orientation histogram (GLOH) descriptor

First 3 steps – same as SIFT

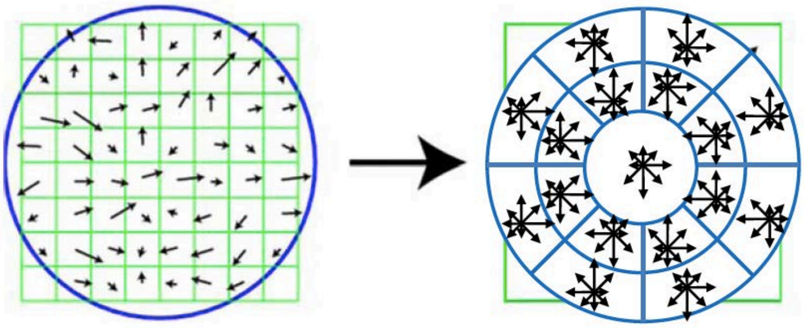
Step 4 – Local image descriptor

- **Consider log-polar location grid with 3 different radii and 8 angular direction for two of them, in total 17 location bin**
- **Form histogram of gradients having 16 bins**
- **Form a feature vector of 272 dimension (17*16)**
- **Perform dimensionality reduction and project the features to a 128 dimensional space.**



Mikolajczyk and Schmid (2005),

Gradient location-orientation histogram (GLOH) descriptor



(a) image gradients (b) keypoint descriptor

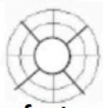
The gradient location-orientation histogram (GLOH) descriptor uses log-polar bins instead of square bins to compute orientation histograms (Mikolajczyk and Schmid 2005).

GLOH

First 3 steps – same as SIFT

Step 4 – Local image descriptor

- **Consider log-polar location grid with 3 different radii and 8 angular direction for two of them, in total 17 location bin**
- **Form histogram of gradients having 16 bins**
- **Form a feature vector of 272 dimension (17*16)**
- **Perform dimensionality reduction and project the features to a 128 dimensional space.**



192 correct matches (yellow) and 208 false matches (blue).

Histogram of Oriented Gradients Descriptors (Hogs)

- Local object appearance and shape within an image are described by the distribution of intensity gradients or edge directions.
- The image is divided into small connected regions called cells, and for the pixels within each cell, a histogram of gradient directions is compiled.
- The descriptor is the concatenation of these histograms.
- For improved accuracy, the local histograms are
- contrast-normalized by calculating a measure of the intensity across a larger region of the image, called a block, and then using this value to normalize all cells within the block.
- This normalization results in better invariance to changes in illumination and shadowing.

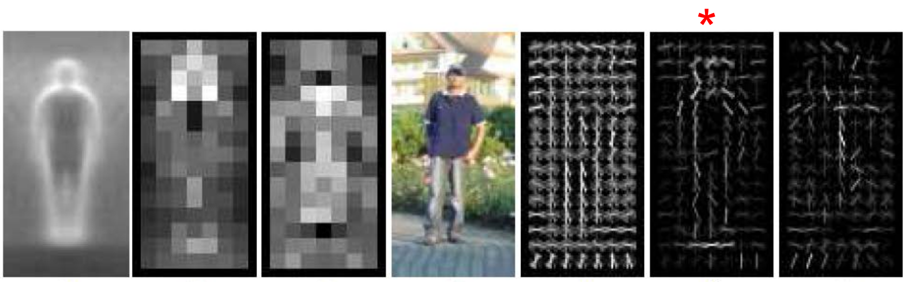
HOGs – Block Normalization

L2-norm: $$f = \frac{v}{\sqrt{\|v\|_2^2 + e^2}}$$

L1-norm: $$f = \frac{v}{(\|v\|_1 + e)}$$

L1-sqrt: $$f = \sqrt{\frac{v}{(\|v\|_1 + e)}}$$

Hogs



(a) (b) (c) (d) (e) (f) (g)

(a) average gradient image over training examples
(b) each “pixel” shows max positive SVM weight in the block centered on that pixel
(c) same as (b) for negative SVM weights
(d) test image
(e) its R-HOG descriptor
(f) R-HOG descriptor weighted by positive SVM weights
(g) R-HOG descriptor weighted by negative SVM weights

HOGs Examples

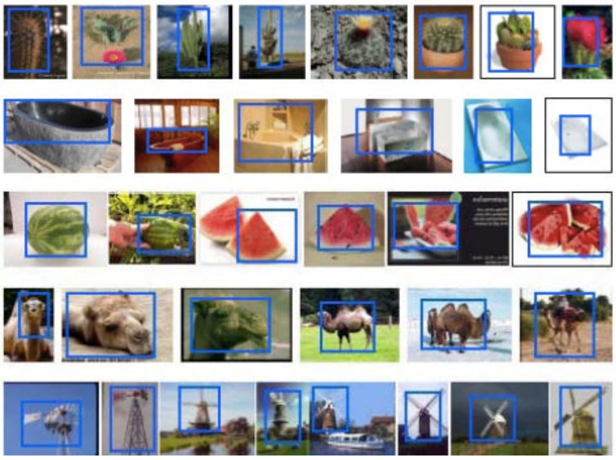


adapted from Fei-Fei Li

SURF example

adapted from Fei-Fei Li

Example: Pyramid Histogram Of Words (PHOW)



