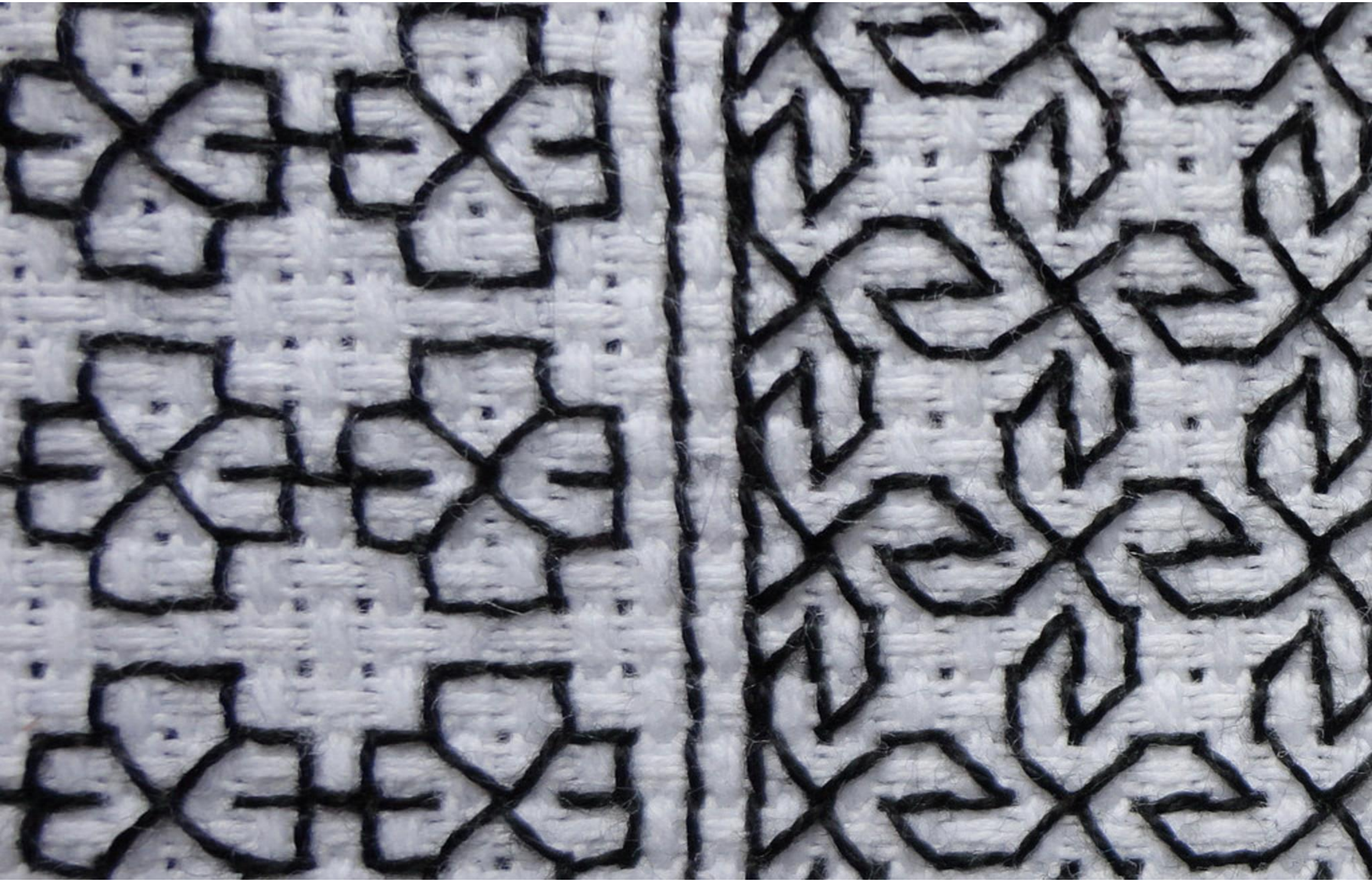


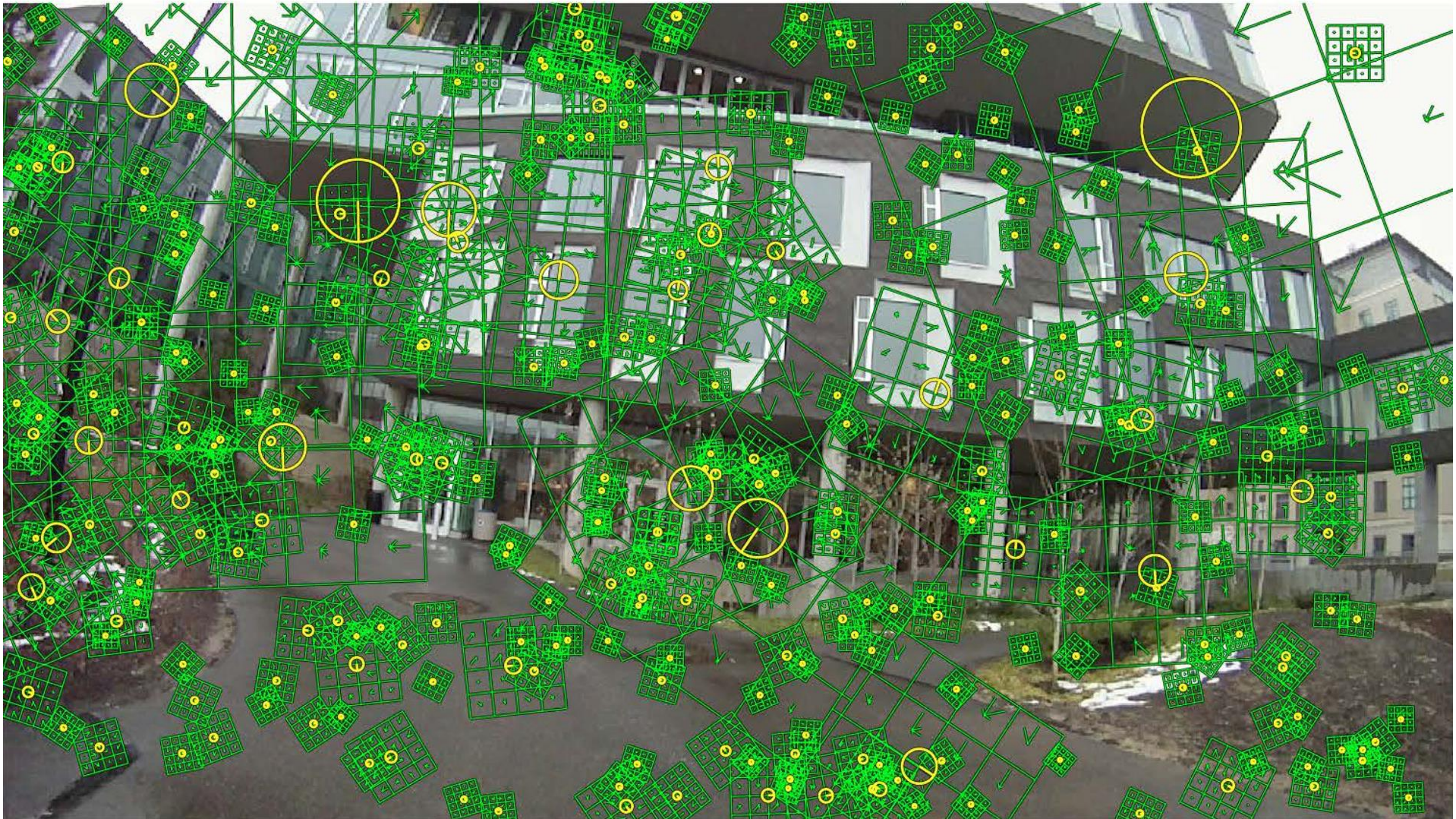
Miriam Blaylock@flickr

Feature Detection

SIFT (Scale Invariant Feature Transform)

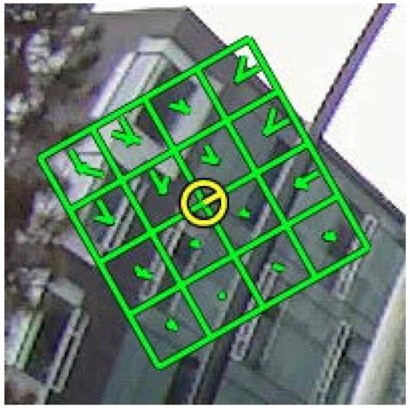
Feature Detection

SIFT (Scale Invariant Feature Transform)



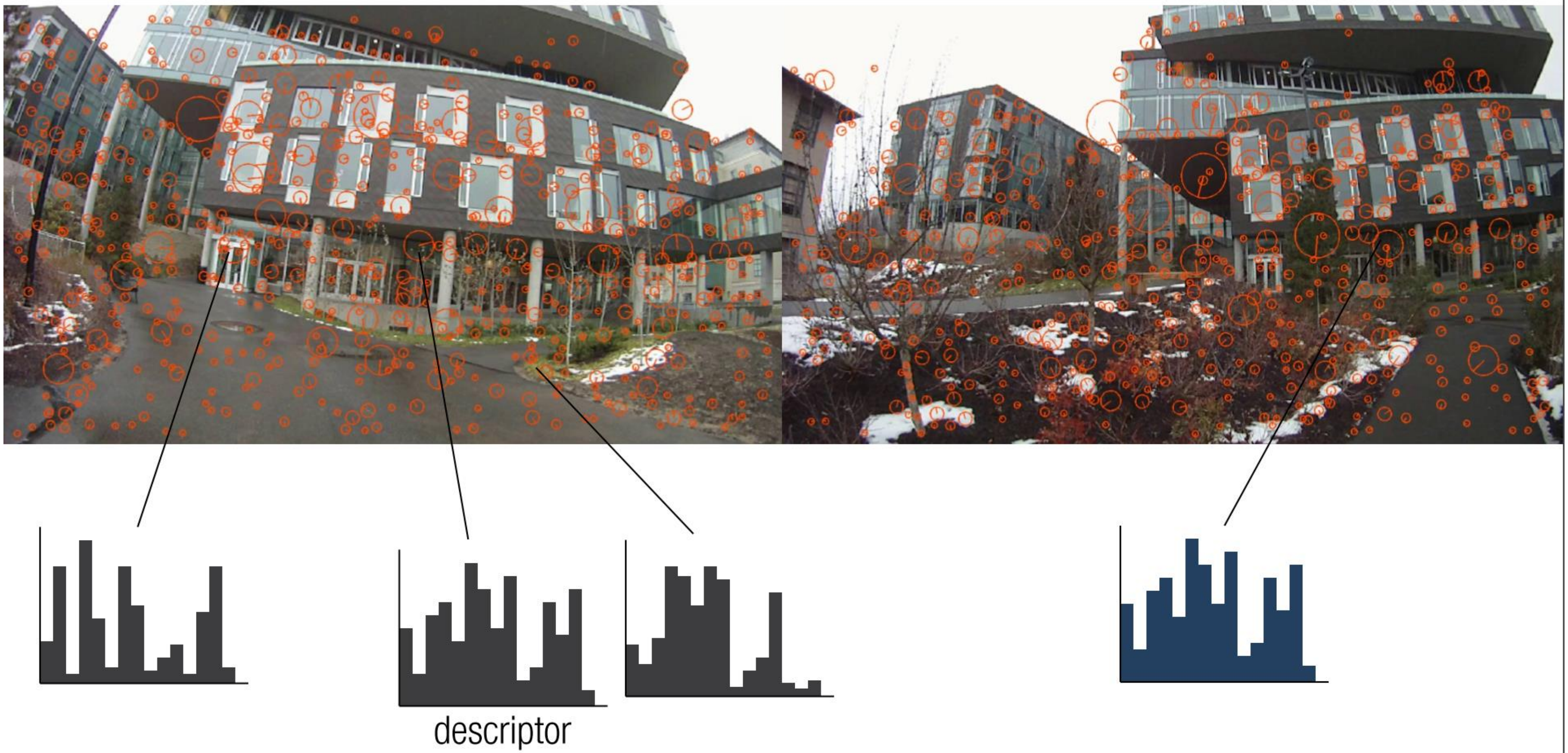
Feature Detection

SIFT (Scale Invariant Feature Transform)



