Scale Invariant Feature Transform

CS 783: Visual Recognition

Overview

- Scale Space Extrema Detection
- Keypoint Localization
- Orientation Assignment
- Keypoint Descriptor

Last Class

- We initially obtained a measure for identifying corners using Harris Corner operator
- This however is not scale invariant
- We then considered the SIFT operator
- In this class, we will understand the SIFT in some more detail

Scale Space

- To obtain a scale space, certain axioms need to be fulfilled such as
    - non-creation of local extrema (zero-crossings)
    - non-enhancement of local extrema in any number of dimensions at spatial maxima and at spatial minima
    - Koenderink in his structure of images showed that this could be obtained through a Gaussian kernel

Scale Space

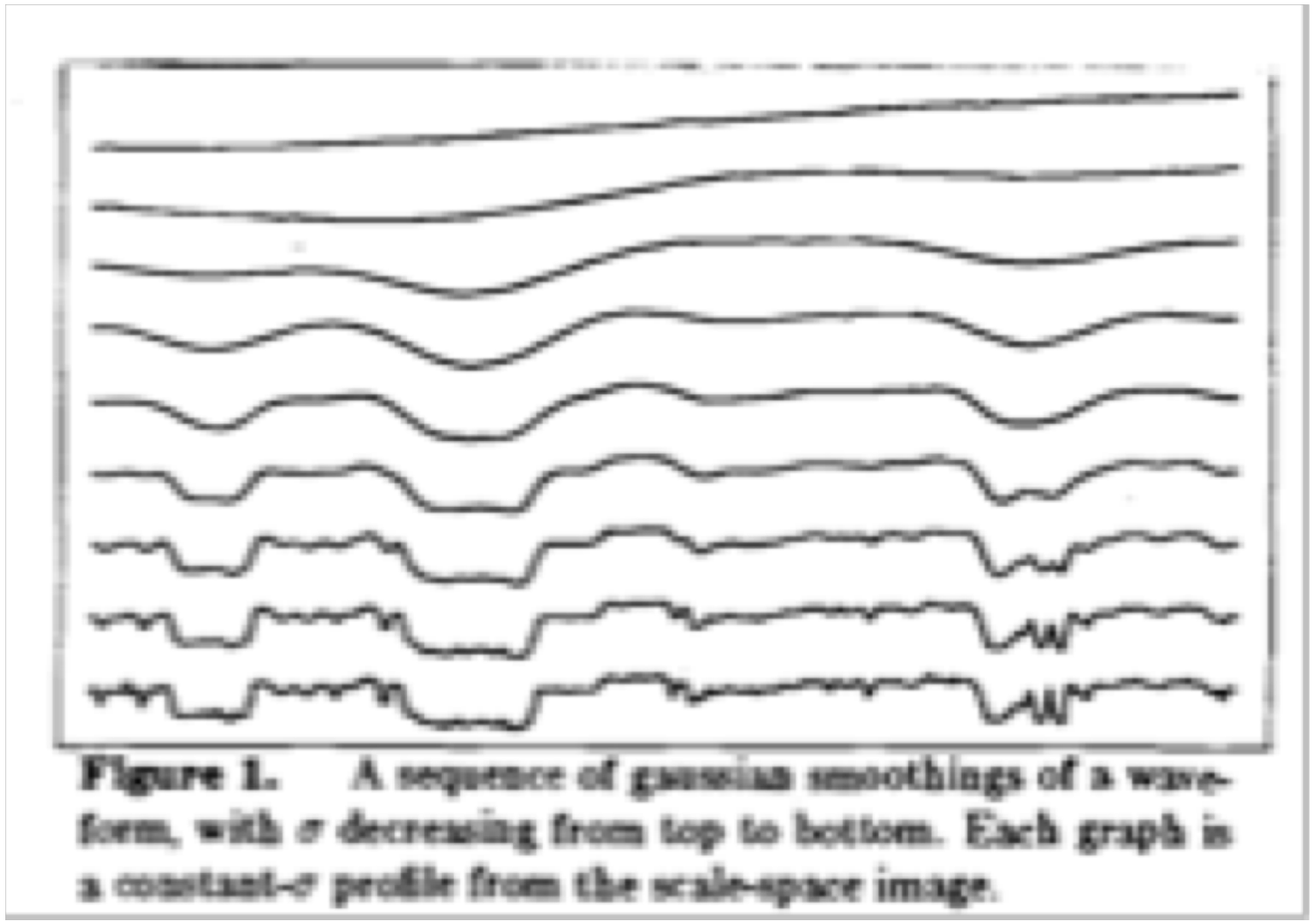
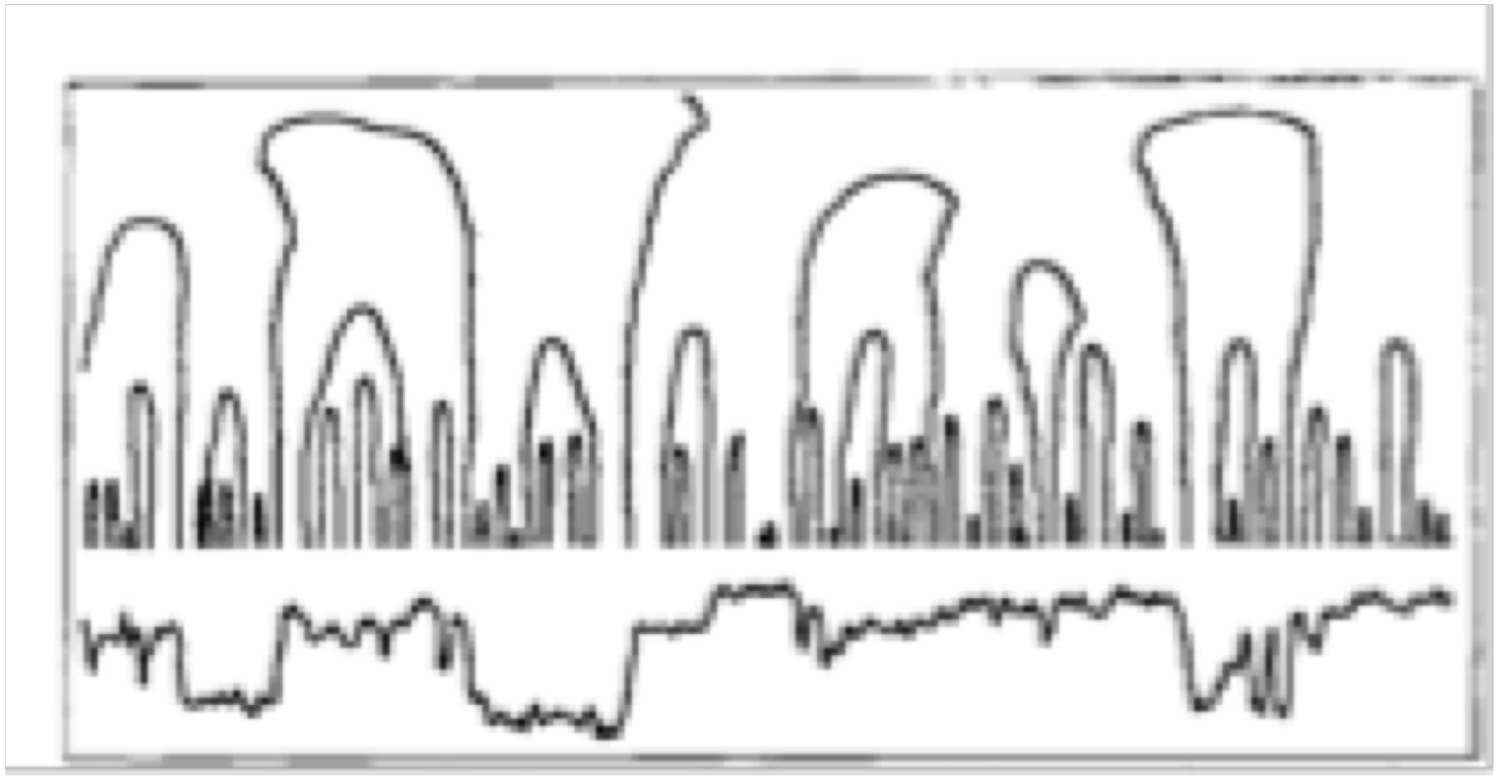


Figure 1. A sequence of gaussian smoothings of a waveform, with σ decreasing from top to bottom. Each graph is a constant-σ profile from the scale-space image.