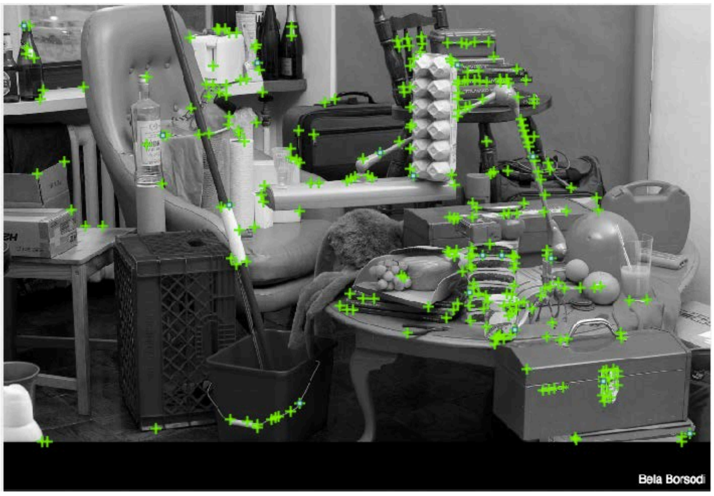
Bosch et al, ICCV 2007 (variant of dense SIFT descriptor)

Features in Matlab

[FEATURES, VALID_POINTS] = extractFeatures(I, POINTS, Name, Value)

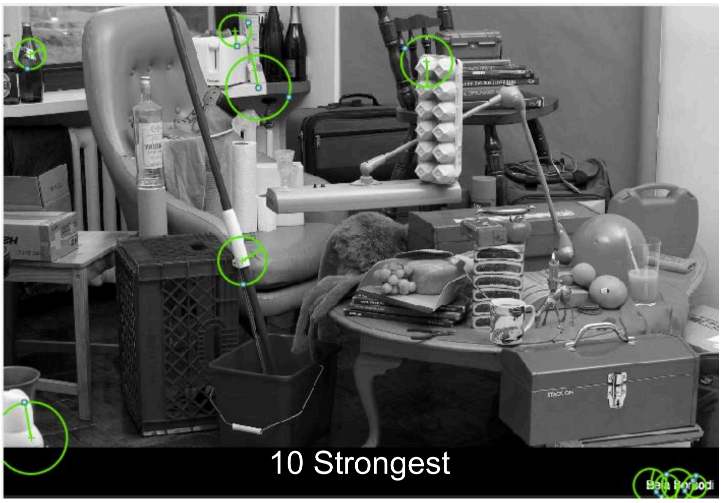
```
Class of POINTS              Descriptor extraction method
---------------              ---------------------------------
- SURFPoints object          - Speeded-Up Robust Features (SURF)
- MSERRegions object         - Speeded-Up Robust Features (SURF)
- cornerPoints object        - Fast Retina Keypoint (FREAK)
- BRISKPoints object         - Fast Retina Keypoint (FREAK)
- M-by-2 matrix of [x y]     - Simple square neighborhood around [x y]
  coordinates                  point location

 Method      Feature vector (descriptor)
 -------     ----------------------------------
 'BRISK'     Binary Robust Invariant Scalable Keypoints (BRISK)
 'FREAK'     Fast Retina Keypoint (FREAK)
 'SURF'      Speeded-Up Robust Features (SURF)
 'Block'     Simple square neighborhood
 'Auto'      Selects the extraction method based on the class of
             input points. See the table above.

             Default: 'Auto'
```

Example: Harris Corner Detector



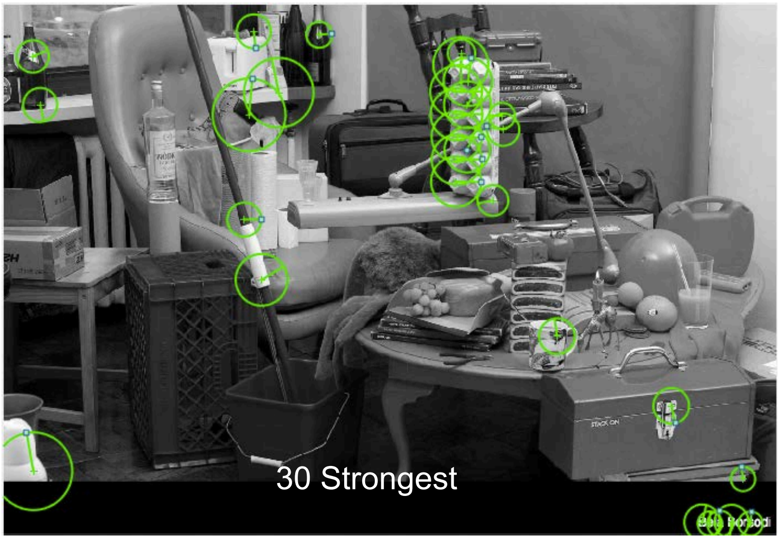
corners = detectHarrisFeatures(I);