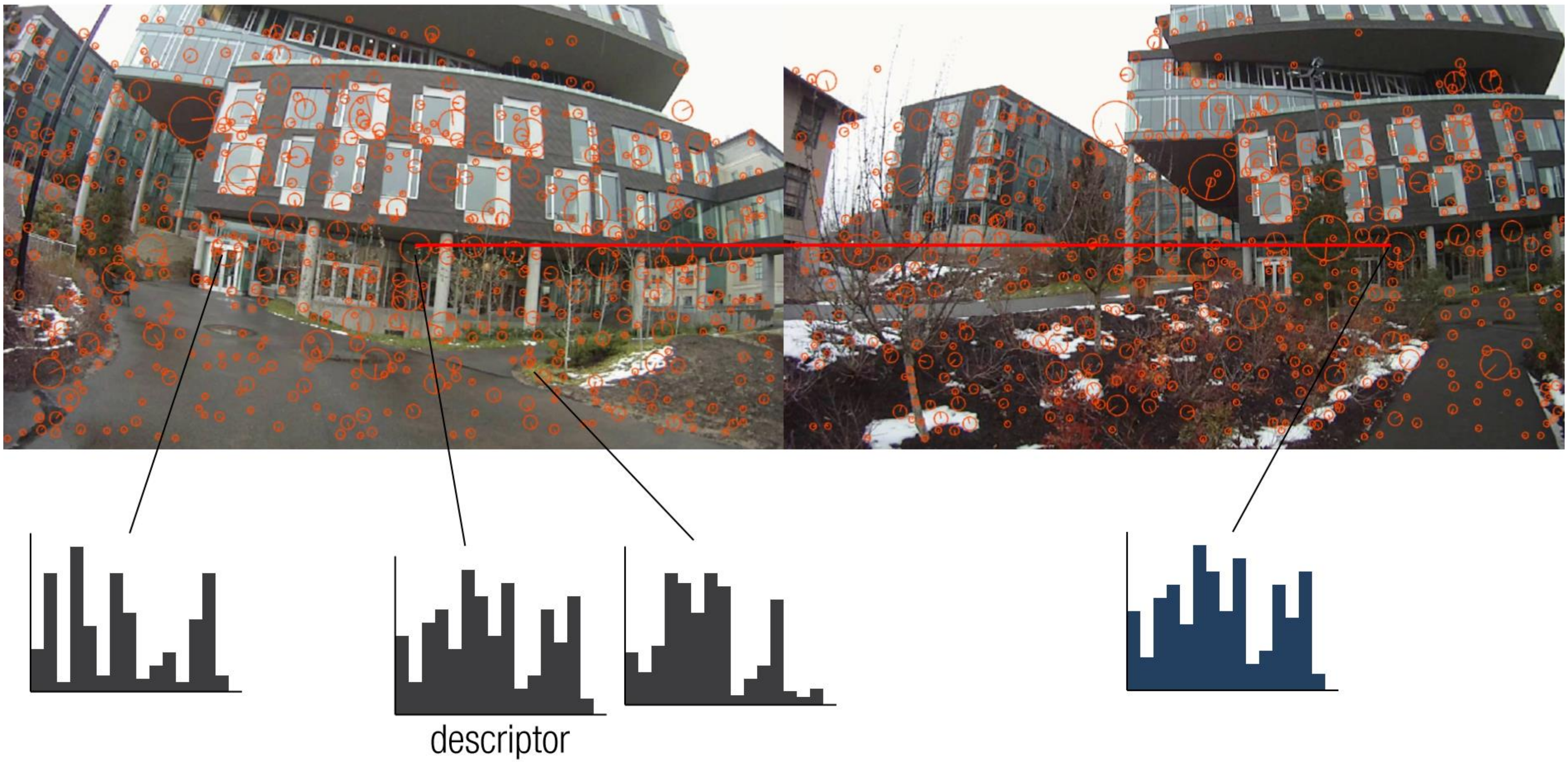
Feature Matching

Feature Descriptor

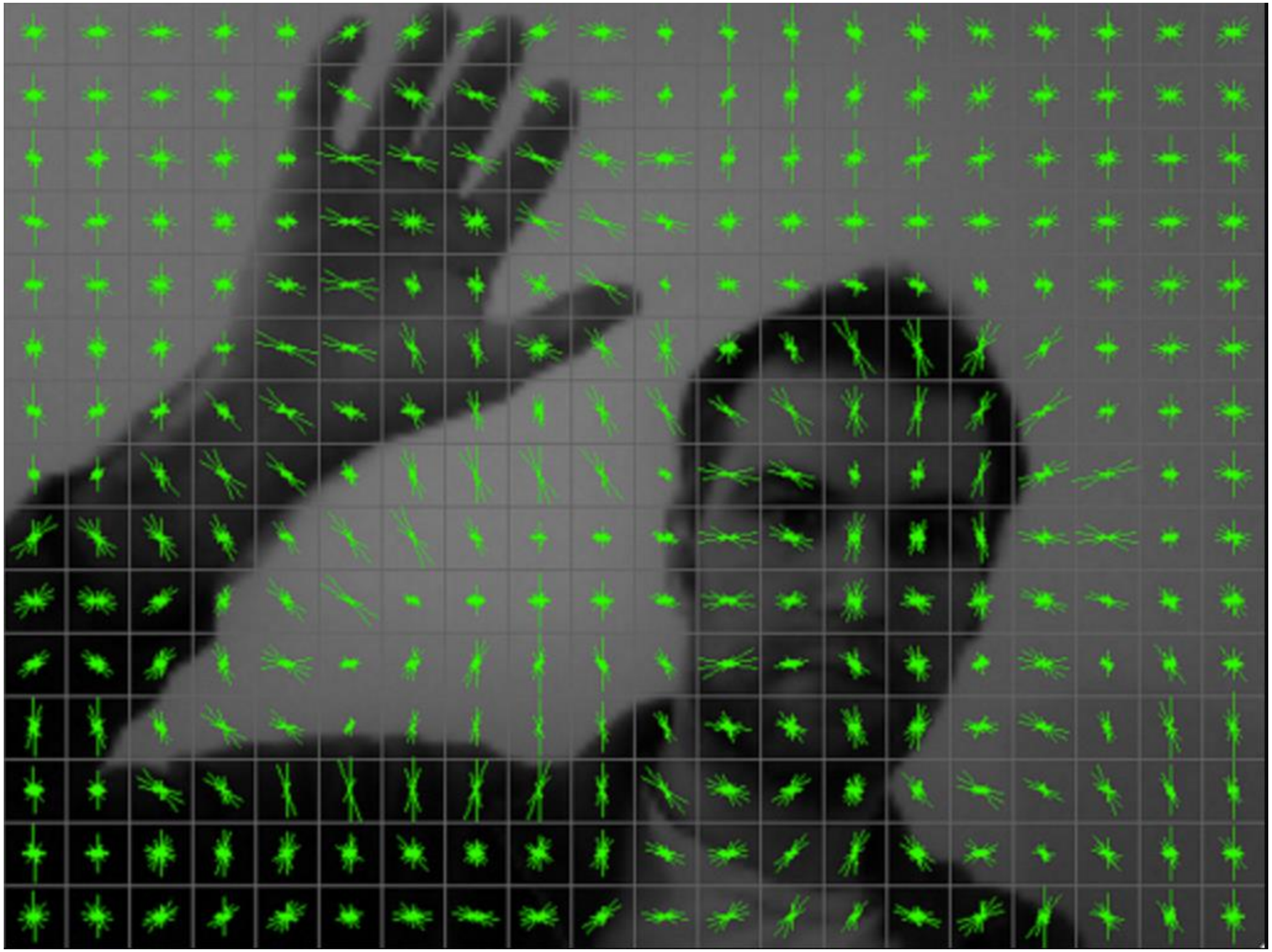


Feature Matching

Nearest Neighbor Search

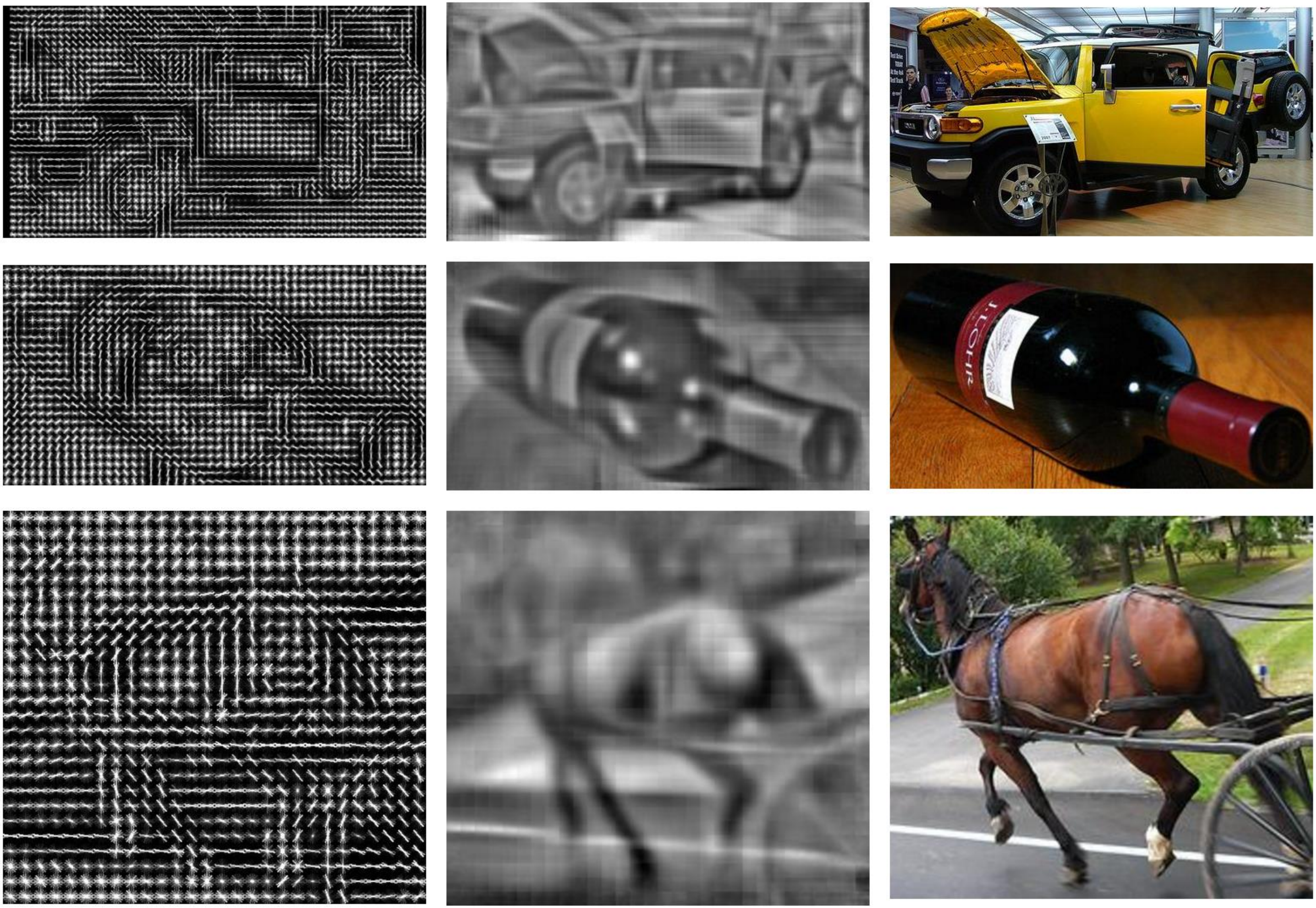


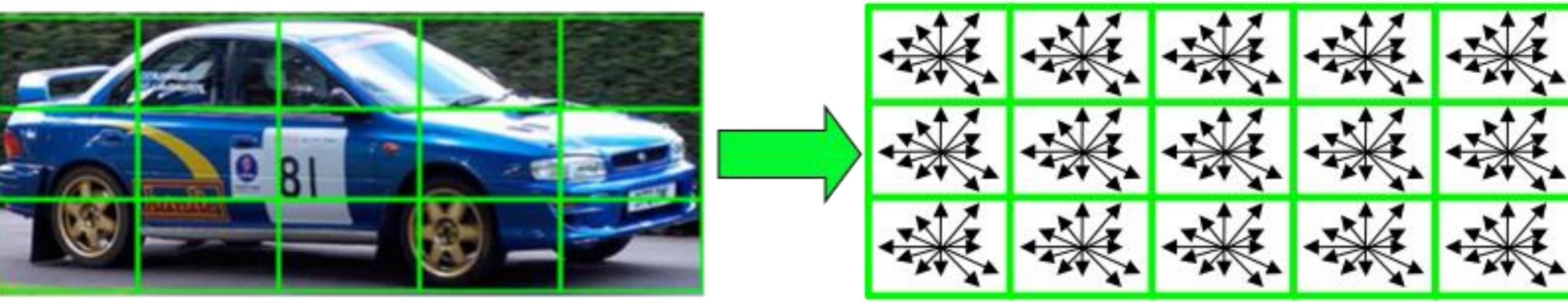
Hog



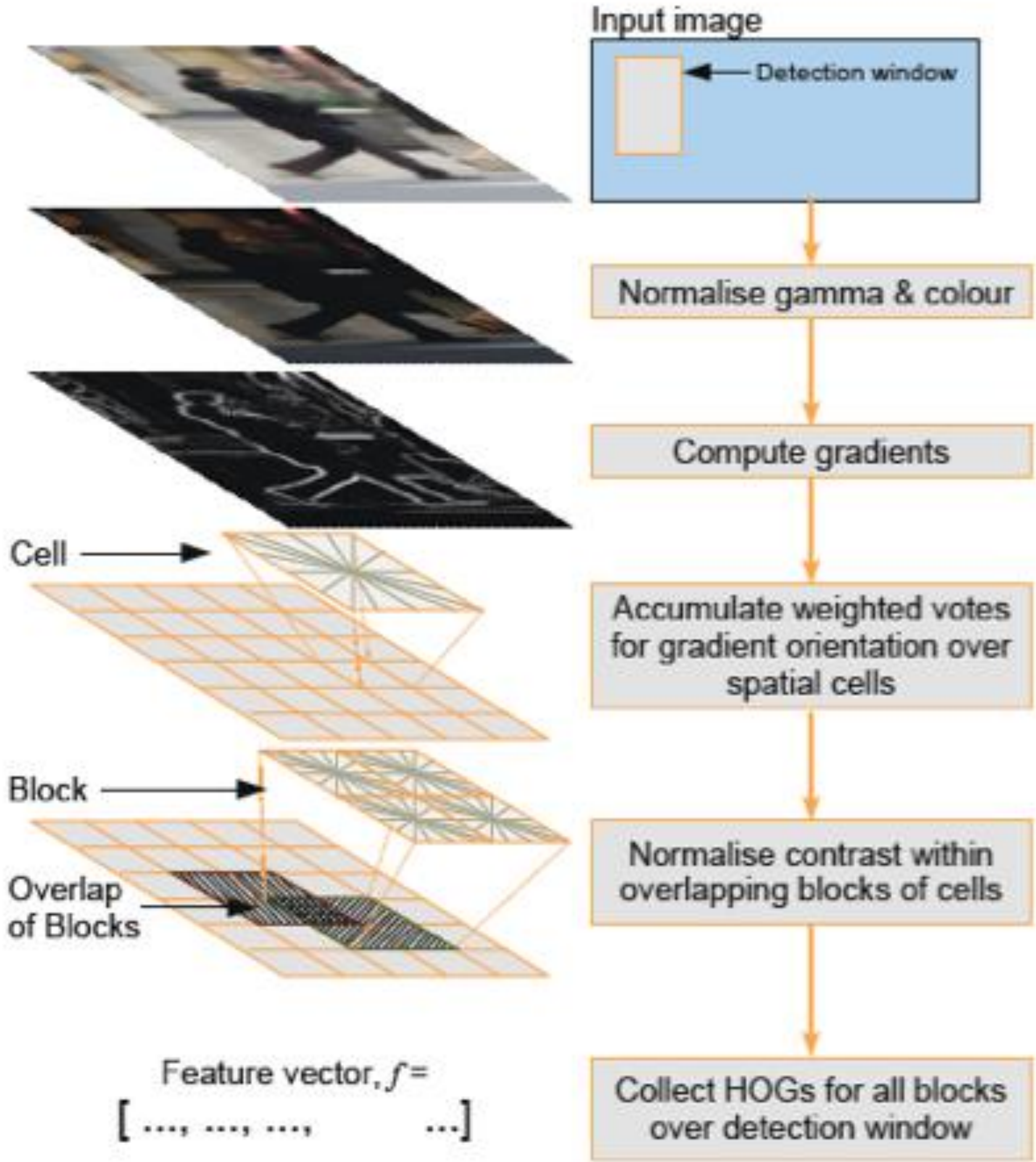
Greg Borenstein@flickr

HOG [1] | Inverse (Us) | Original




81

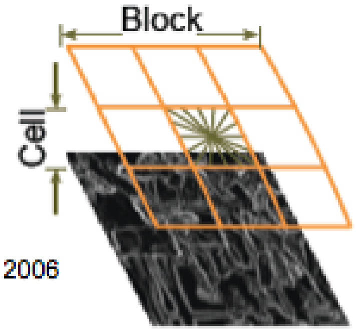
HOG Features



- Create cell histograms
  - Each pixel in cell casts weighted vote based on gradient magnitude centered there
    - Weighted by applying a Gaussian spatial window to each pixel before accumulating orientation votes into cells → (σ = .5*block width)
  - Votes are accumulated in 9 Histogram channels (orientation bins) spread evenly over 0-180 degrees (Or 0-360 degrees if signed values desired)

"Human Detection PHD Thesis" Navneet Dalal 2006

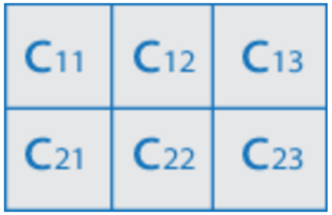
- Descriptor Blocks
  - To account for illumination/ contrast changes the cells must be grouped into "blocks" and normalized
  - HOG descriptor is a vector of components of normalized cell histograms from all the block regions
  - Author's optimum R-HOG (10% miss rate)
    - 3 parameters
      - 3x3 cell blocks
      - 6x6 pixel cells
      - 9 histogram channels (orientation bins)



(a) R-HOG/SIFT

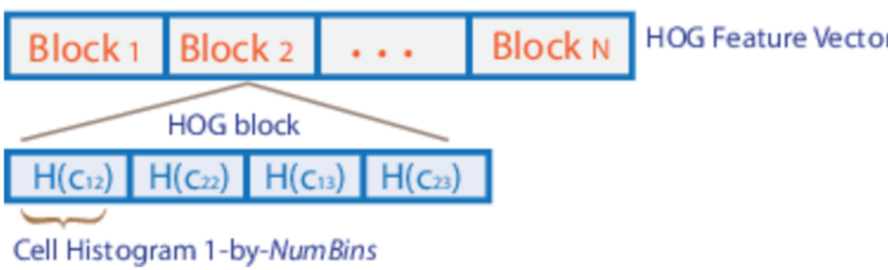
Arrangement of Histograms in HOG Feature Vectors

The figure below shows an image with six cells.