Scale Space filtering by Andrew Within (1983) showed also that the scale space could be obtained using Gaussian. The figure above shows the extrema changing with changing sigma (increasing sigma from bottom to top)

Scale Space



Scale Space filtering by Andrew Within (1983) showed also that the scale space could be obtained using Gaussian. The figure above shows the extrema changing with changing sigma (increasing sigma from bottom to top)

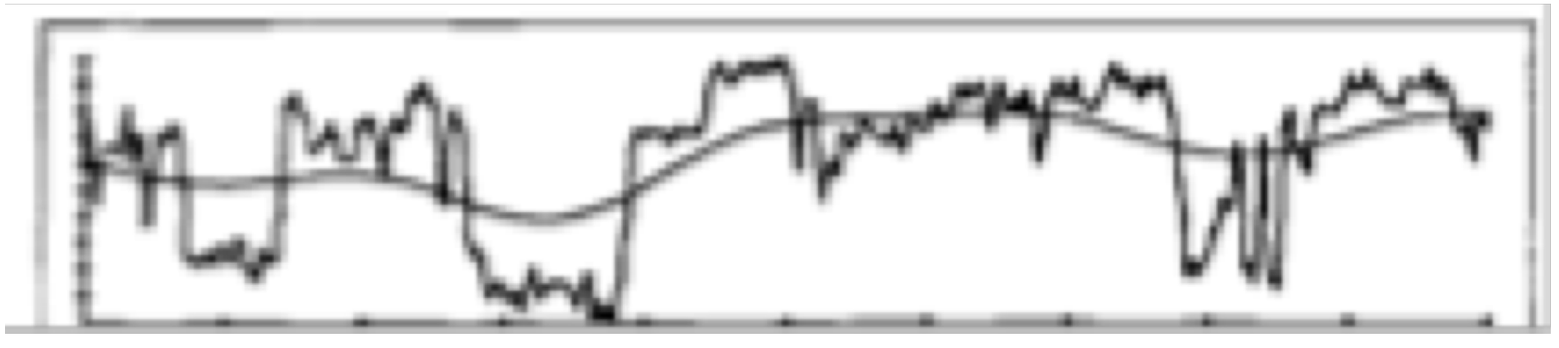


Figure shows a waveform with its coarser Gaussian super-imposed (Witkin 1983)

Scale Space

- Scale space in SIFT is constructed using Difference of Gaussians

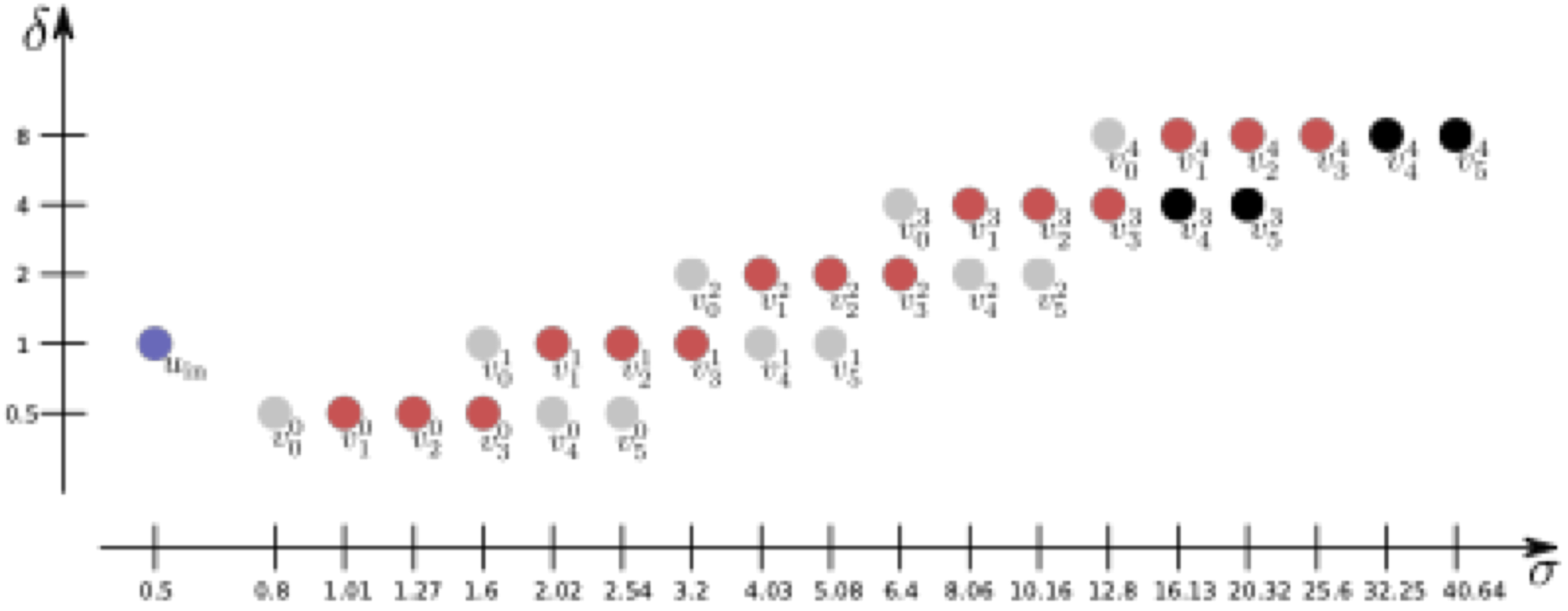
- Truncated discrete Gaussian filter is convolved with the image to obtain the set of Difference of Gaussian images

- After computation of the scale space, extrema is located by detecting the point having local maxima in a 3x3x3 neighbourhood space

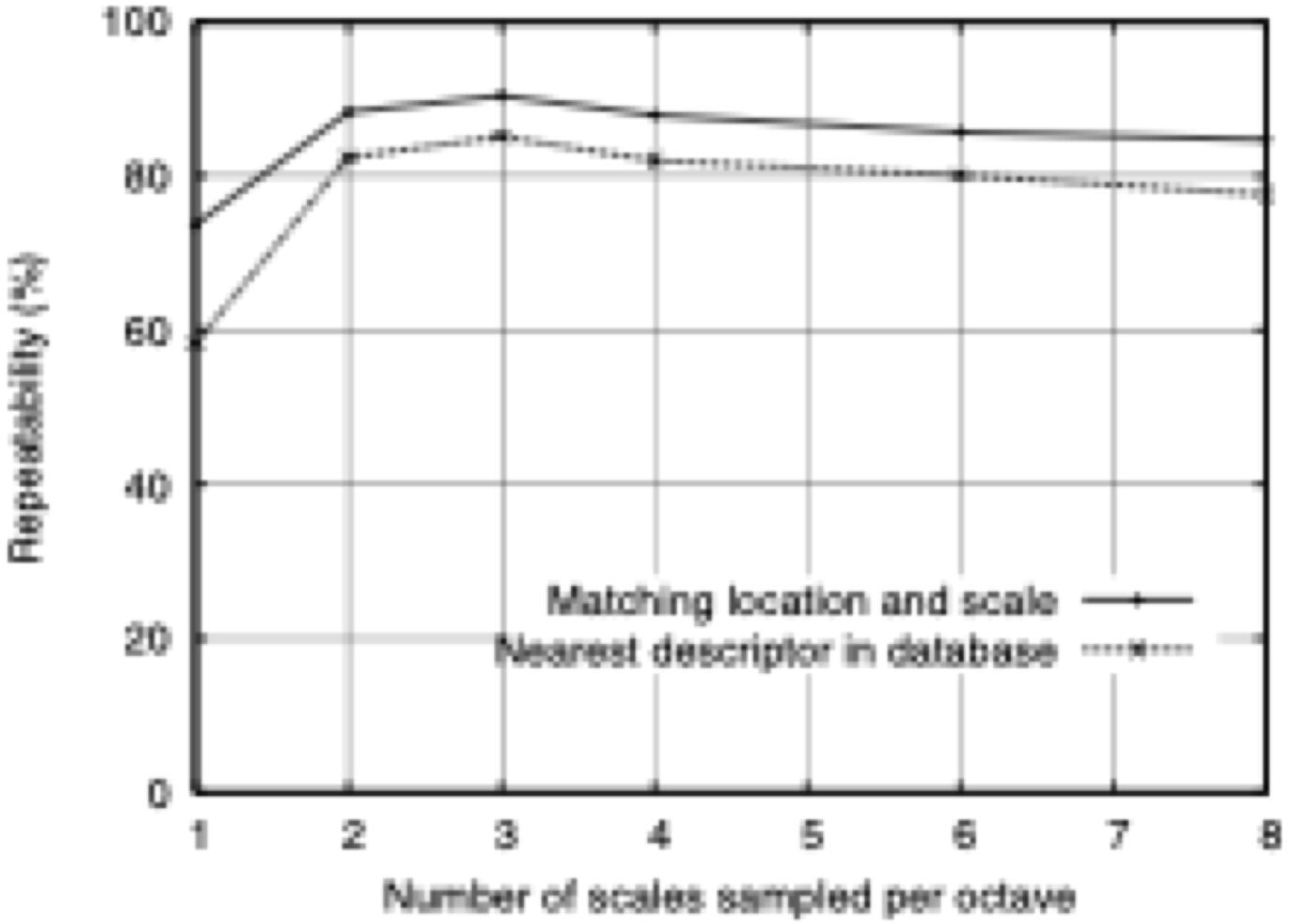
Scale Space Generation

- Scale space generation is done on the basis of octaves
- Each octave denotes a doubling of the σ
- Once a complete octave has been processed, the image is subsampled by half (using bilinear interpolation)
- The accuracy for sampling σ is not different while the computation is reduced