Example: SURF features

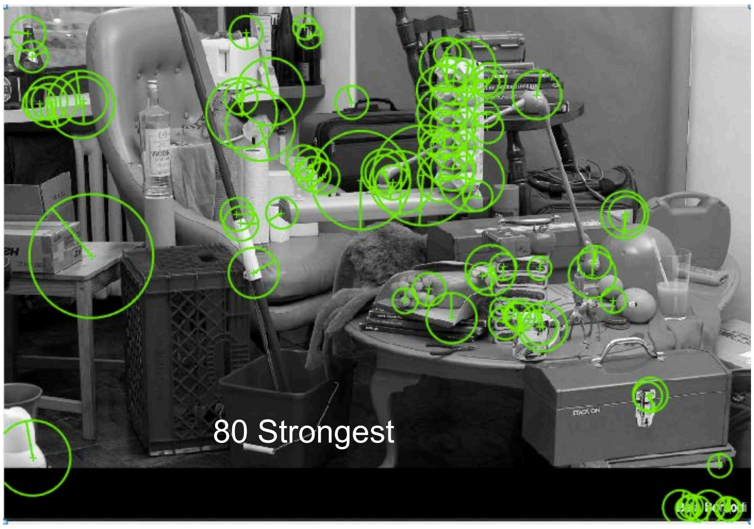


points = detectSURFFeatures(I);

Example: SURF features



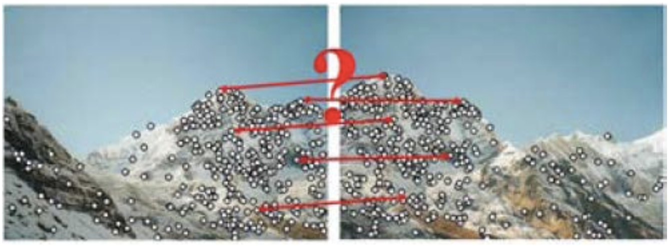
Example: SURF features



Example: MSER with upright SURF feature descriptor

regions = detectMSERFeatures(I);

Feature Matching



how can we extract local descriptors that are invariant
to inter-image variations and yet still discriminative enough to establish
correct correspondences?

Matching strategy and error rates

- Context and application dependent
  - matching a pair of images with large overlap
  - object detection
- Euclidean distances in feature space can be directly used for ranking potential matches.
- Thresholding

Performance quantification of matching algorithms

TP: true positives, i.e., number of correct matches;

FN: false negatives, matches that were not correctly detected;

FP: false positives, proposed matches that are incorrect;

TN: true negatives, non-matches that were correctly rejected.

Performance quantification of matching algorithms

- true positive rate (TPR),

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP{+}FN}} = \frac{\mathrm{TP}}{\mathrm{P}};$$

- false positive rate (FPR),

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP{+}TN}} = \frac{\mathrm{FP}}{\mathrm{N}};$$

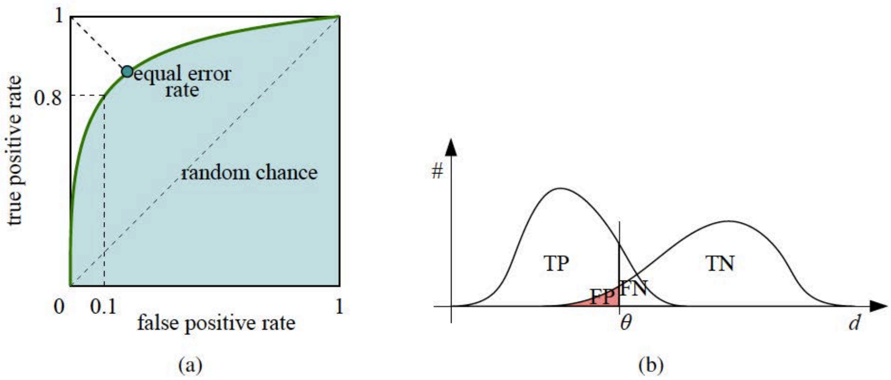
- positive predictive value (PPV),

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP{+}FP}} = \frac{\mathrm{TP}}{\mathrm{P'}};$$

- accuracy (ACC),

$$\mathrm{ACC} = \frac{\mathrm{TP{+}TN}}{\mathrm{P{+}N}}.$$

ROC curve and its related rates



Efficient Matching

- Multi-dimensional search tree
- Hash table