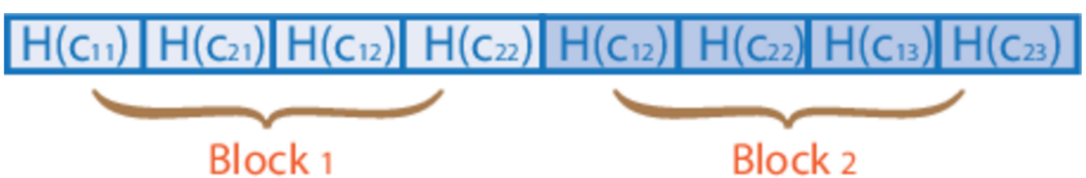


If you set the `BlockSize` to `[2 2]`, it would make the size of each HOG block, 2-by-2 cells. The size of the cells are in pixels.

The HOG feature vector is arranged by HOG blocks. The cell histogram, $H(C_{yx})$, is 1-by-`NumBins`.

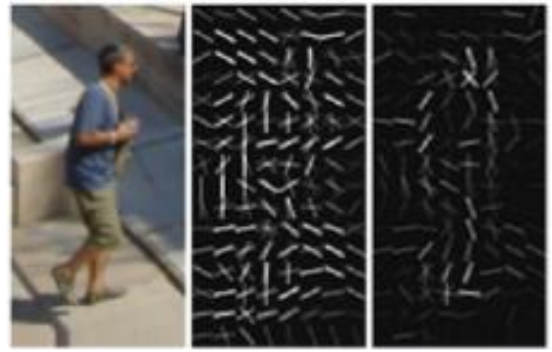


The figure below shows the HOG feature vector with a 1-by-1 cell overlap between blocks.

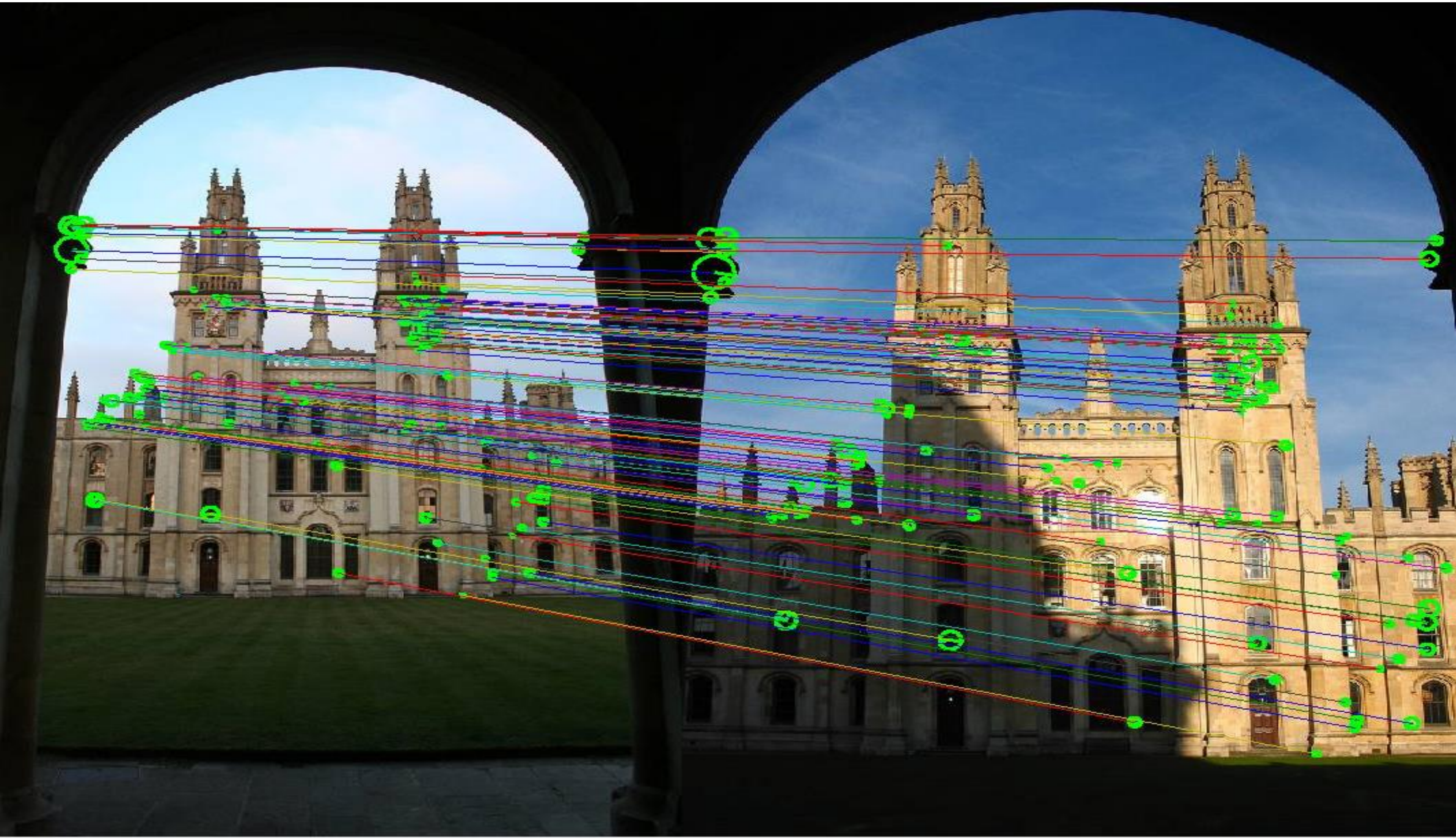


- Normalize the Blocks
  - V is vector containing non-normalized histogram data and e is a small constant (Not very important over the larger ranges – 1e^-3 to 5e^-2)
- Typical Detector Window
  - Authors used 64x128 detection window
  - 16 pixels of margin around person on all four sides
  - Decreasing window size or person size in image decreases performance

Importance weighted responses



SIFT



SIFT vector formation

- Thresholded image gradients are sampled over 16x16 array of locations in scale space
- Create array of orientation histograms
- 8 orientations x 4x4 histogram array = 128 dimensions

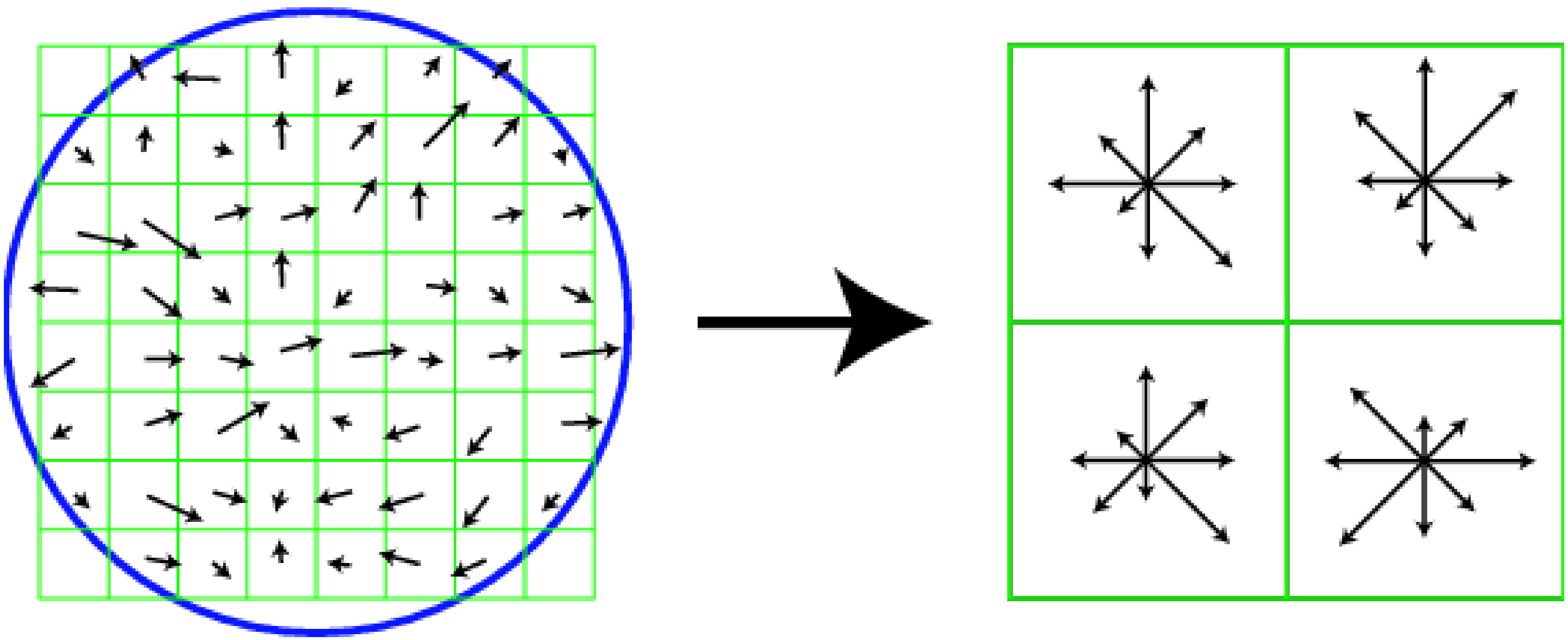


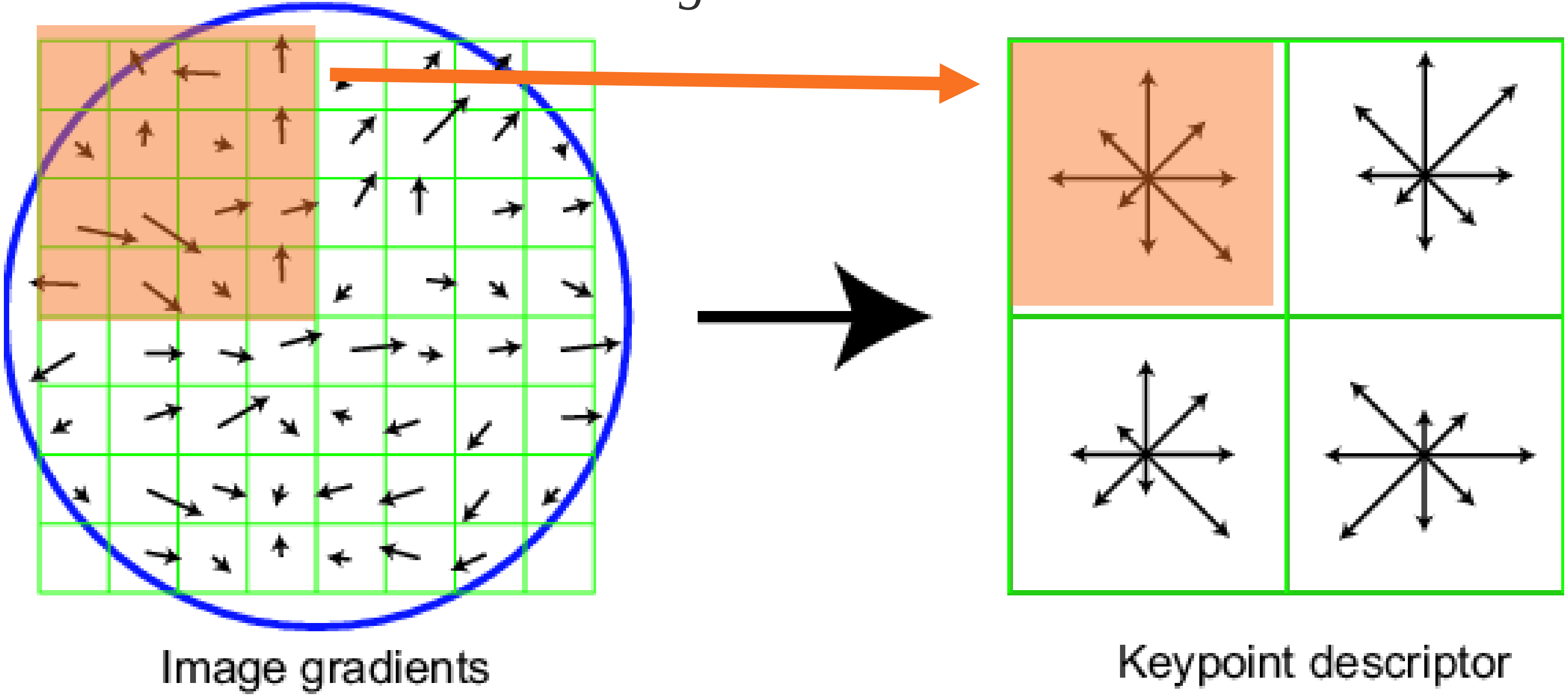
Image gradients Keypoint descriptor

SIFT vector formation

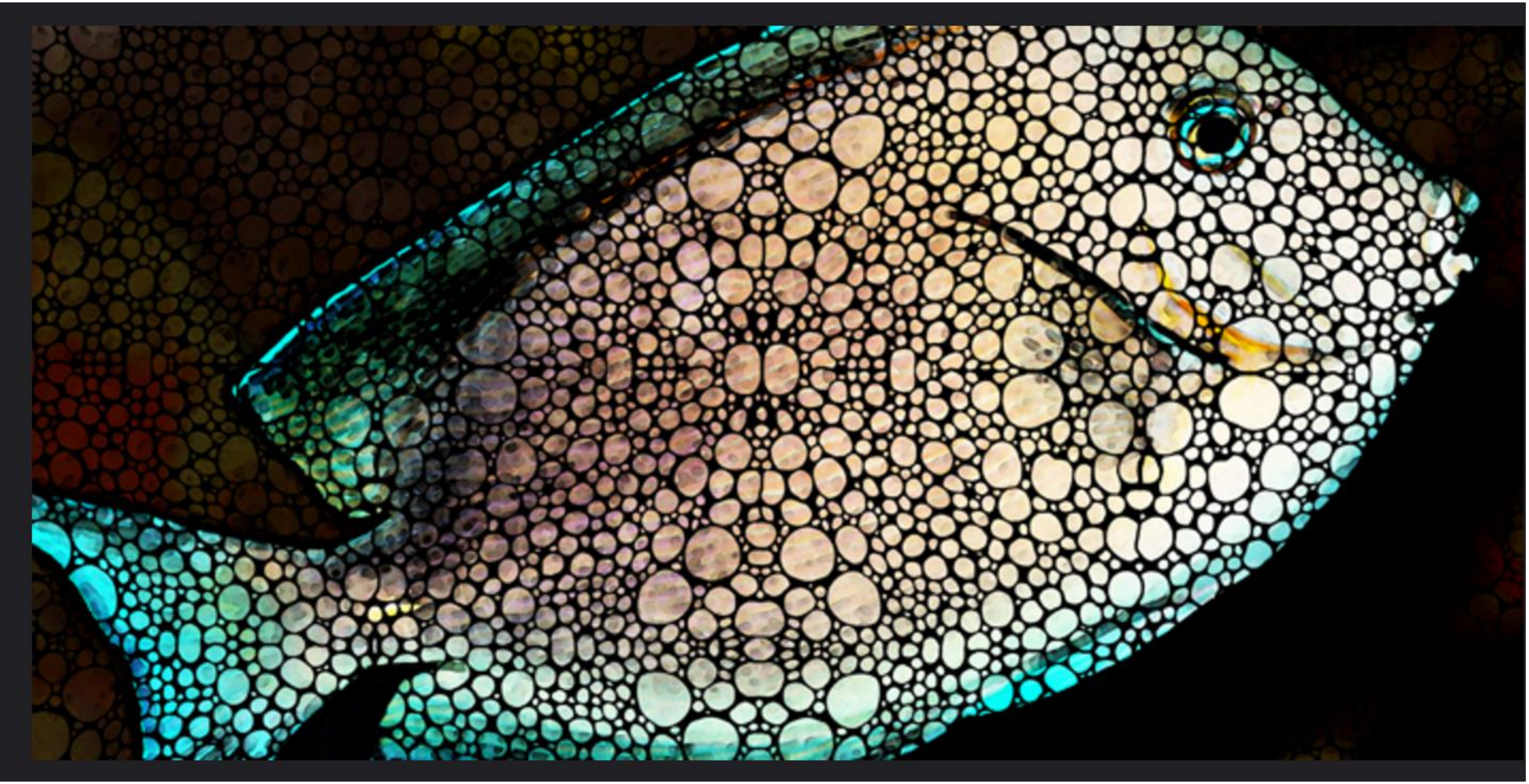
- Orientation is defined relative to the orientation of the detected Sift feature