Scale Space Generation



Number of scales per octave



Scale Space Generation

$\mathbf{v}_1^1$
$\delta_1 = 0.5$
$\sigma_1^1 = 1.0$

$\mathbf{v}_2^2$
$\delta_2 = 1.0$
$\sigma_2^2 = 2.5$

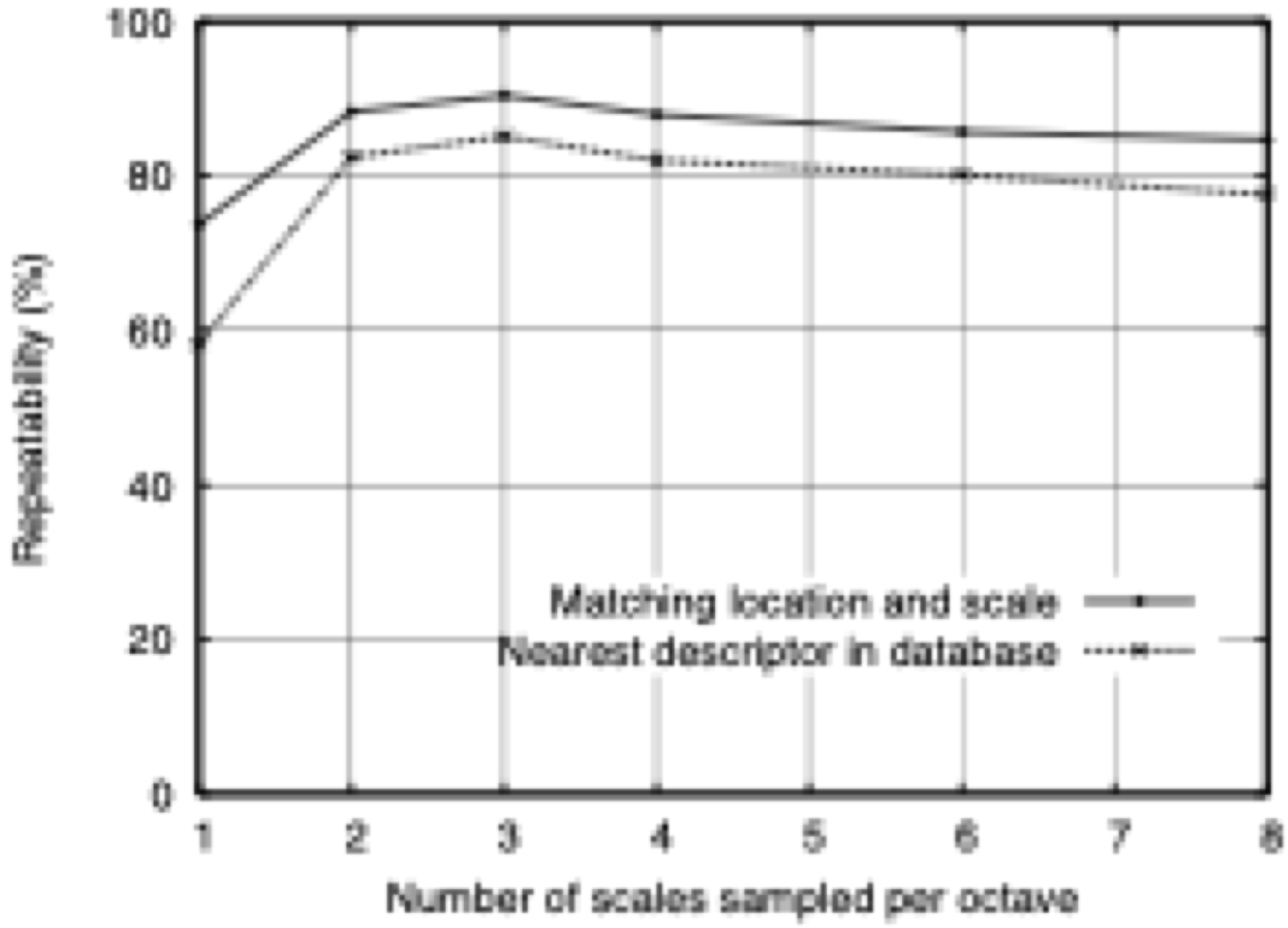
$\mathbf{v}_2^3$
$\delta_3 = 2.0$
$\sigma_2^3 = 5.1$

$\mathbf{v}_2^4$
$\delta_4 = 4.0$
$\sigma_2^4 = 10.2$

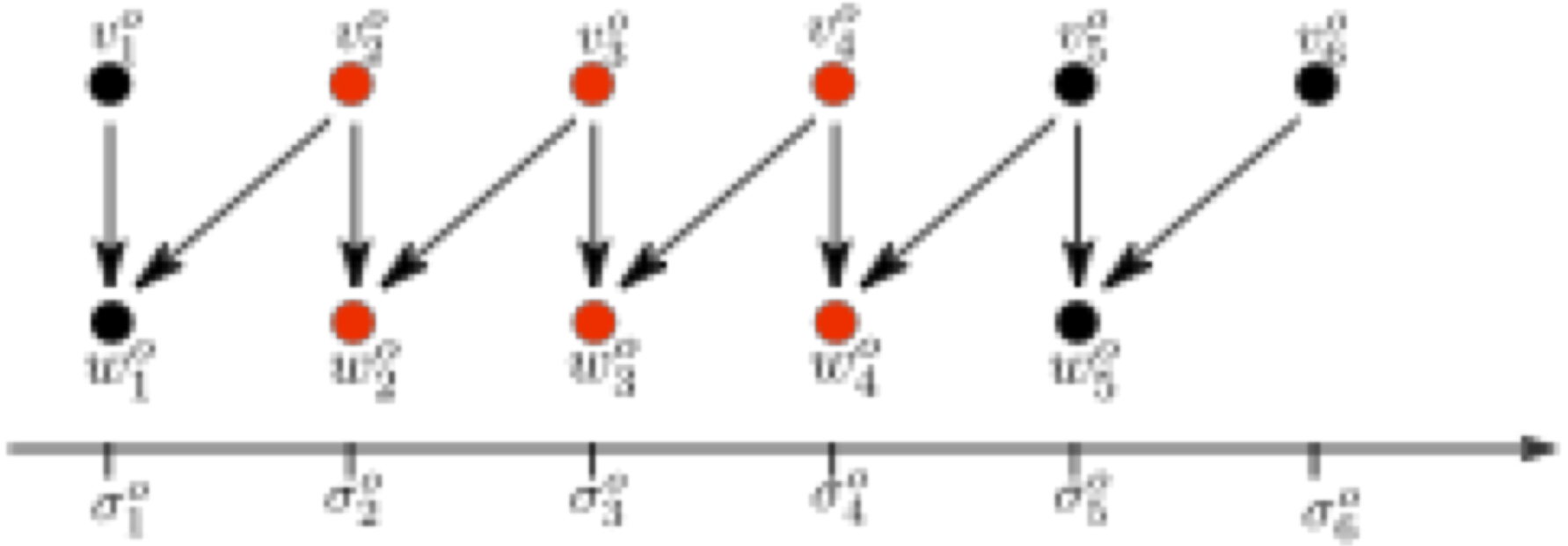
$\mathbf{v}_3^5$
$\delta_5 = 8.0$
$\sigma_3^5 = 25.6$

