SIFT (Scale Invariant Feature Transform)

Reference – Distinctive Image Features from Scale-Invariant Key points (David G. Lowe)

1. Scale Space Extrema Detection

Theory:

Scale space of an image is defined as a function, L(x, y, σ), that is produced from the convolution of a variable-scale Gaussian, G(x, y, σ), with an input image, I(x, y):

$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y),$

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

$D(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y) = L(x,y,k\sigma) - L(x,y,\sigma).$

DoG is used because:
- It is computationally efficient and it is simply calculated by difference of Gaussian smoothed images.
- It provides a close approximation to the scale-normalized Laplacian of Gaussian, $\sigma^2 \nabla^2 G$

Local Extrema Detection: Choose those points which are either maxima or minima among 26 neighbors (8 neighbors on the same scale, 9 on the above scale, 9 on the below scale)

I used 4 octaves and 3 scales in the implementation to build the Difference of Gaussian pyramid. Also, Value of sigma for the lower scale is 1.6 and incremented by $\sqrt{2}$ at successive scale in the given octave. When jump from one octave to another, multiply sigma by 2.

2. Key point Localization

Reject Low Contrast points and less stable or poorly localized points.

Theory:

Taylor expansion (up to the quadratic terms) of the scale-space function, D(x, y, σ), shifted so that the origin is at the sample point, where D and its derivatives are evaluated at the sample point and $x = (x, y, \sigma)^T$ is the offset from this point. The location of the extremum, x^, is determined by taking the derivative of this function with respect to x and setting it to zero.

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

Hessian Matrix:

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \qquad \frac{\mathrm{Tr}(\mathbf{H})^2}{\mathrm{Det}(\mathbf{H})} < \frac{(r+1)^2}{r}.$$

r > 10 for eliminating poor edge responses.
Rejecting unstable extrema with low contrast, |D_x_hat| > 0.03
If x_hat > 0.5 in any of the directions, then it means that it is closer to a different sample point.

3. Orientation Assignment

Create a histogram of gradient directions, within a region around the key point, at selected scale. Histogram indices are scaled values of angle (36 bins – 0 to 360 degrees) and histogram values are weights obtained by Gaussian window (sigma = 1.5 times the scale of image) and gradient magnitude of image

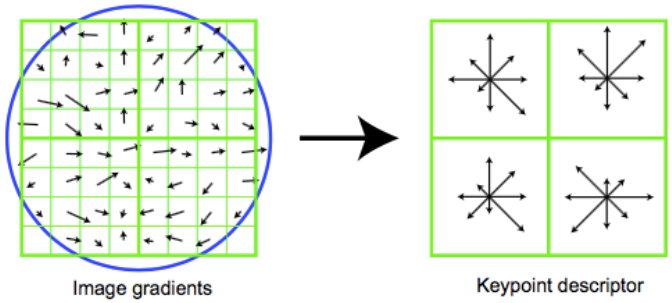
$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2}$$

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

4. Key point Descriptor

A keypoint descriptor is created by first computing the gradient magnitude and orientation at each image sample point in a region around the keypoint location, as shown on the left. These are weighted by a Gaussian window, indicated by the overlaid circle. These samples are then accumulated into orientation histograms summarizing the contents over 4x4 subregions, as shown on the right, with the length of each arrow corresponding to the sum of the gradient magnitudes near that direction within the region. This figure shows a 2x2 descriptor array computed from an 8x8 set of samples, I use 4x4 descriptors computed from a 16x16 sample array.

It generates a feature vector of length 128.



Advantages of SIFT:
Locality – It is local, so robust to occlusion and clutter.
Distinctiveness – Individual features can be matched to a large database of objects.
Quantitative – Number of key points generated for even a small object is huge.
Efficiency – Close to real time performance