Local orientation

histogram

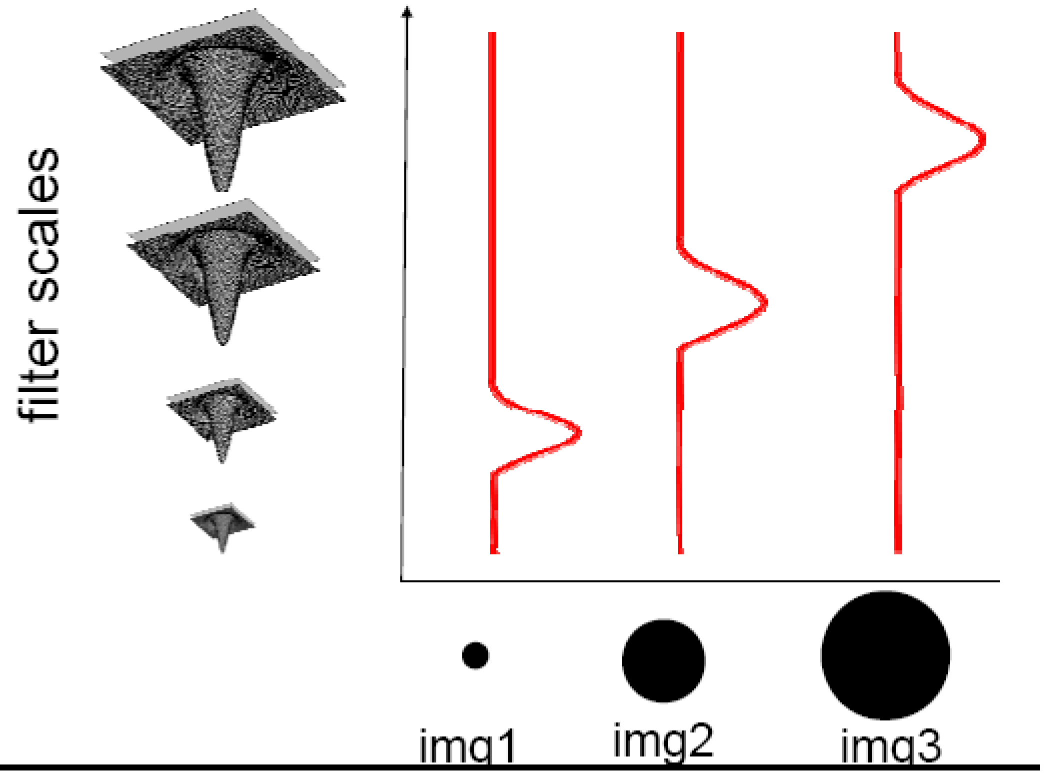


SIFT Fish



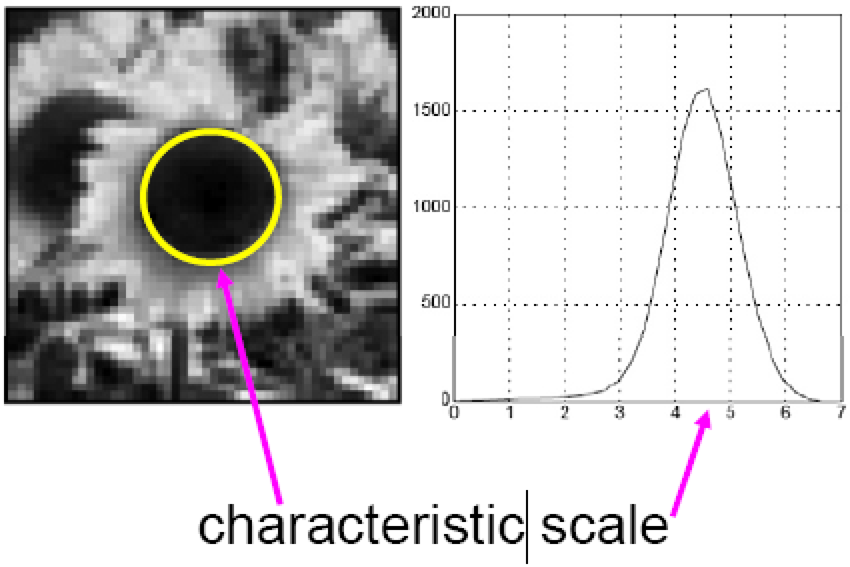
Sharon Cummings @ flickr

Laplacian-of-Gaussian = "***blob***" detector

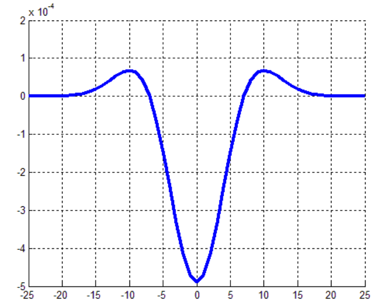


At a given point in the image:

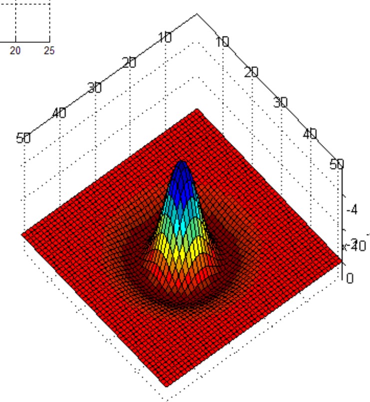
- We define the *characteristic scale* as the scale that produces peak of Laplacian response

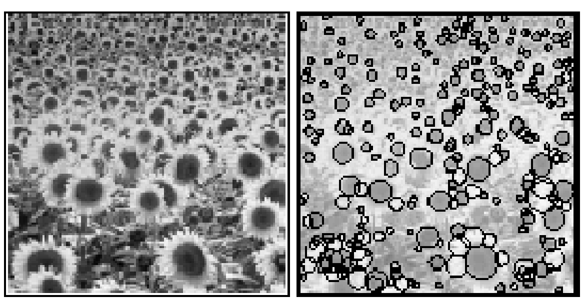


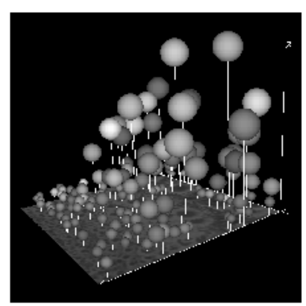
Lowe's Scale-space Interest Points



- **Laplacian of Gaussian** kernel
- Scale-space detection
  - Find local maxima across scale/space
  - A good "blob" detector



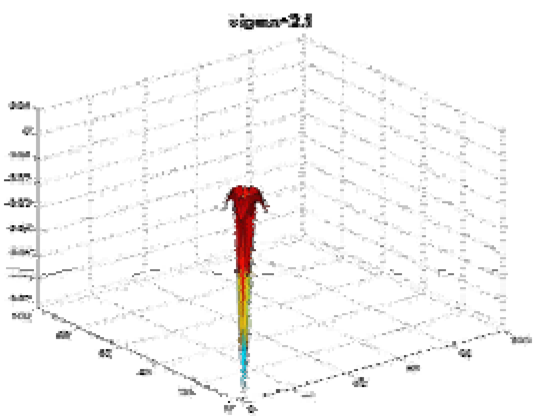


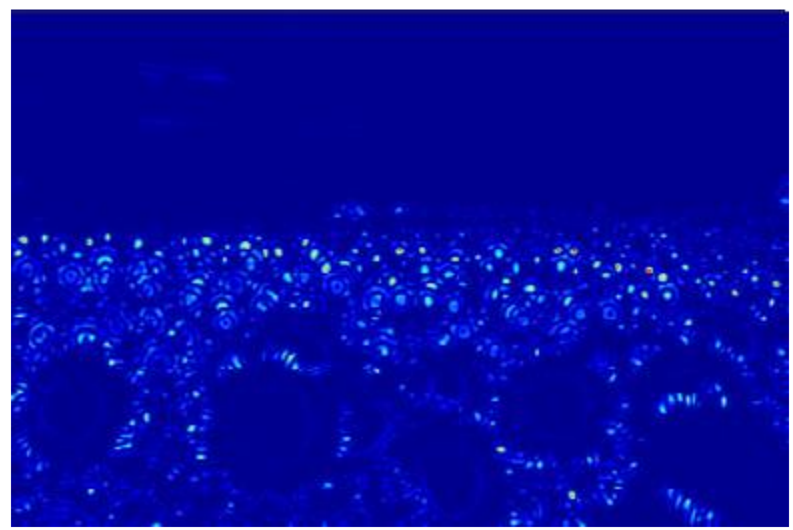


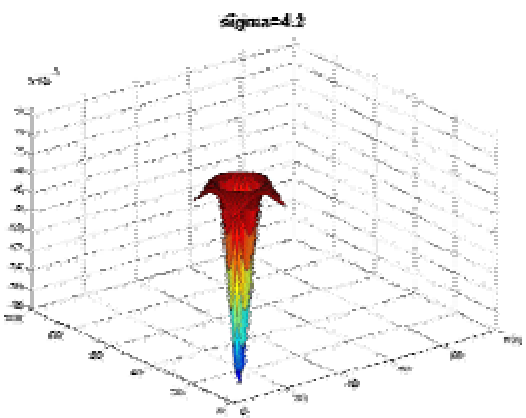
$$G(x, y, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{x^2+y^2}{\sigma^2}}$$

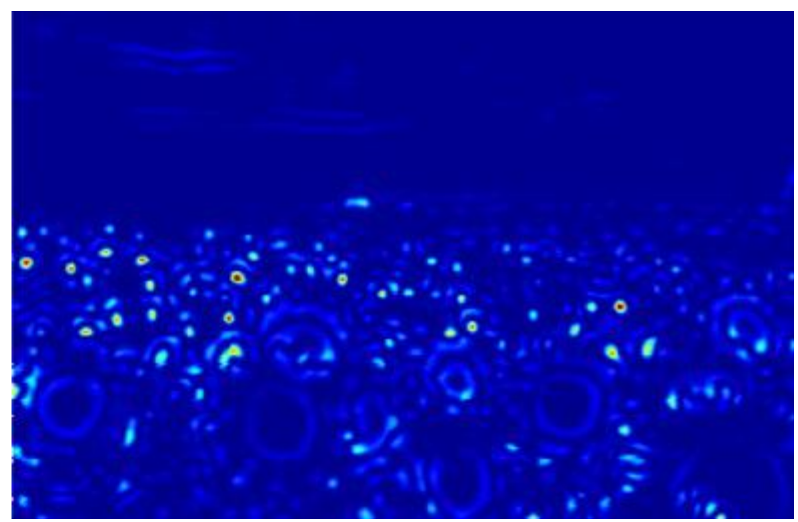
$$\nabla^2 G(x, y, \sigma) = \frac{\partial^2 G}{\partial x^2} + \frac{\partial^2 G}{\partial y^2}$$

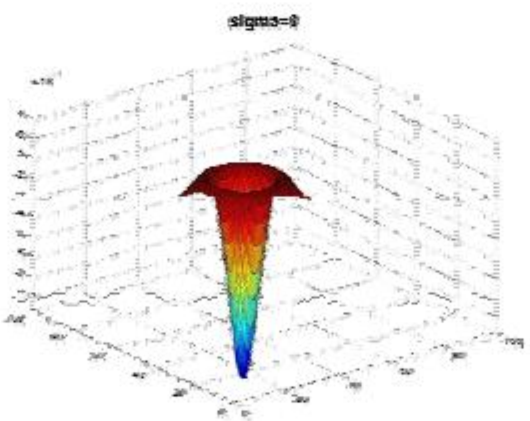
[T. Lindeberg IJCV 1998]