$\mathbf{v}_5^5$
$\delta_5 = 8.0$
$\sigma_5^5 = 40.6$

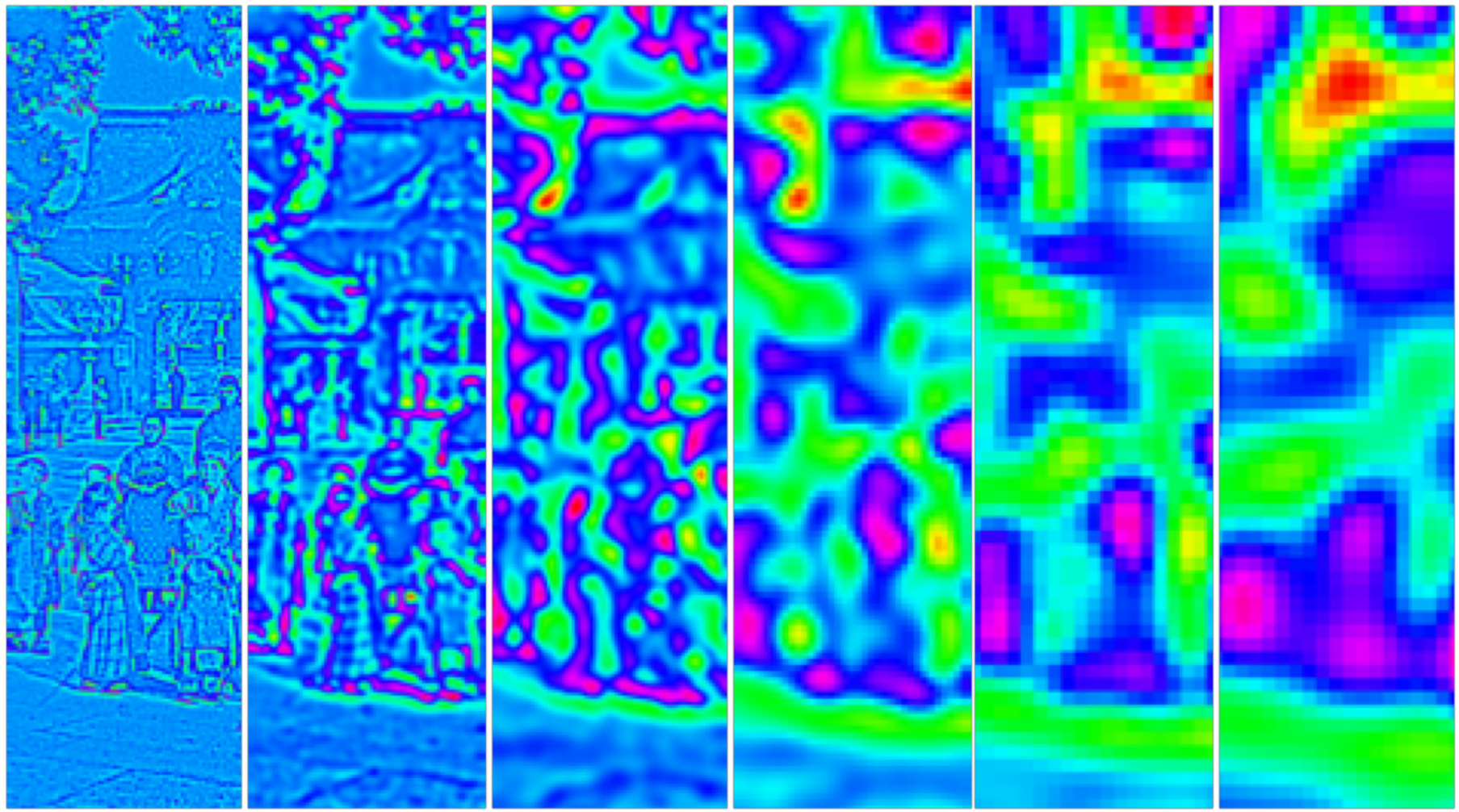
Scale Space Generation



Scale Space DoG



Scale Space DoG



$\mathbf{w}_1^1$
$\delta_1 = 0.5$
$\sigma_1^1 = 1.0$

$\mathbf{w}_2^2$
$\delta_2 = 1.0$
$\sigma_2^2 = 2.5$

$\mathbf{w}_2^3$
$\delta_3 = 2.0$
$\sigma_2^3 = 5.1$

$\mathbf{w}_2^4$
$\delta_4 = 4.0$
$\sigma_2^4 = 10.2$

$\mathbf{w}_1^5$
$\delta_5 = 8.0$
$\sigma_1^5 = 16.1$

$\mathbf{w}_3^5$
$\delta_5 = 8.0$
$\sigma_3^5 = 25.6$

Scale Space Extrema Extraction

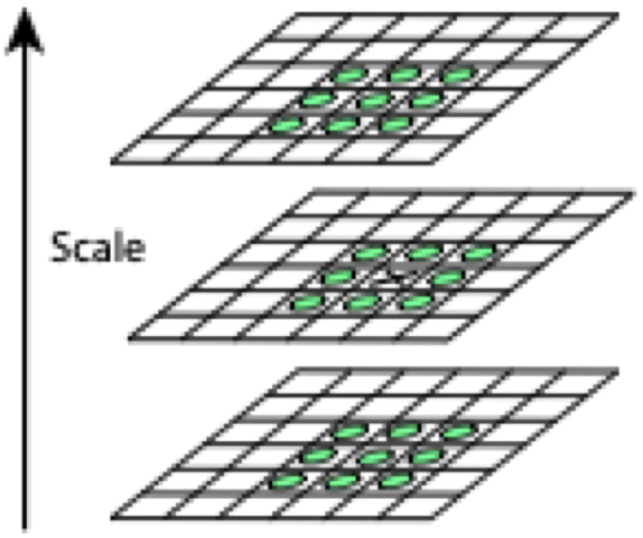


Figure 2: Maxima and minima of the difference-of-Gaussian images are detected by comparing a pixel (marked with X) to its 26 neighbors in 3x3 regions at the current and adjacent scales (marked with circles).

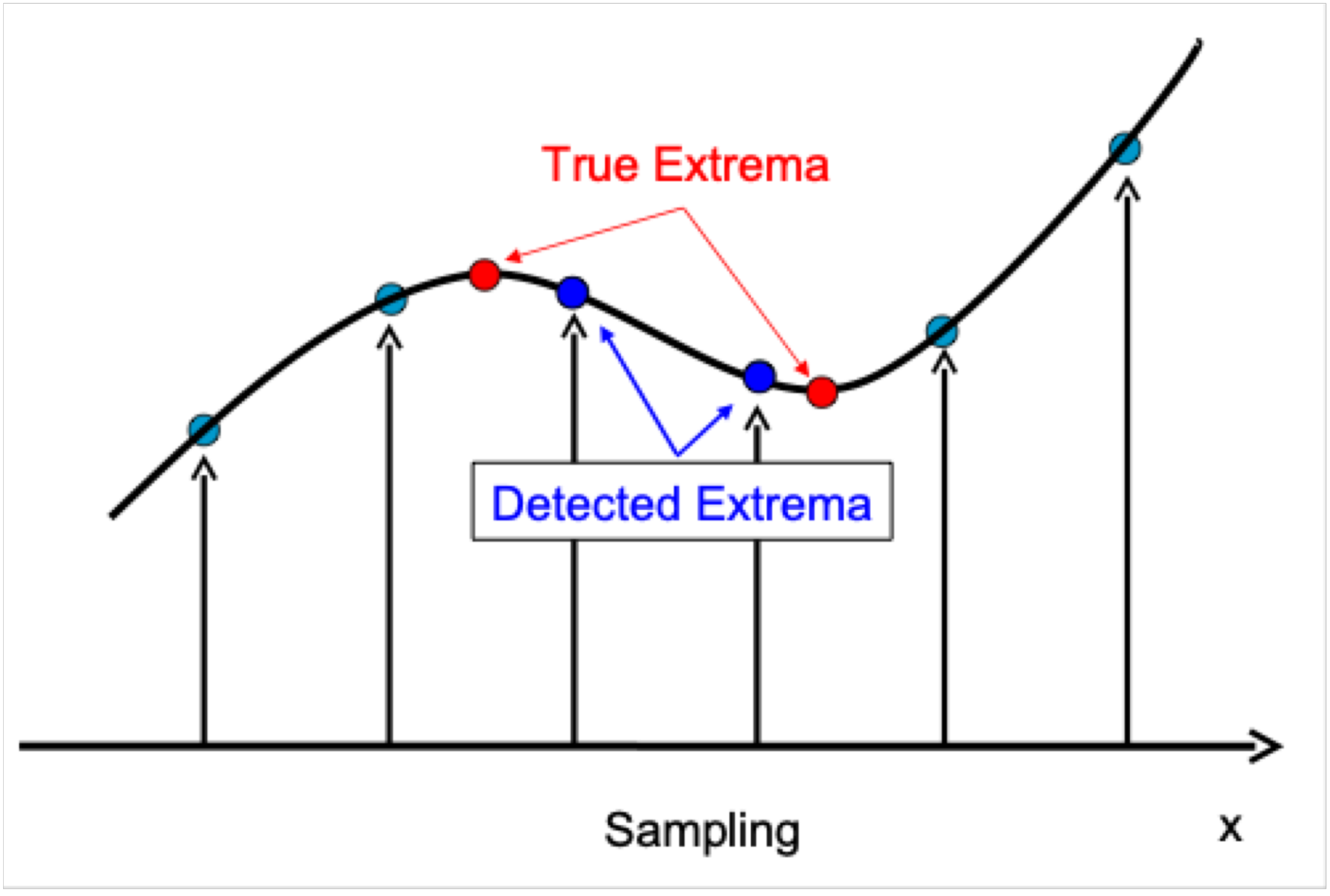
Scale Space Extrema Extraction

- Initial version only used the maxima in the 26 neighbours

- Further, it was refined by fitting a 3D quadratic function to the local sample points to locate the interpolated maximum location

- After computation of the scale space, extrema is located by detecting the point having local maxima in a 3x3x3 neighbourhood space

Scale Space Extrema Extraction



Scale Space Extrema Extraction

We denote by $\boldsymbol{\omega}^o_{s,m,n}(\boldsymbol{\alpha})$ the quadratic function at sample point (s,m,n) in the octave o, given by

$$\boldsymbol{\omega}^o_{s,m,n}(\boldsymbol{\alpha}) = \mathbf{w}^o_{s,m,n} + \boldsymbol{\alpha}^T \bar{g}^o_{s,m,n} + \frac{1}{2}\boldsymbol{\alpha}^T \bar{H}^o_{s,m,n}\boldsymbol{\alpha}, \tag{12}$$

$$\text{where } \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T \in [-1/2, 1/2]^3$$

$$\bar{g}^o_{s,m,n} = \begin{bmatrix} (\mathbf{w}^o_{s+1,m,n} - \mathbf{w}^o_{s-1,m,n})/2 \\ (\mathbf{w}^o_{s,m+1,n} - \mathbf{w}^o_{s,m-1,n})/2 \\ (\mathbf{w}^o_{s,m,n+1} - \mathbf{w}^o_{s,m,n-1})/2 \end{bmatrix}, \quad \bar{H}^o_{s,m,n} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{12} & h_{22} & h_{23} \\ h_{13} & h_{23} & h_{33} \end{bmatrix}$$

$$h_{11} = \mathbf{w}^o_{s+1,m,n} + \mathbf{w}^o_{s-1,m,n} - 2\mathbf{w}^o_{s,m,n}, \qquad h_{12} = (\mathbf{w}^o_{s+1,m+1,n} - \mathbf{w}^o_{s+1,m-1,n} - \mathbf{w}^o_{s-1,m+1,n} + \mathbf{w}^o_{s-1,m-1,n})/4,$$
$$h_{22} = \mathbf{w}^o_{s,m+1,n} + \mathbf{w}^o_{s,m-1,n} - 2\mathbf{w}^o_{s,m,n}, \qquad h_{13} = (\mathbf{w}^o_{s+1,m,n+1} - \mathbf{w}^o_{s+1,m,n-1} - \mathbf{w}^o_{s-1,m,n+1} + \mathbf{w}^o_{s-1,m,n-1})/4,$$
$$h_{33} = \mathbf{w}^o_{s,m,n+1} + \mathbf{w}^o_{s,m,n-1} - 2\mathbf{w}^o_{s,m,n}, \qquad h_{23} = (\mathbf{w}^o_{s,m+1,n+1} - \mathbf{w}^o_{s,m+1,n-1} - \mathbf{w}^o_{s,m-1,n+1} + \mathbf{w}^o_{s,m-1,n-1})/4.$$

$$\text{Compute } \boldsymbol{\alpha}^* = -\left(\bar{H}^o_{s,m,n}\right)^{-1} \bar{g}^o_{s,m,n}$$