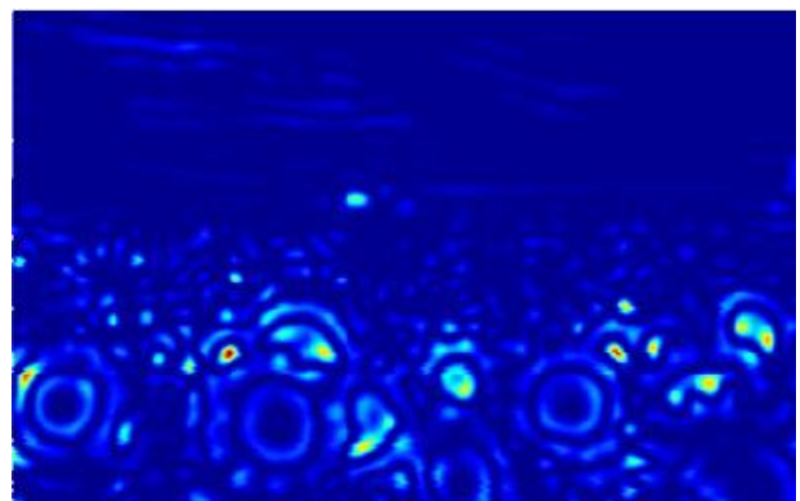


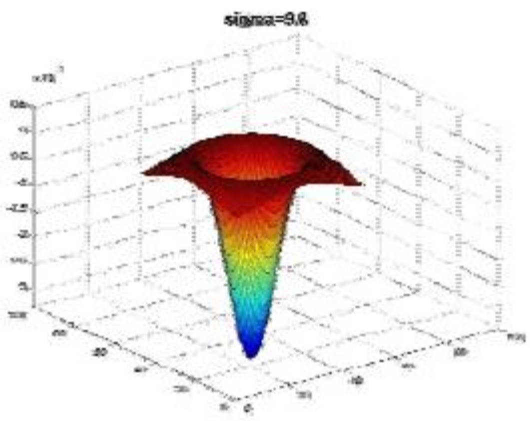


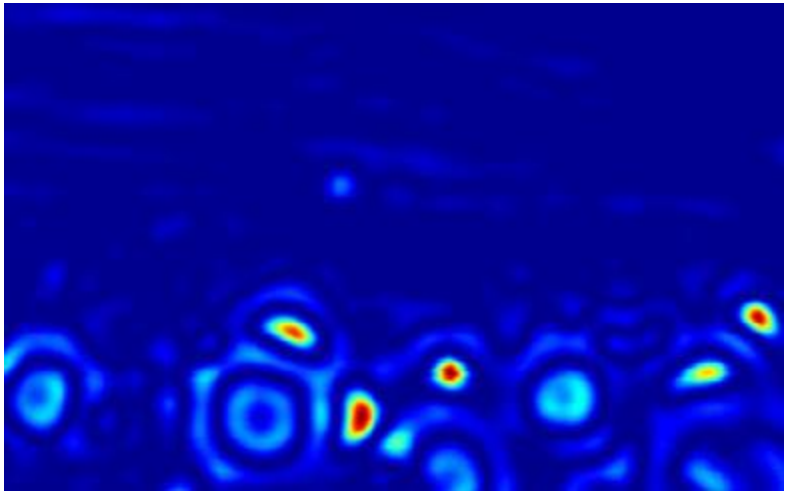

sigma=4.2

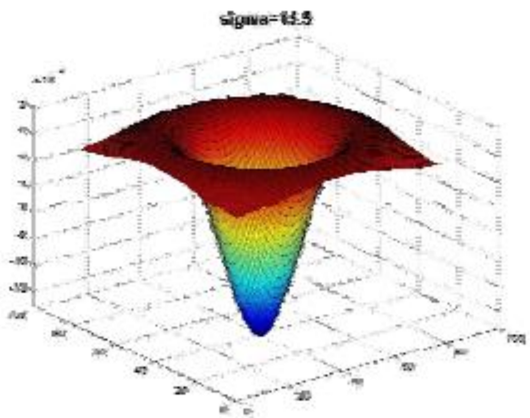


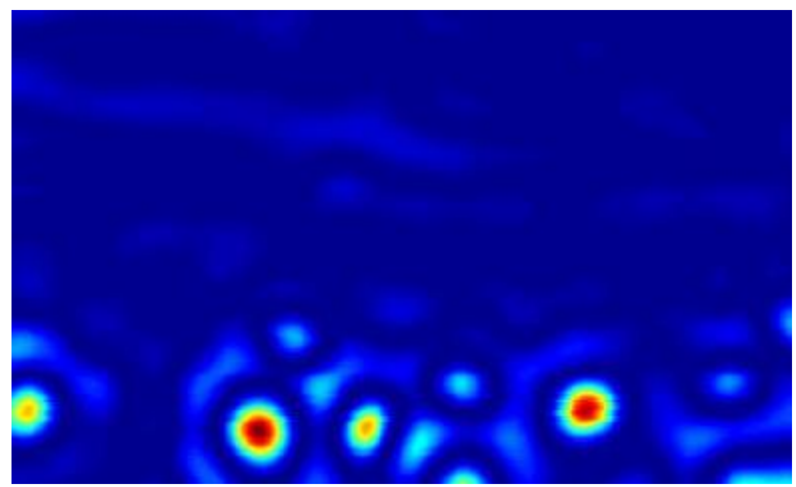


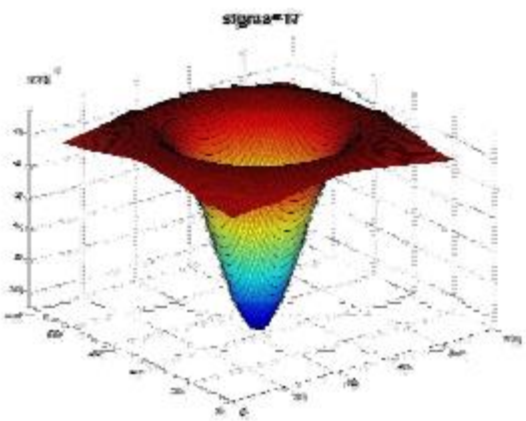


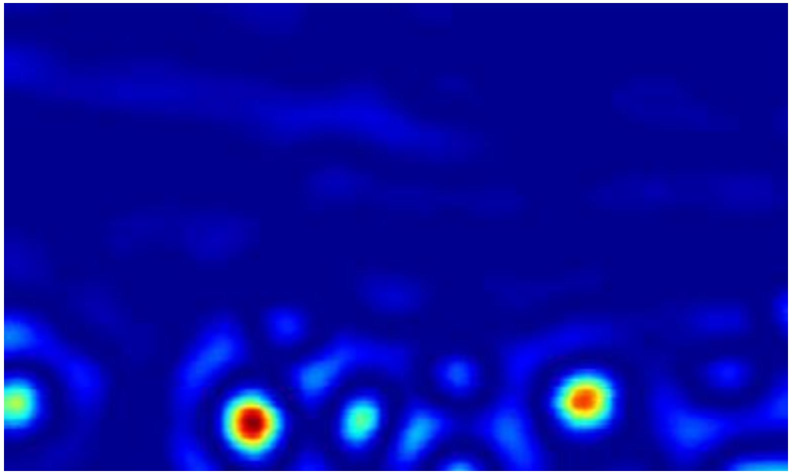




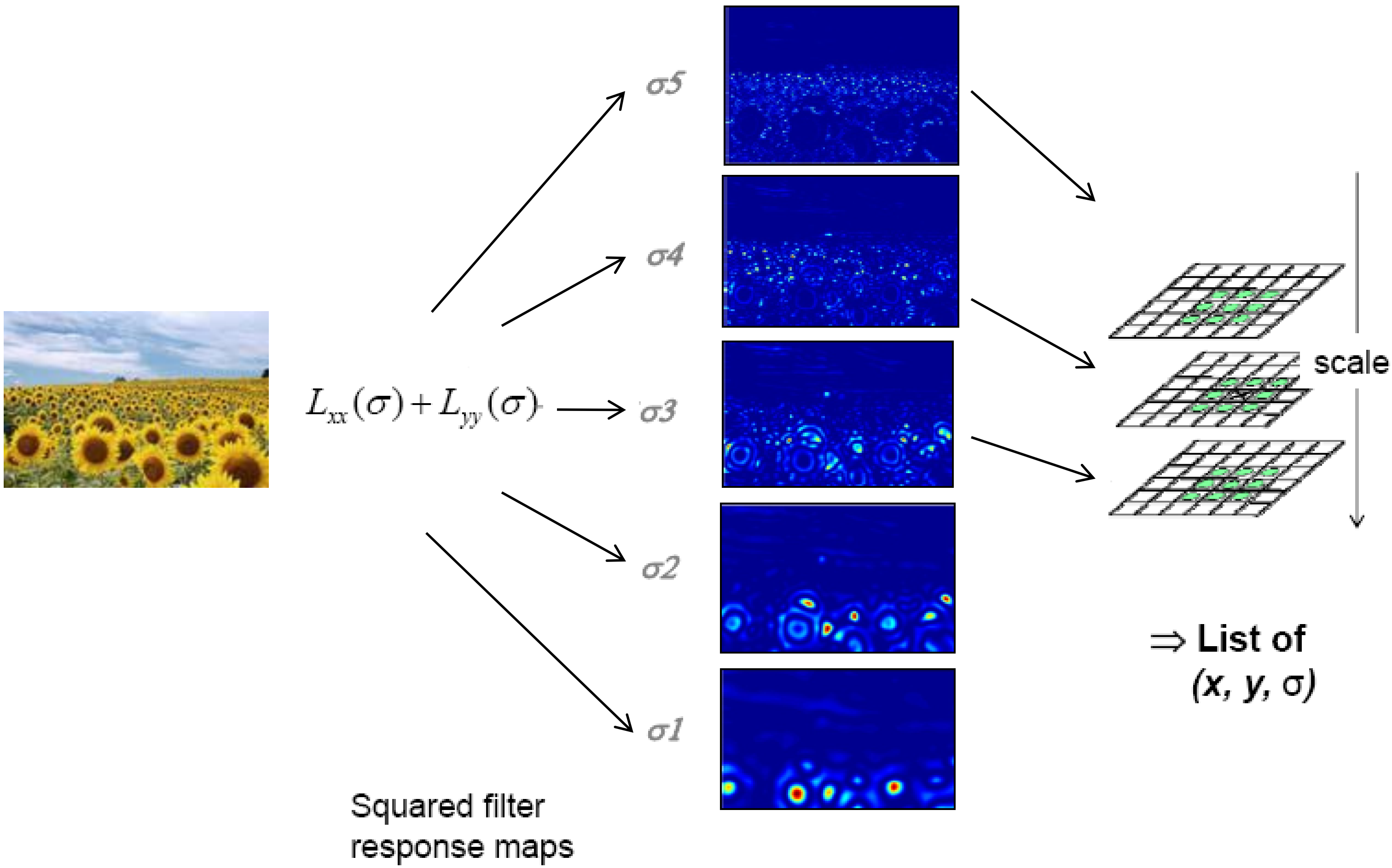




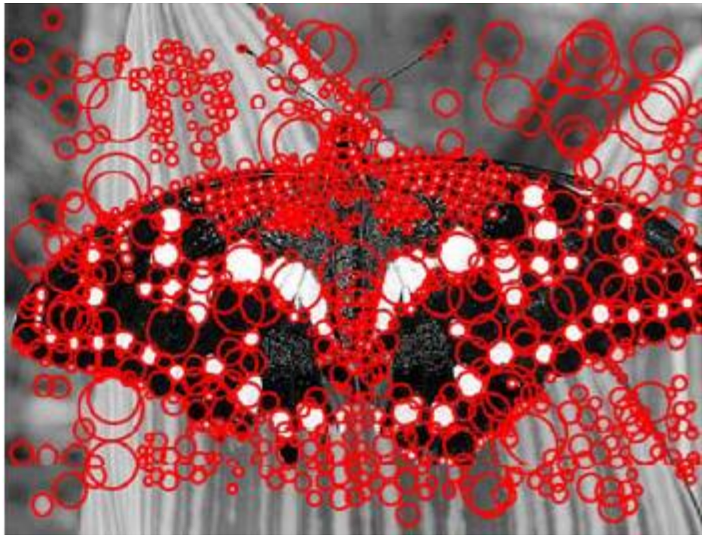




Scale-space blob detection



Scale-space blob detector: Example

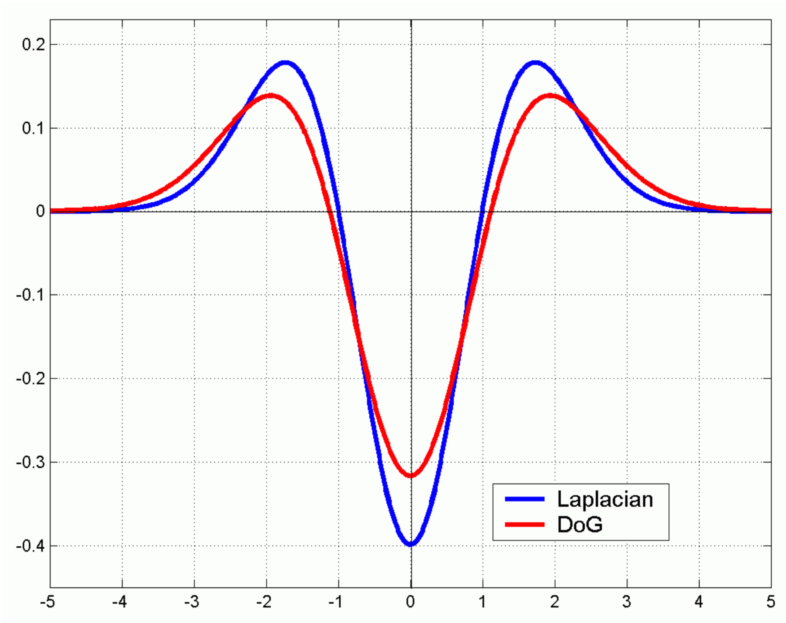


LoG V.S. DoG

$$\nabla^2 G_\sigma(x,y) = \left(\frac{x^2+y^2}{\sigma^4} - \frac{2}{\sigma^2} \right) G_\sigma(x,y)$$

$$\underbrace{G(x,y,k\sigma) - G(x,y,\sigma)}_{\text{DoG}} \approx (k-1)\sigma \frac{\partial G}{\partial \sigma} = (k-1)\sigma^2 \underbrace{\nabla^2 G}_{\text{LoG}}$$

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

Technical detail

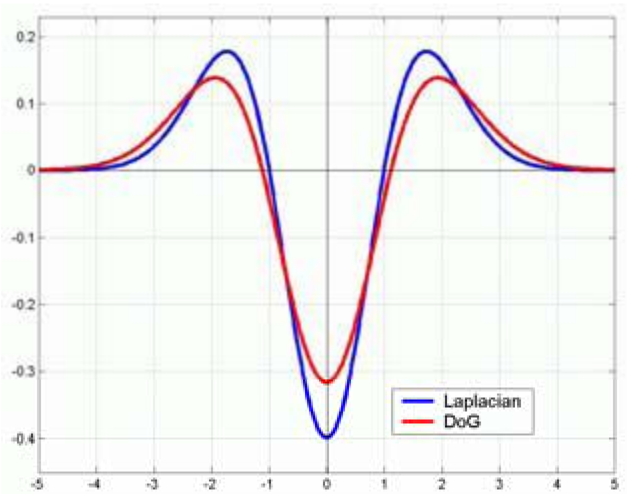
- We can approximate the Laplacian with a difference of Gaussians; more efficient to implement.

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

(Laplacian)

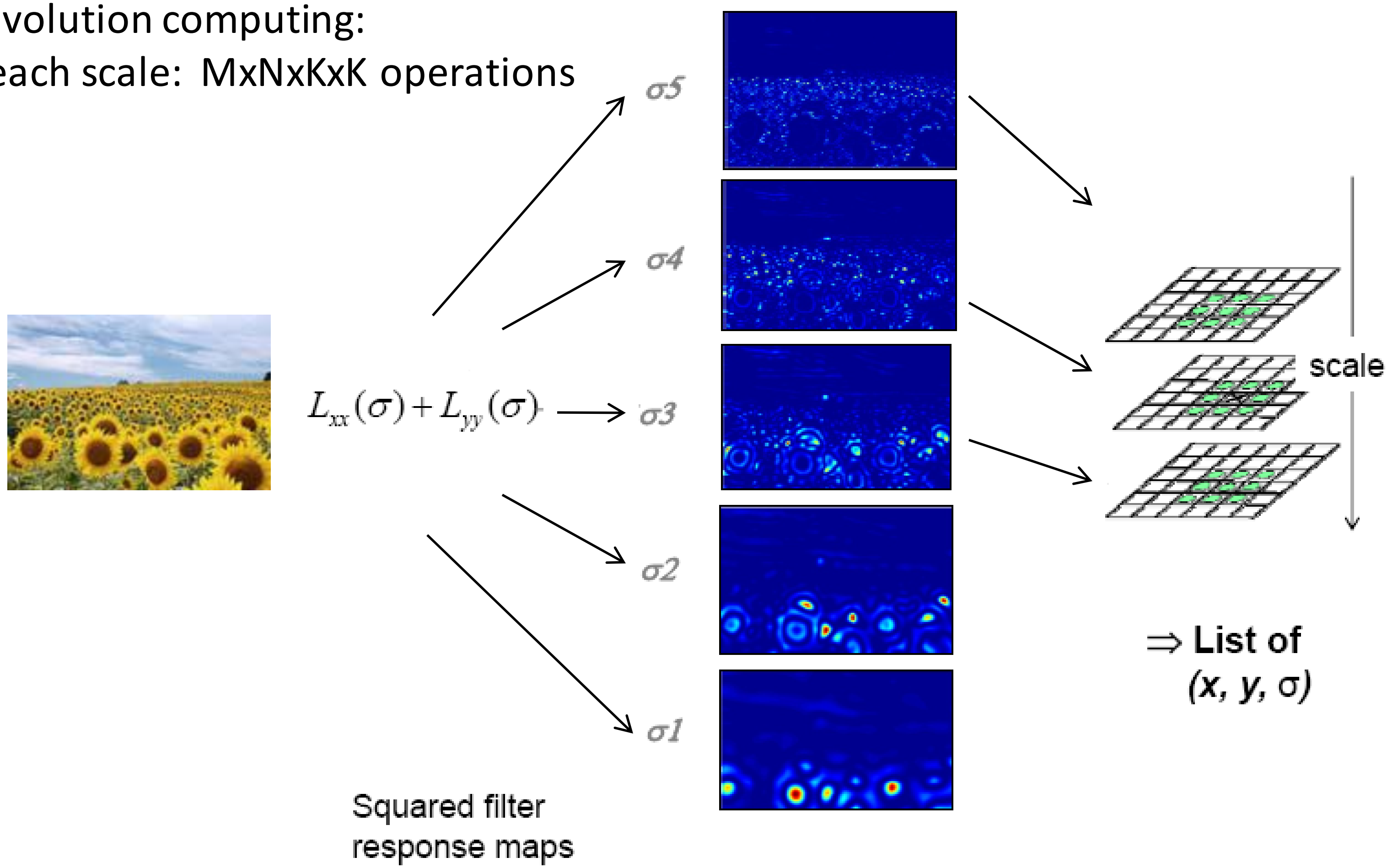
$$DoG = G(x, y, k\sigma) - G(x, y, \sigma)$$

(Difference of Gaussians)



How many scales?