Pruning Scale Space Extrema

- The scale space extrema obtained through interpolation are then subject to a series of checks

- In the first check, low contrast points. Those having value in the DoG space $|w| < 0.03$ are discarded as weak extrema or low contrast points

Pruning Scale Space Extrema

original image 233x189 pixel

832 initial key points

729 key points after pruning low contrast key points

Eliminating Edge Responses

- Further key points are refined by eliminating key points located on edges or those that are not distinctive enough

- This is similar to measure used for Harris Corner. Here the author uses 2x2 Hessian matrix

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

Eliminating Edge Responses

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

where alpha and beta are the large and small eigenvalues
Let r be the ratio between the eigen values

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

This ratio is a minimum only when the two eigenvalues are equal and increases otherwise. r > 10 is used to eliminate any other points

Eliminating Edge Responses

original image
233x189 pixel

832 initial
key points

729 key points after
pruning low
contrast key points

536 key points after
pruning based on
ratio of eigenvalues

Orientation Assignment

- The Gaussian smoothed image L with the appropriate scale is chosen
- At this scale the gradient magnitude and orientation are computed

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

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

- An orientation histogram is computed with 36 bins representing the 360 degree angles orientation around the key point

Orientation Assignment

- Each sample added to the histogram is weighted by its gradient magnitude and by a Gaussian window that is 1.5 times that of the scale of the key point

- Peaks in the histogram correspond to dominant directions of orientation.

- The highest peak and any peak within 80% of the highest peak are detected and key points for these peaks are formed (i.e. there could be multiple key points for the multiple peaks in orientation)

- Only 15% points have multiple orientations but they contribute significantly

- Finally a parabola is fit to the 3 histogram values closest to each peak to interpolate the peak position for better accuracy

Keypoint description

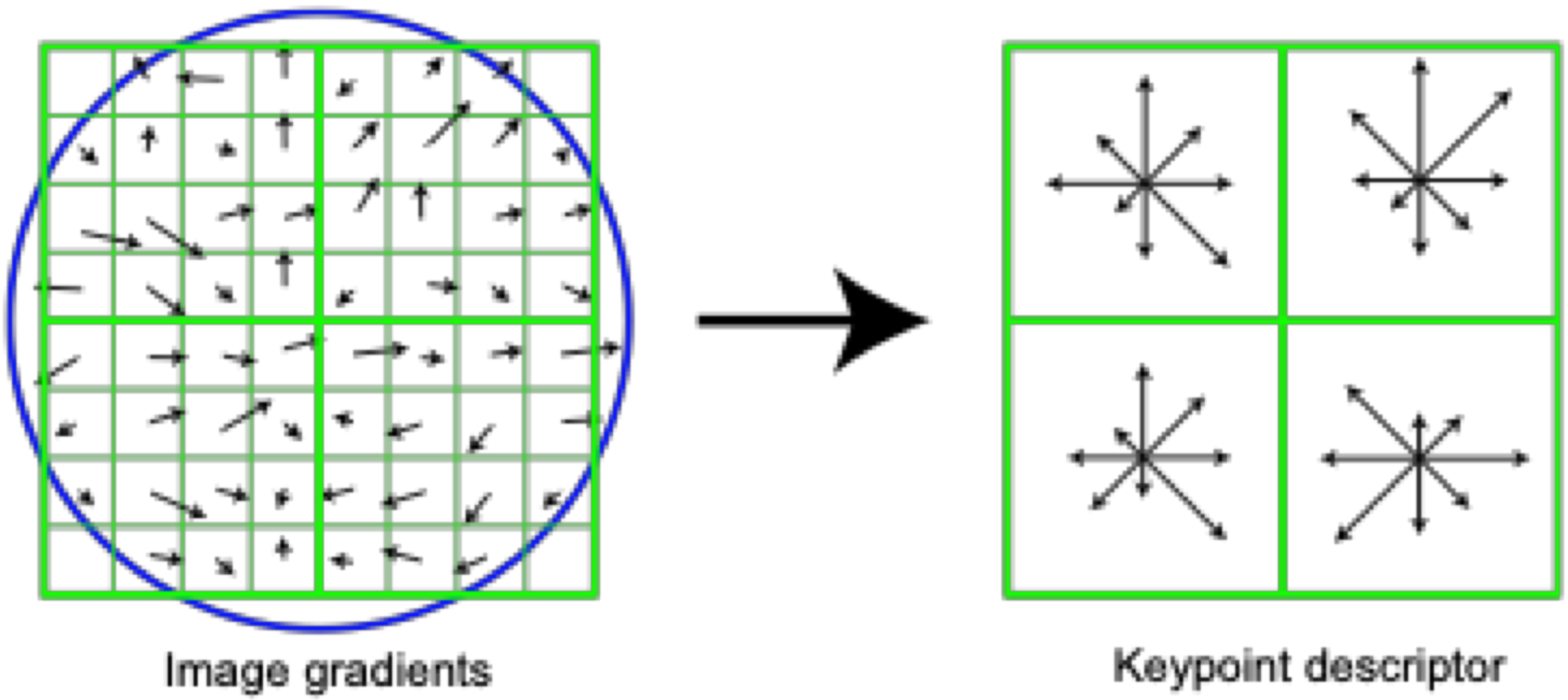
- The key point description is obtained by creating orientation histograms over 4x4 sample regions

- Each orientation histogram has 8 bins

- Each sample is weighted by a circularly symmetric Gaussian around the centre so that gradients far from the centre have less effect

- The resultant 4x4x8 length feature vector is obtained as the description of the key point

Keypoint description



Keypoint description

- The key point feature vector is normalised to unit length
- A change in image contrast will therefore not effect the feature vector due to normalisation
- Brighness change also does not affect the descriptor as the gradients are obtained by pixel differences
- For illumination changes due to 3D orientation, the descriptors with large gradient magnitude get effected. To reduce this, gradient magnitude greater than 0.2 are thresholded and the feature vector renormalised

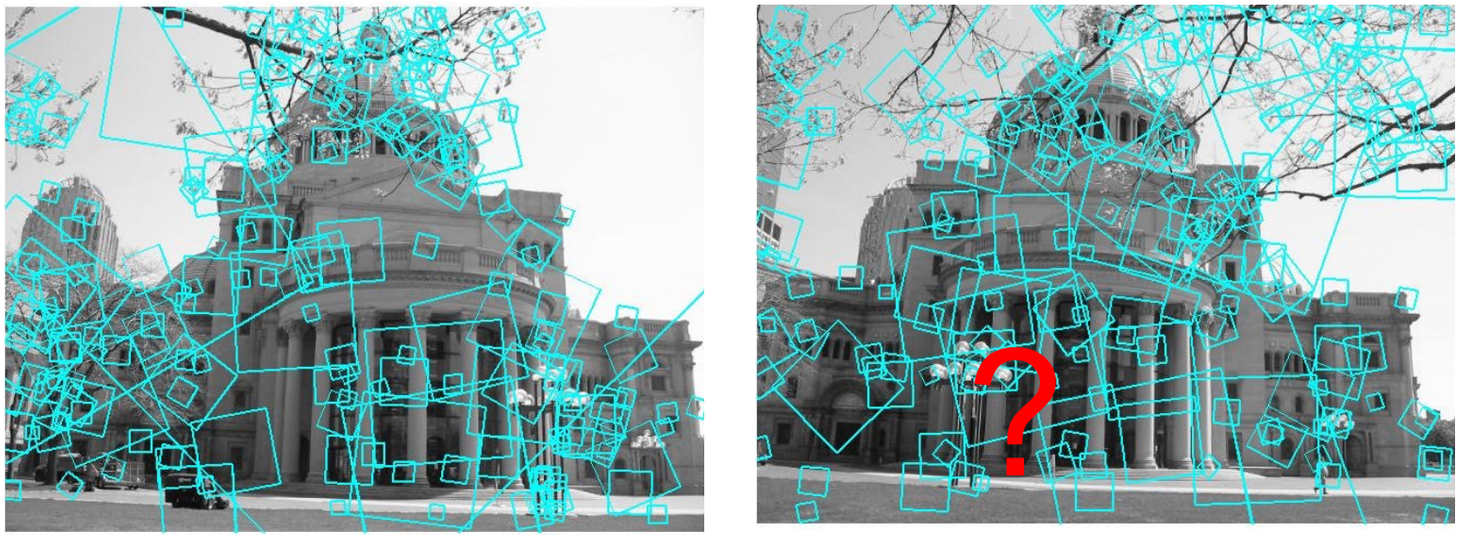
Feature Matching

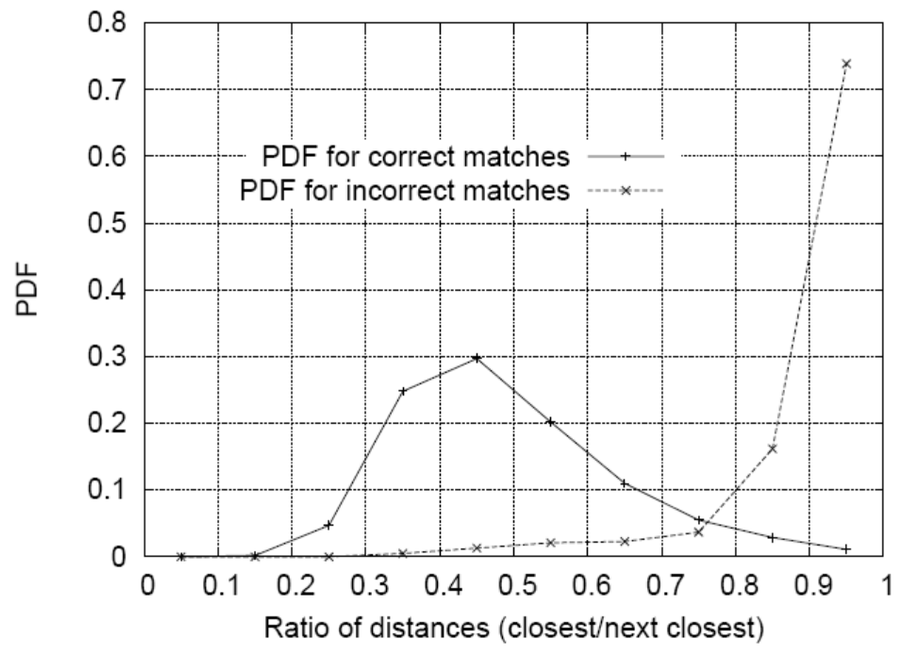
- Feature matching is obtained by using minimum Euclidean distance

- $\|x_a - x_b\|_2$

Feature Matching

- Feature matching is obtained by using minimum Euclidean distance
- $\|x_a - x_b\|_2$
- For robustness ratio of distances between best match and second best match is considered





SIFT

- Scale Space Extrema Detection
- Keypoint Localization
- Orientation Assignment
- Keypoint Descriptor

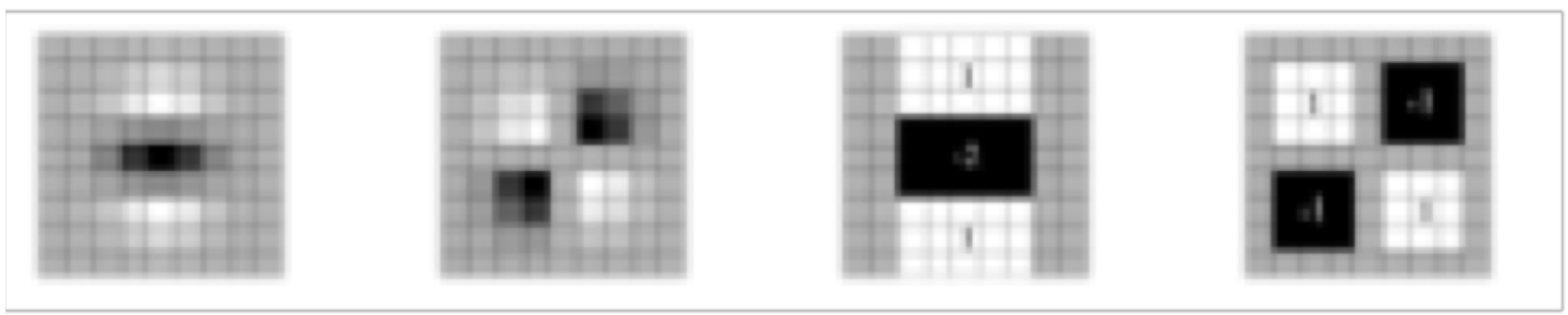
Very robust

– 80% Repeatability at:

- 10% image noise
- 45° viewing angle
- 1k-100k keypoints in database

Other feature: SURF

- Uses Box filters instead of Gaussian for faster approximation
- Scale space constructed by difference of box filters
- Hessian matrix based interest points



Other feature: SURF

- Determinant of Hessian matrix used to determine the appropriate scale
- Faster computation using the idea of Integral images or summed area table

1	2	2	4	1
3	4	1	5	2
2	3	3	2	4
4	1	5	4	6
6	3	2	1	3

input image

0	0	0	0	0	0
0	1	3	5	9	10
0	4	10	13	22	25
0	6	15	21	32	39
0	10	20	31	46	59
0	16	29	42	58	74

integral image

Other feature: MSER

- Maximally stable extremal regions considers stable regions as obtained by considering different thresholds for a watershed segmentation routine.

- Those segments that do not change for a wide range of thresholds are considered candidate segments

- They are extremal in the sense that the internal intensities are either higher or lower than the boundary

> **Extremal Region** $\mathcal{Q} \subset D$ **is a region such that for all** $p \in \mathcal{Q}, q \in \partial\mathcal{Q} : I(p) > I(q)$ **(maximum intensity region) or** $I(p) < I(q)$ **(minimum intensity region).**

- Initially showed application by Matas et al for wide baseline matching of patches

Other feature: MSER

Object Categorisation

CS 783: Visual Recognition

Topics covered so far

- Instance Recognition

- Local features - visual words, SIFT

Object Categorisation

What is Object Categorisation?

- Ability to label objects
- In category theory, a branch of mathematics, an initial object of a category C is an object I in C such that for every object X in C, there exists precisely one morphism I → X

was ist das?

এটা কি?

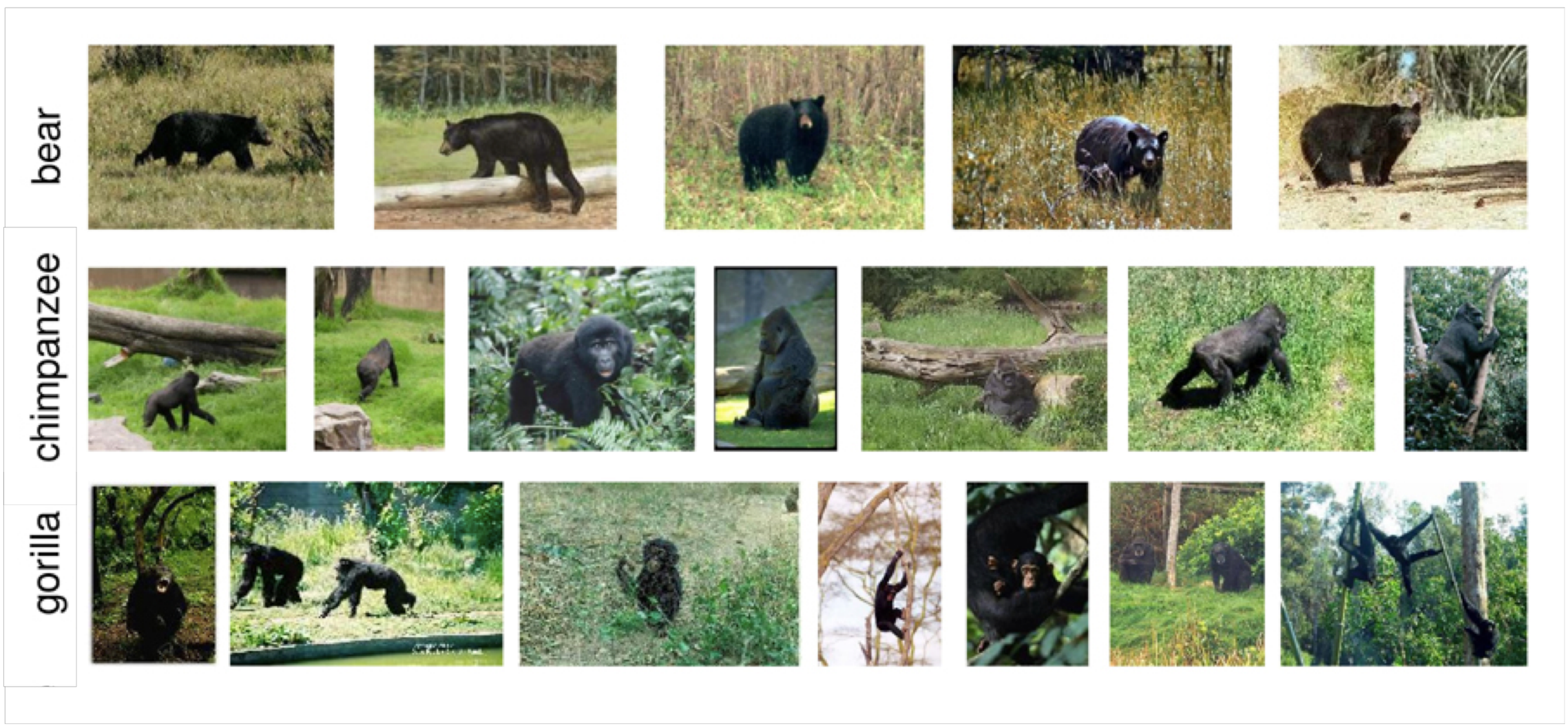
वो क्या है?

ഇത് എന്താണ്?

wat is dat?