convolution computing:
At each scale: MxNxKxK operations



DoG Image Pyramid

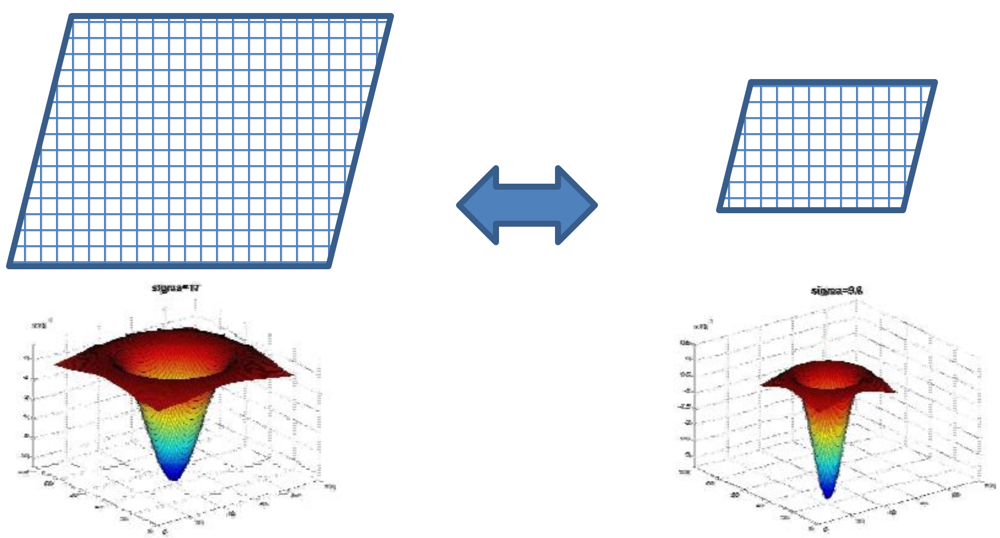
$$\sigma_0, k\sigma_0, k^2\sigma_0, k^3\sigma_0, k^4\sigma_0, k^5\sigma_0, k^6\sigma_0, \ldots$$

$$\sigma_0 \to 2\sigma_0$$

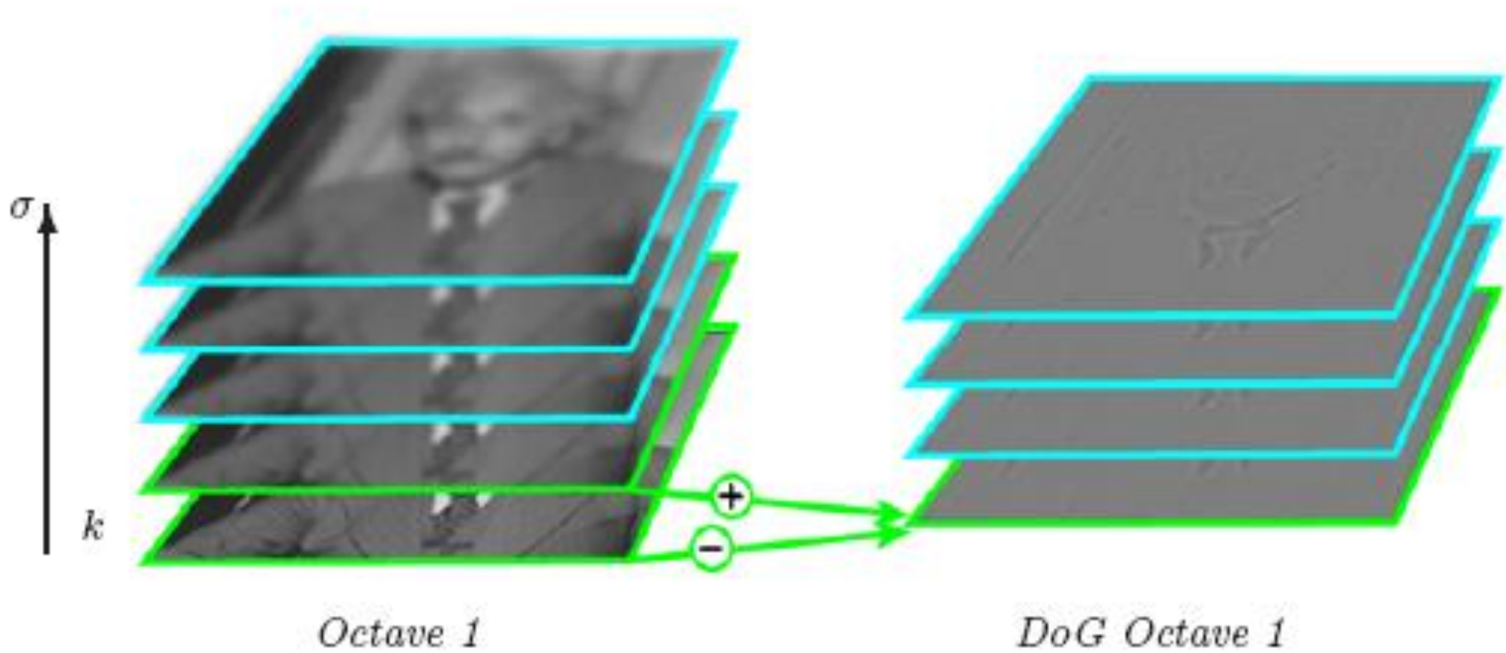
image MxN, filter 2Kx2K

$\downarrow$

image M/2xN/2, filter, KxK

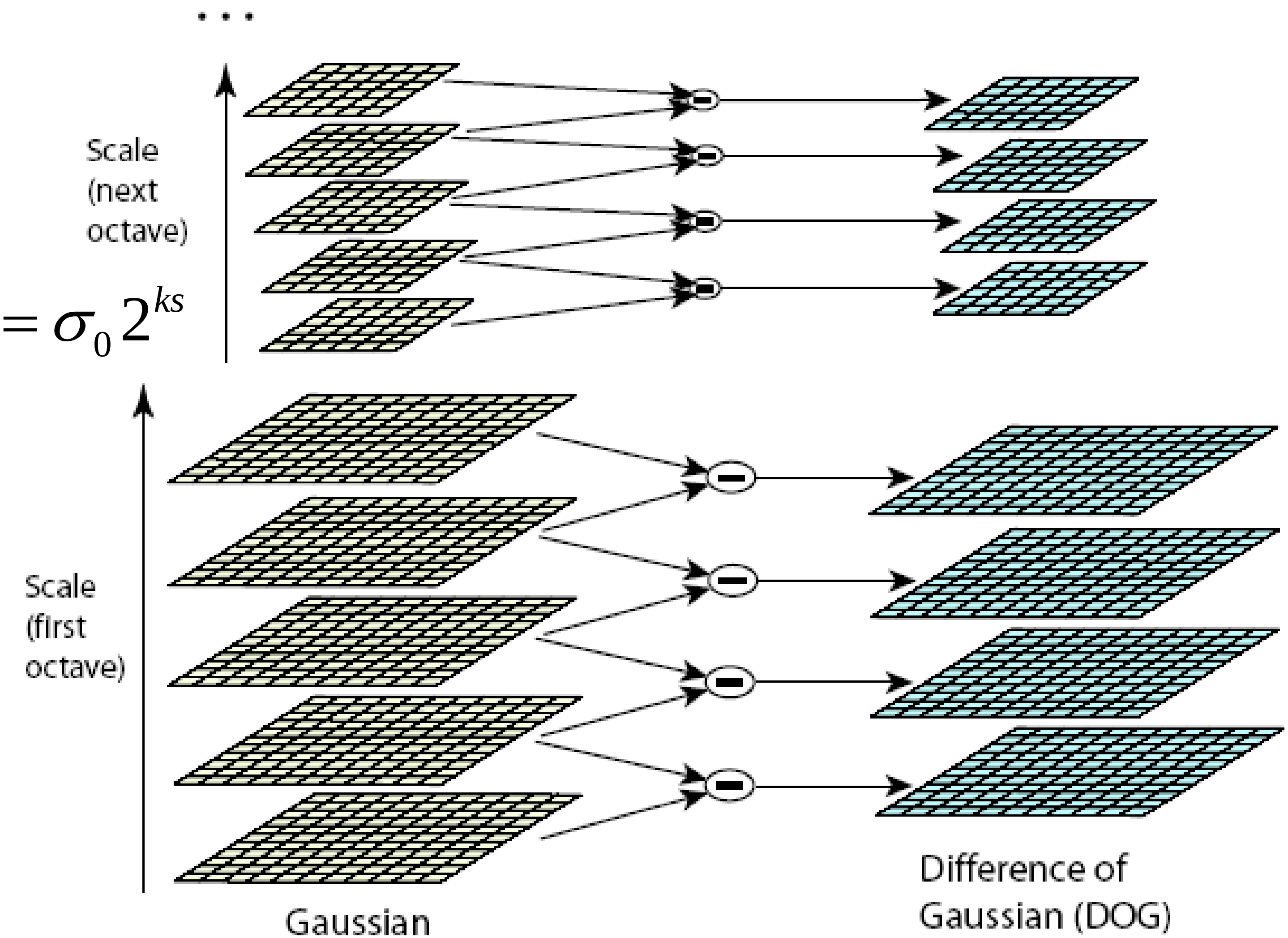


DoG Image Pyramid



$$\sigma_0 \rightarrow 2\sigma_0$$

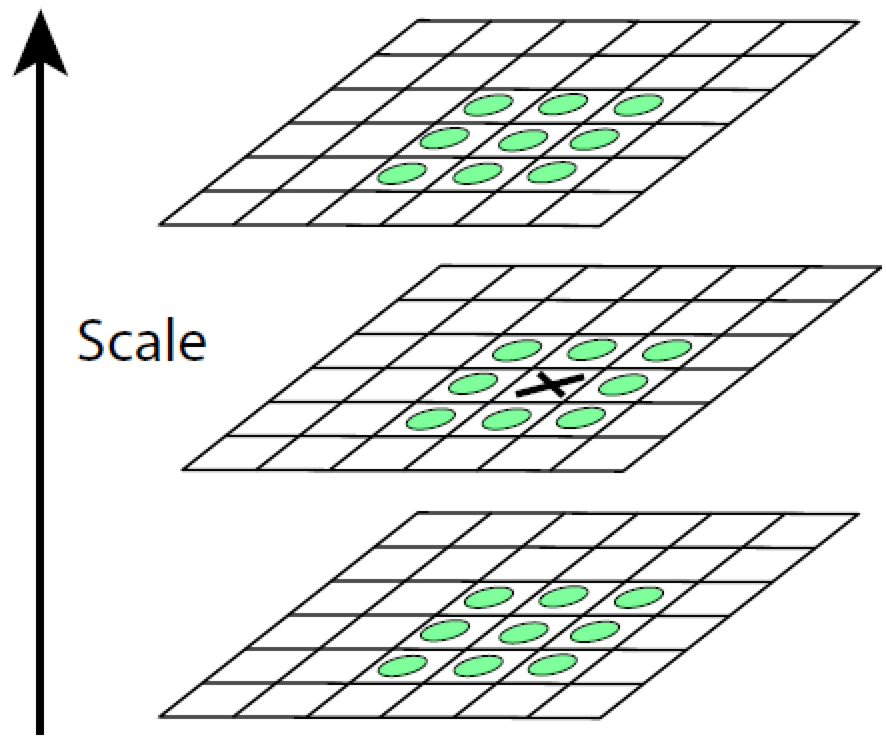
$$\sigma_{o,s} = \sigma_0 2^{ks}$$



$$\sigma_{o,s} = \sigma_0 2^{o+ks}$$

Local Extrema Detection

- Maxima and minima
- Compare x with its 26 neighbors at 3 scales



Frequency of sampling in scale

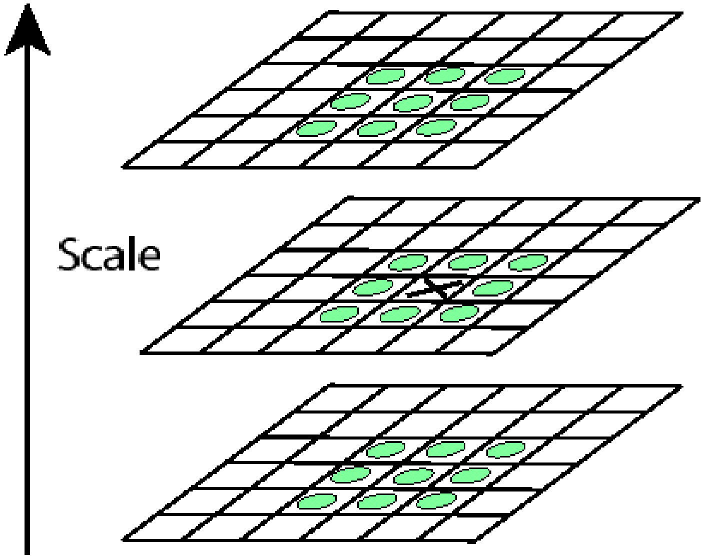
- s: intervals in each octave of scale space ($\sigma_0 \rightarrow 2\sigma_0$)
  - k=2^{1/s}

$$\sigma_{o,s} = \sigma_0 2^o k^s$$

- In order to cover a complete octave for extrema detection
  - S = s+3 Gaussian images are produced for each octave
    - s: {-1,S+1}
  - s+2 DoG images
  - s scales for extrema detection

SIFT Key point localization

- Detect maxima and minima of difference-of-Gaussian in scale space
- Fit a quadratic to surrounding values for sub-pixel and sub-scale interpolation (Brown & Lowe, 2002)
- Taylor expansion around point:
- Offset of extremum (use finite differences for derivatives):

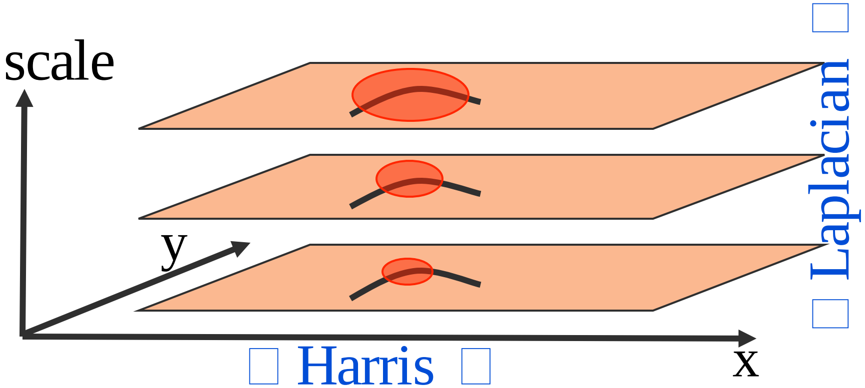


$$D(\mathbf{x}) = D + \frac{\partial D}{\partial \mathbf{x}}^T \mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathbf{T}} \frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x}$$

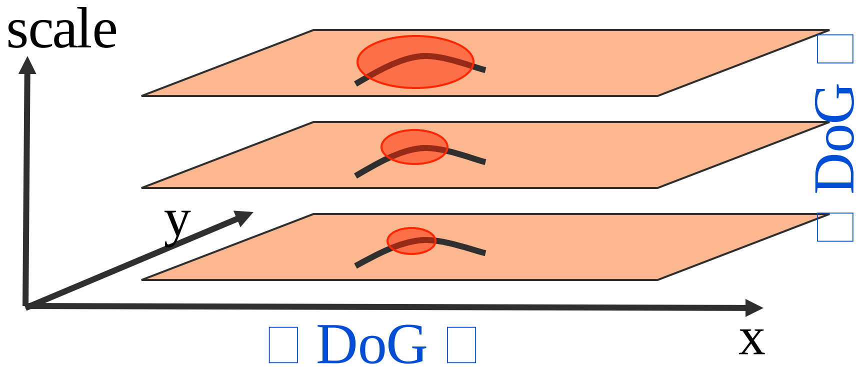
$$\hat{\mathbf{x}} = -\frac{\partial^2 D}{\partial \mathbf{x}^2}^{-1} \frac{\partial D}{\partial \mathbf{x}}$$

Scale Invariant Detectors

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



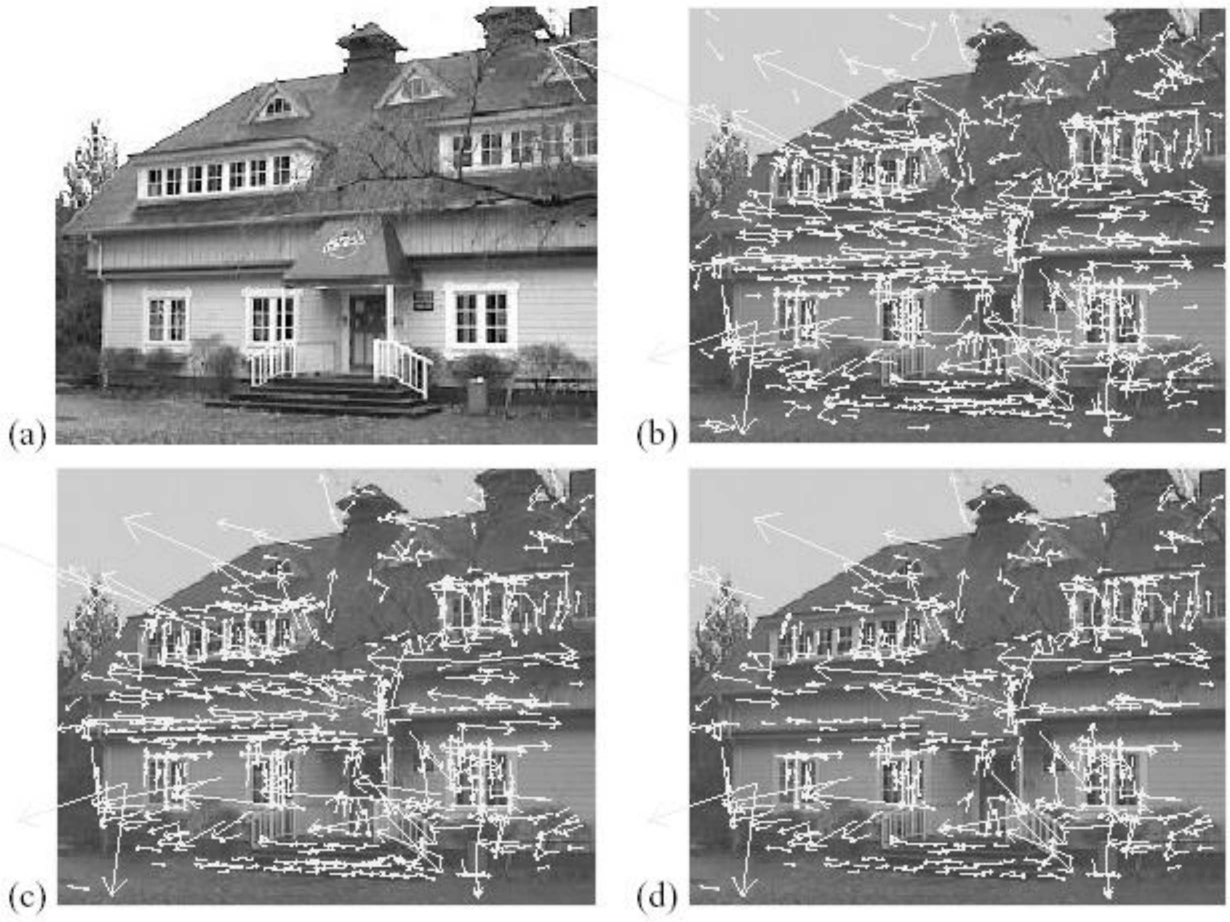
- SIFT (Lowe)[2]
  Find local maximum of:
  - Difference of Gaussians in space and scale



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

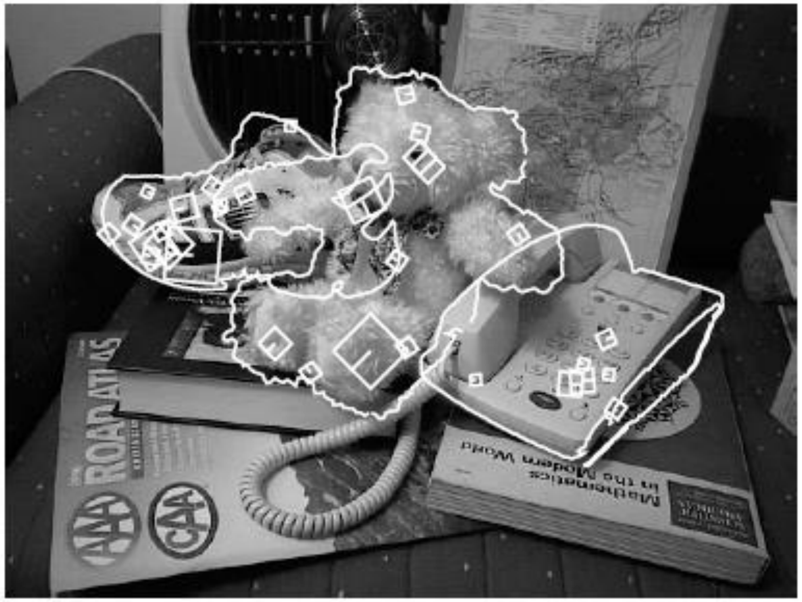
Example of keypoint detection

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


Mathematics in the Modern World


Mathematics
in the Modern World

ROAD ATLAS
Mathematics
in the Modern World

SIFT vector formation

- Thresholded image gradients are sampled over 16x16 array of locations in scale space
- Create array of orientation histograms
- 8 orientations x 4x4 histogram array = 128 dimensions

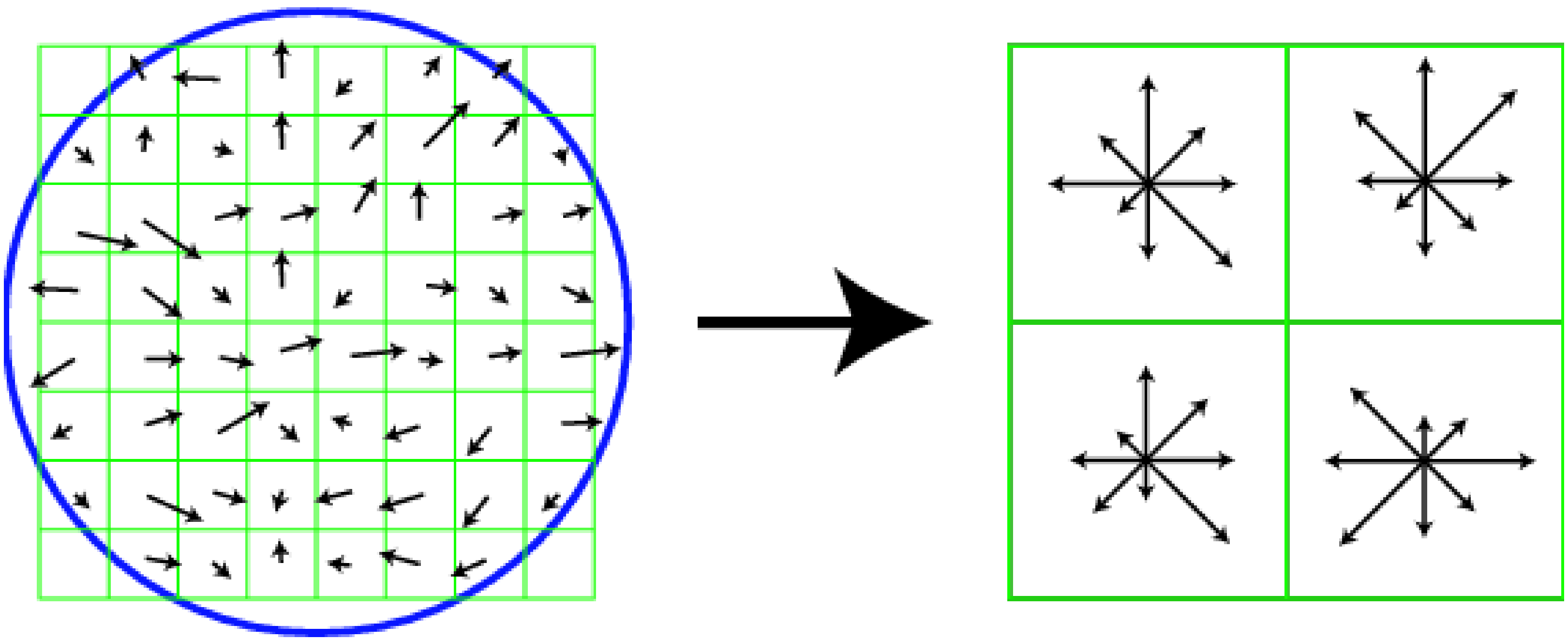


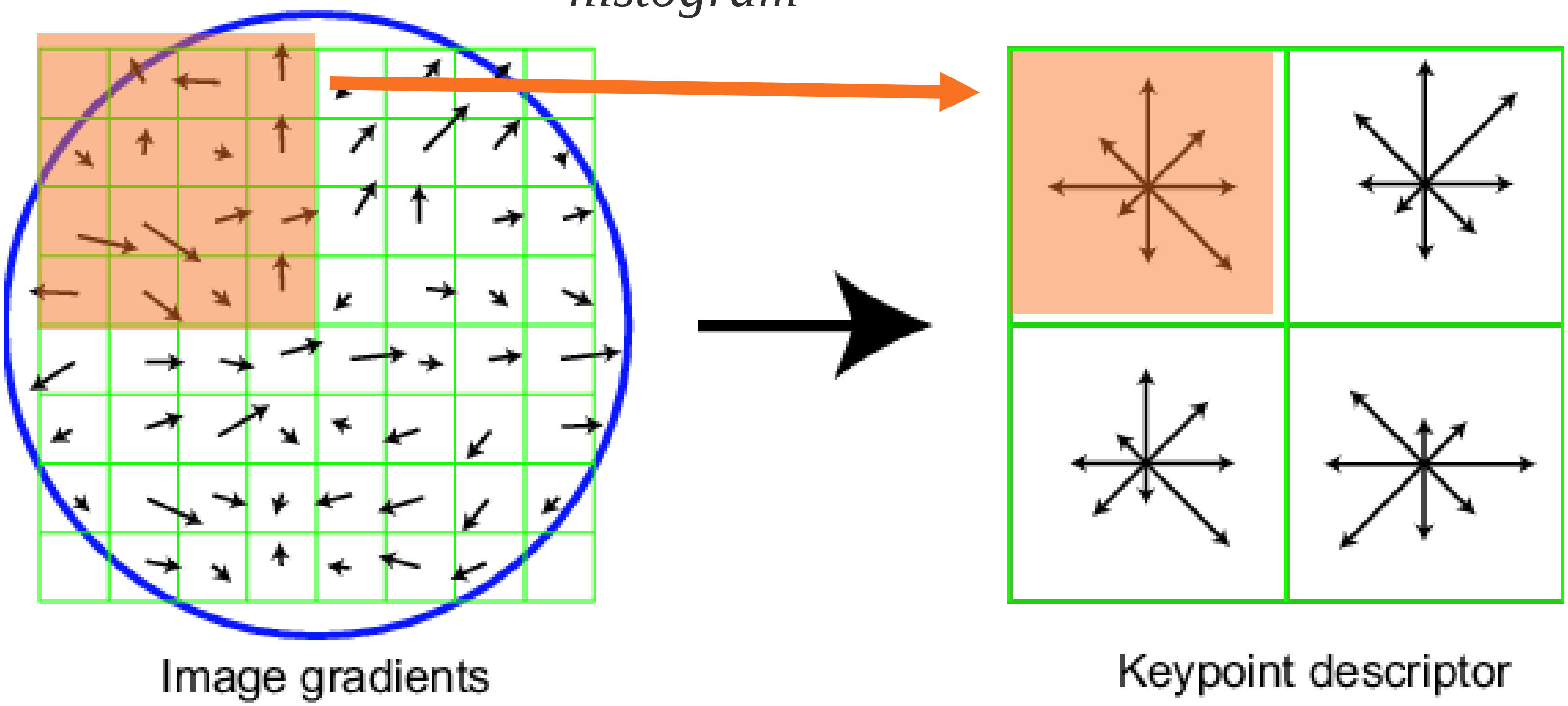
Image gradients   Keypoint descriptor

SIFT vector formation

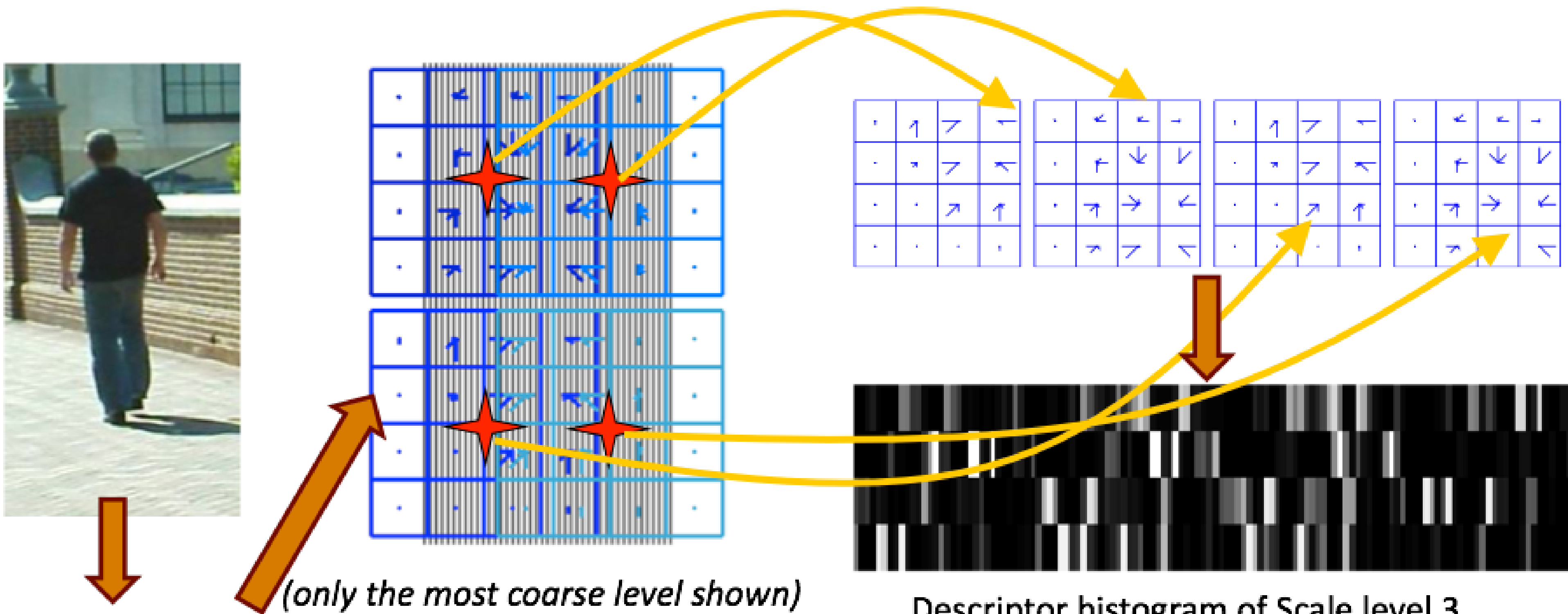
- Orientation is defined relative to the orientation of the detected Sift feature