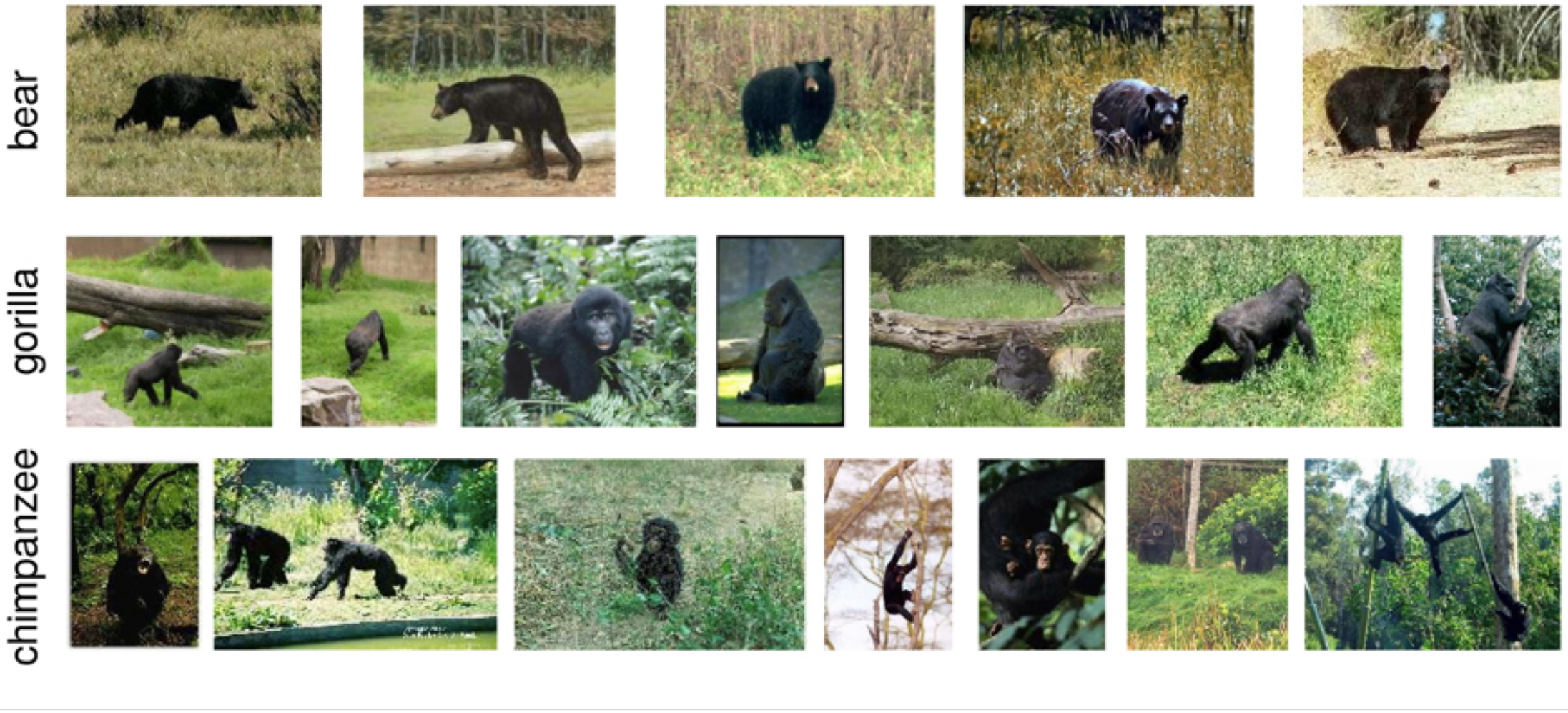
Qu'est-ce que c'est?

Object Categorisation



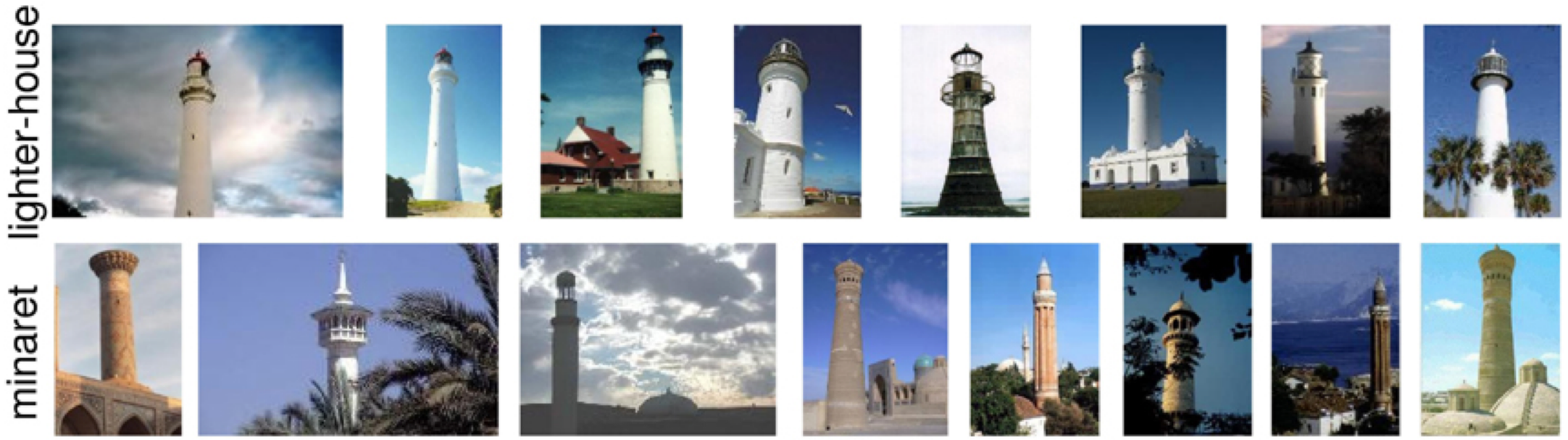
http://www.robots.ox.ac.uk/~vgg/research/caltech/C256_ambiguities.html

Object Categorisation



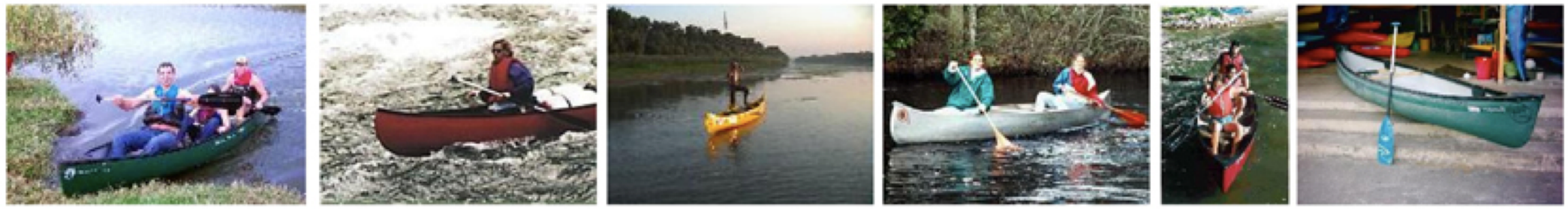
http://www.robots.ox.ac.uk/~vgg/research/caltech/C256_ambiguities.html

Object Categorisation



Object Categorisation

canoe



kayac

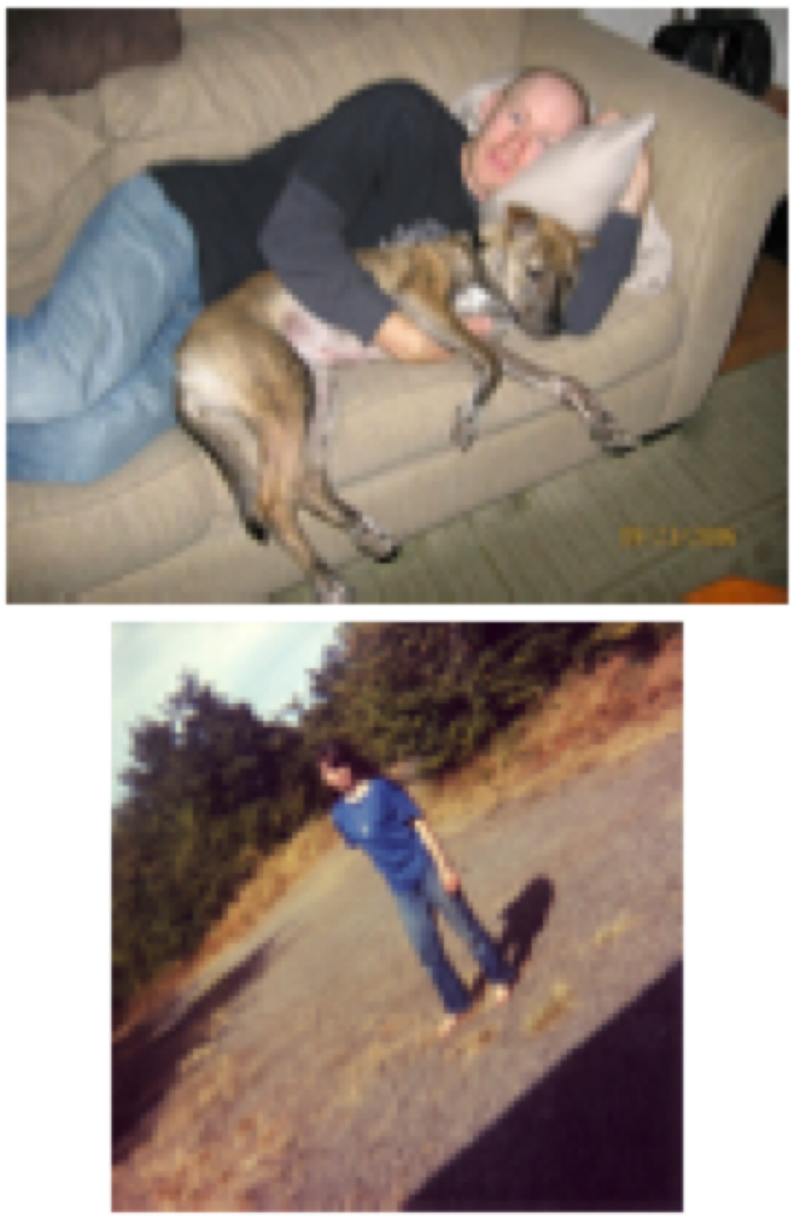
Object Categorisation

Challenges

Illumination

image credit: Hakan Bilen

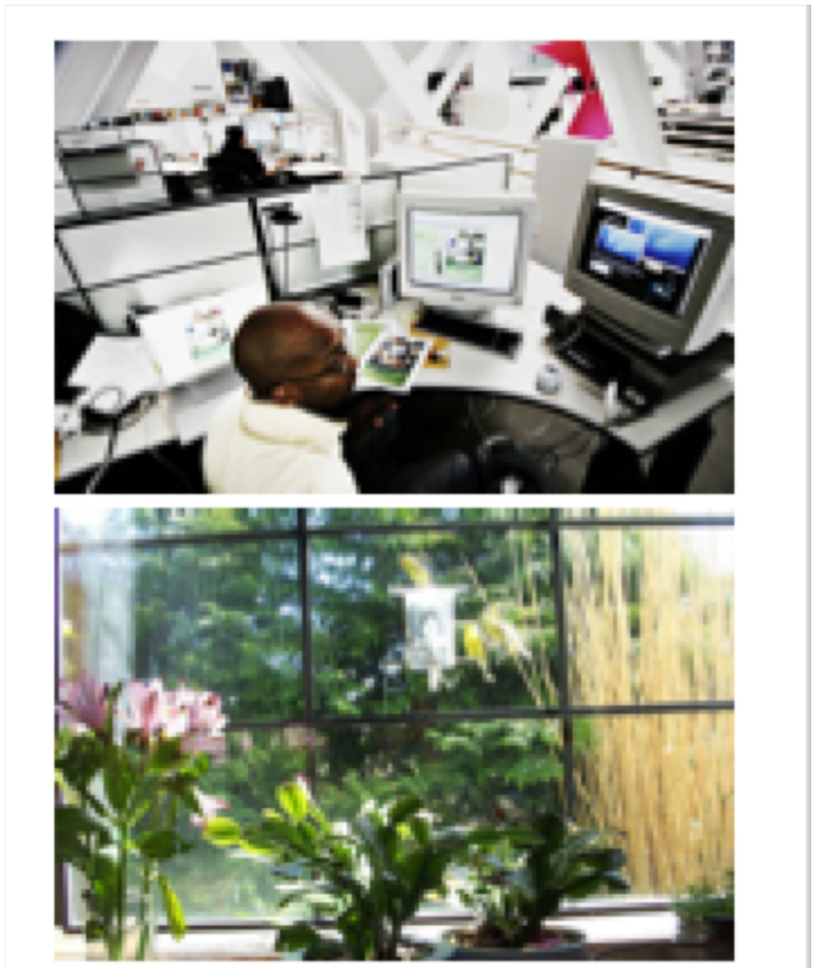
Challenges



Viewpoint

image credit: Hakan Bilen

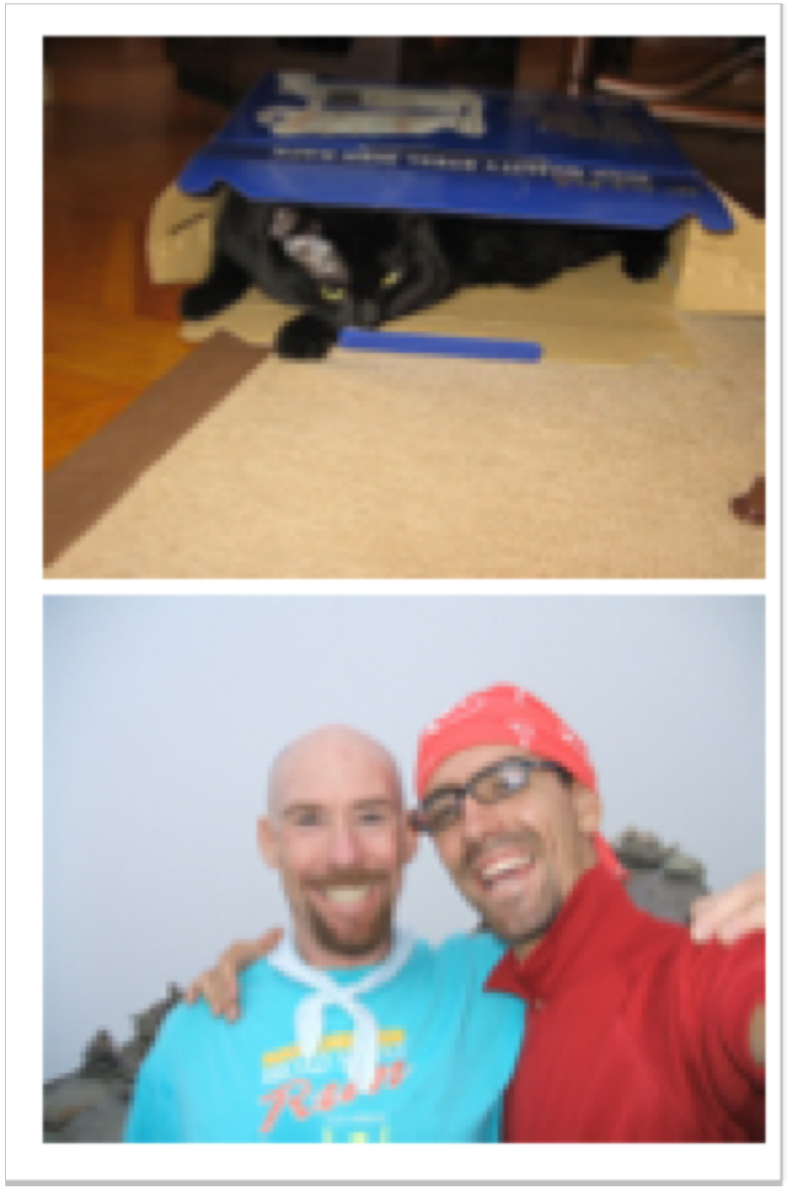
Challenges



Background clutter

image credit: Hakan Bilen

Challenges



Occlusion

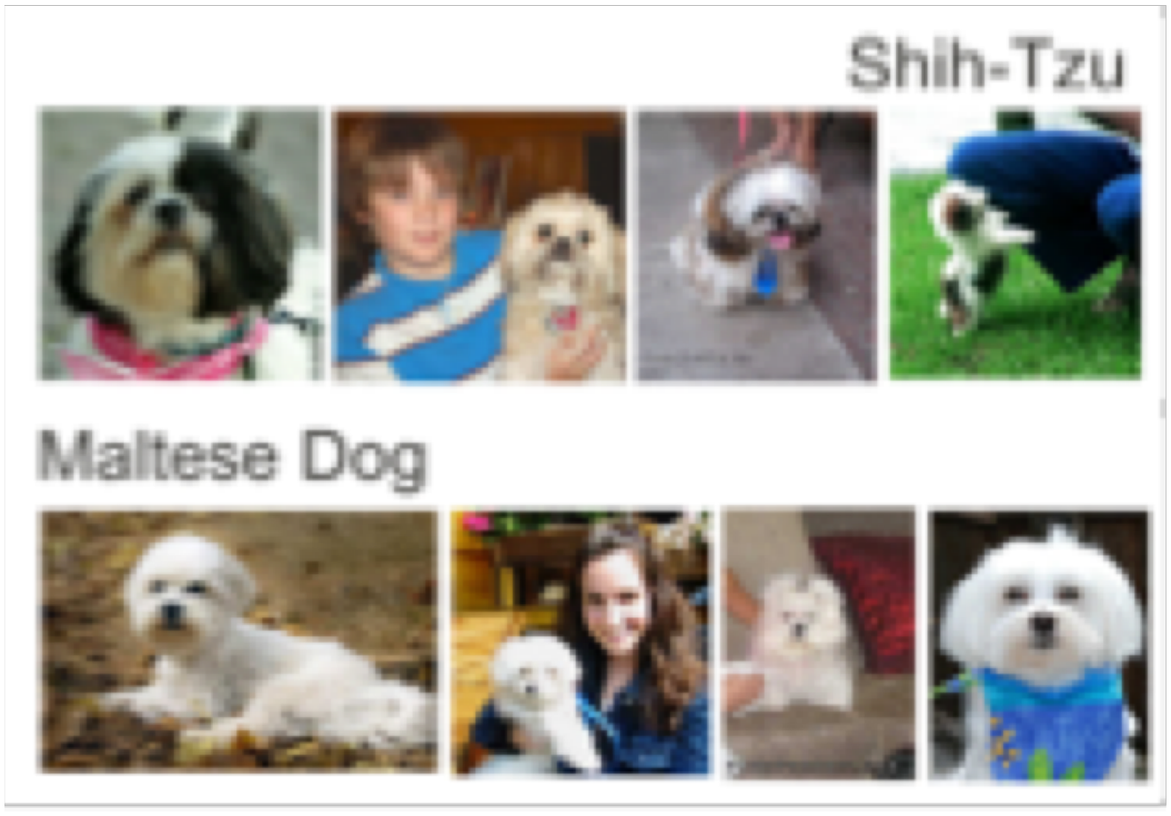
image credit: Hakan Bilen

Challenges

Intra-class variation

image credit: Hakan Bilen

Challenges



Intra-class vs inter-class variability

image credit: Hakan Bilen

Approach

Problem

- Given: Set of positive training images that contain images from a particular object class
- And a set of negative images that do not contain the particular object class
- Predict given a test image whether the image contains the class or not

Approach

- Represent each image as a bag of visual words to obtain a feature
- Train a classifier to classify the images
- Use the trained model to predict the test image