Local orientation

histogram

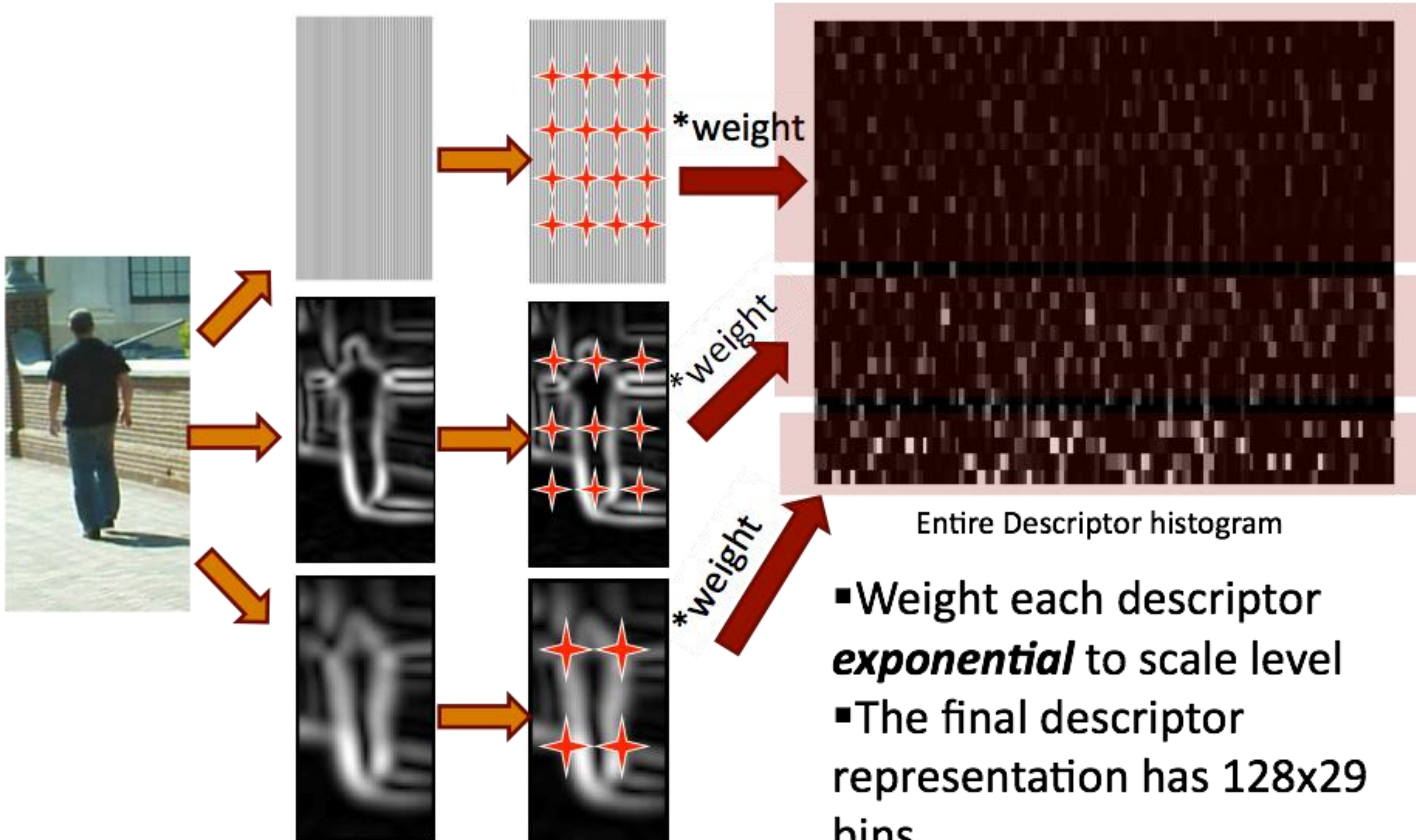


Feature Extraction for Image Classifier



(only the most coarse level shown)

Descriptor histogram of Scale level 3

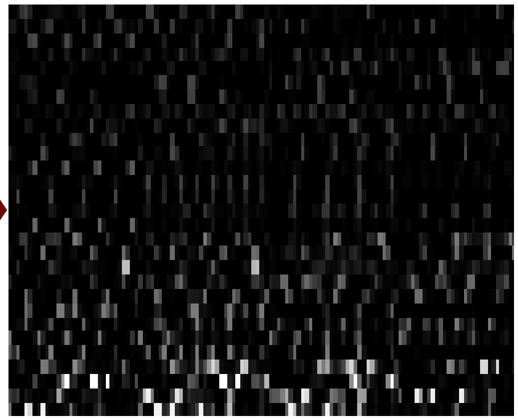


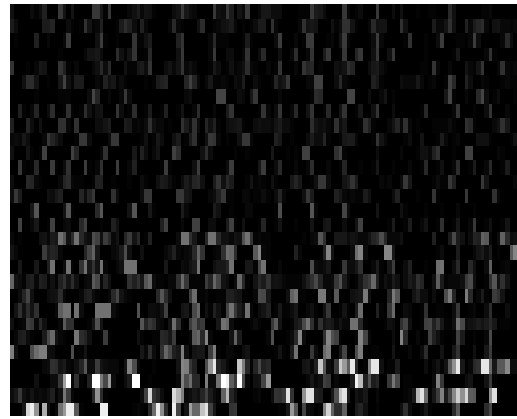
- Weight each descriptor ***exponential*** to scale level
- The final descriptor representation has 128x29 bins

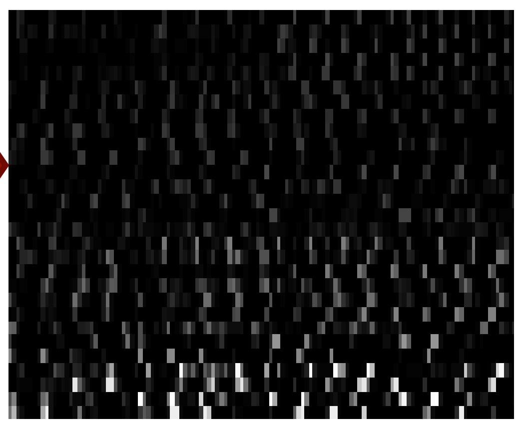
Feature Extraction for Image Classifier

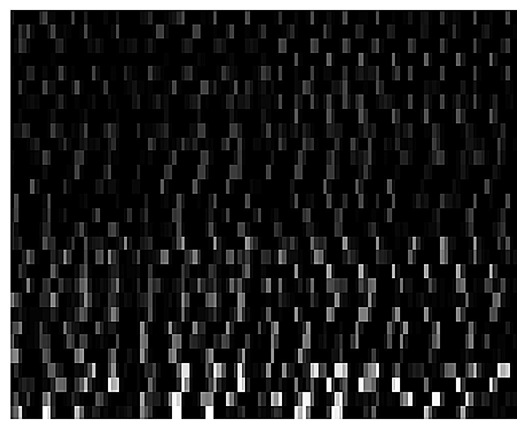
Examples from training set

Examples from test set

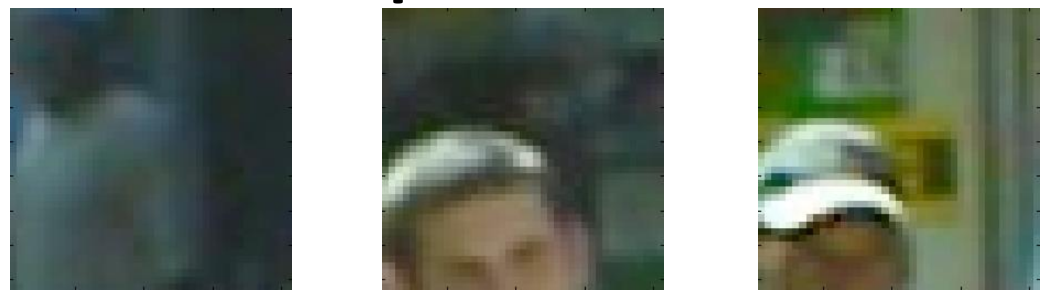




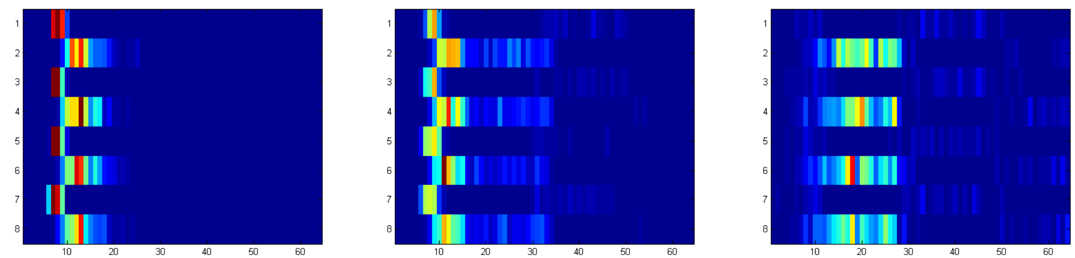




Features Sample

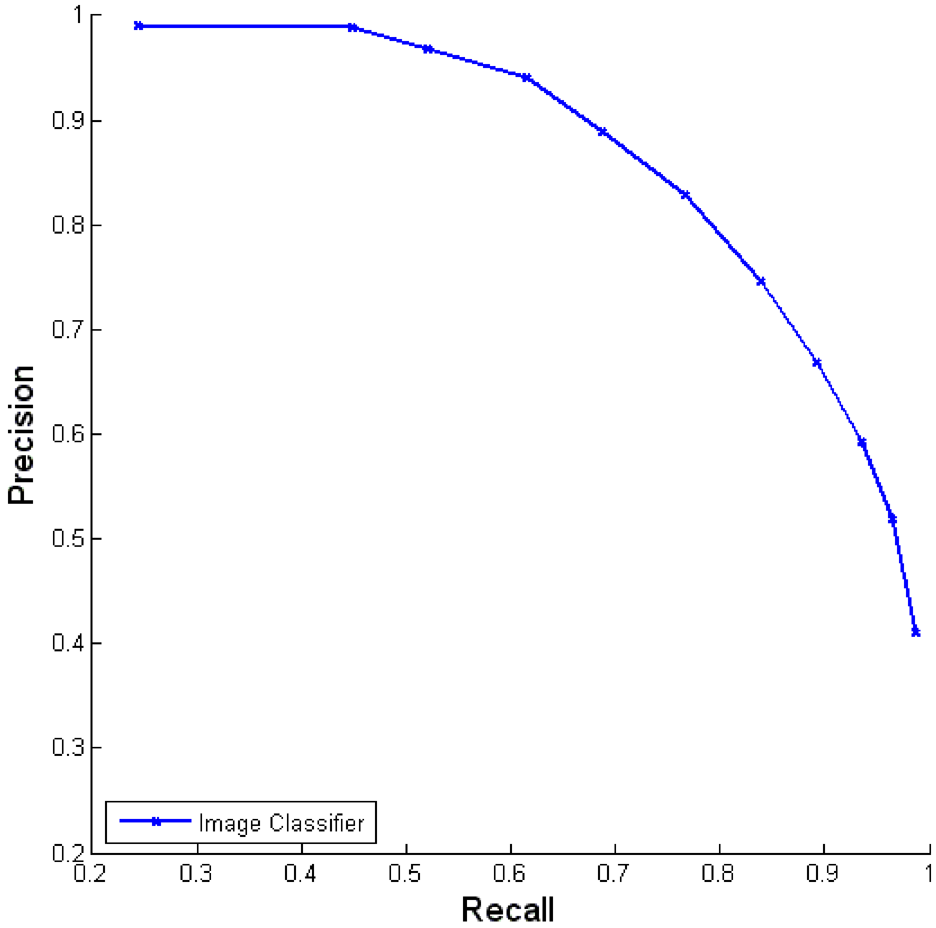


1. Color Histogram



2. HOG feature

Image Classifier Result



- Precision-Recall curve of our Image classifier
- Trained/Tested on image set generated from stereo detection
- RBF kernel SVM used

Typical Missed Detections

- ***Occlusions***

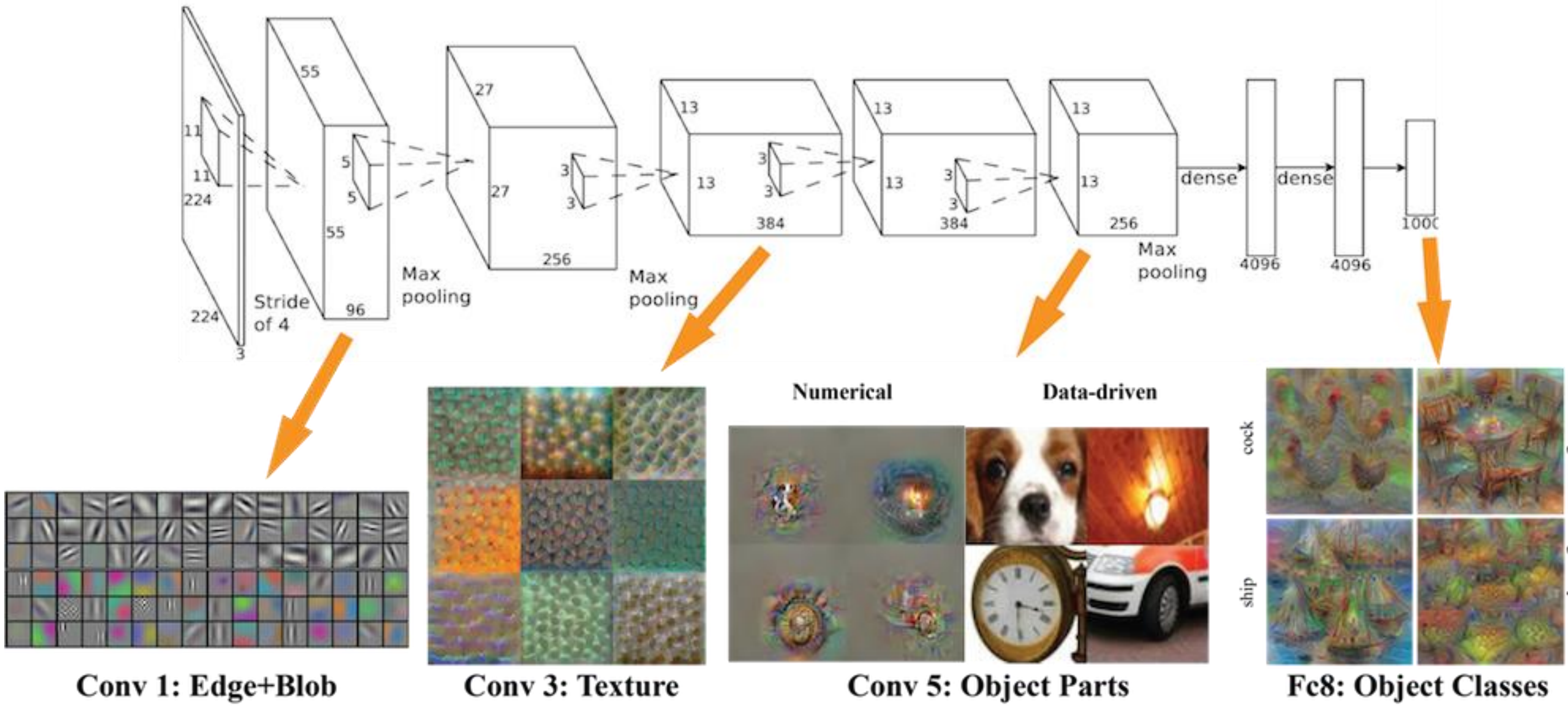
- ***Incomplete stereo detection***

- ***Lack of training data***

Typical False Positives

- ***Human-like Shapes & Clutters***

Deep…



224
224
3
11
11
Stride of 4
55
55
96
5
5
Max pooling
27
27
256
3
3
Max pooling
13
13
384
3
3
13
13
384
3
3
13
13
256
Max pooling
dense
dense
4096
4096
1000
Numerical
Data-driven
cock
ship
dinning table
grocery store
Conv 1: Edge+Blob
Conv 3: Texture
Conv 5: Object Parts
Fc8: Object Classes