1. Reviewer:

쿠쿠

2. Article:

David G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision 60(2), pp.91-110, 2004.

3. SIFT 알고리즘 설명:

Detection of Scale-Space Extrema

Create Scale-Space:

scale-space 이론의 목적은 영상의 multi-scale 성질을 모델링 하는 것임.

Scale-Space of an image:

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \qquad \text{where} \quad G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

DoG (Difference-of-Gaussian):

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

(x,y): 공간좌표, σ: scale 좌표

Why DoG??? 1) 계산이 간단함, 2) scale-normalized LoG (Laplacian of Gaussian)의 approximation임.

영상 피라미드: 영상 피라미드는 O 개의 octave가 있고, 각각의 octave에는 $S+3$ 개의 Gaussian blurring 영상 $L(x,y,\sigma)$가 있음, 다음 octave는 앞의 octave의 영상을 factor 2로 다운샘플링 하여 얻음. DoG는 영상 피라미드에서 인접한 두 영상의 차영상임.

Octave: 하나의 octave는 scale σ 값을 doubling하는 것에 상응함.

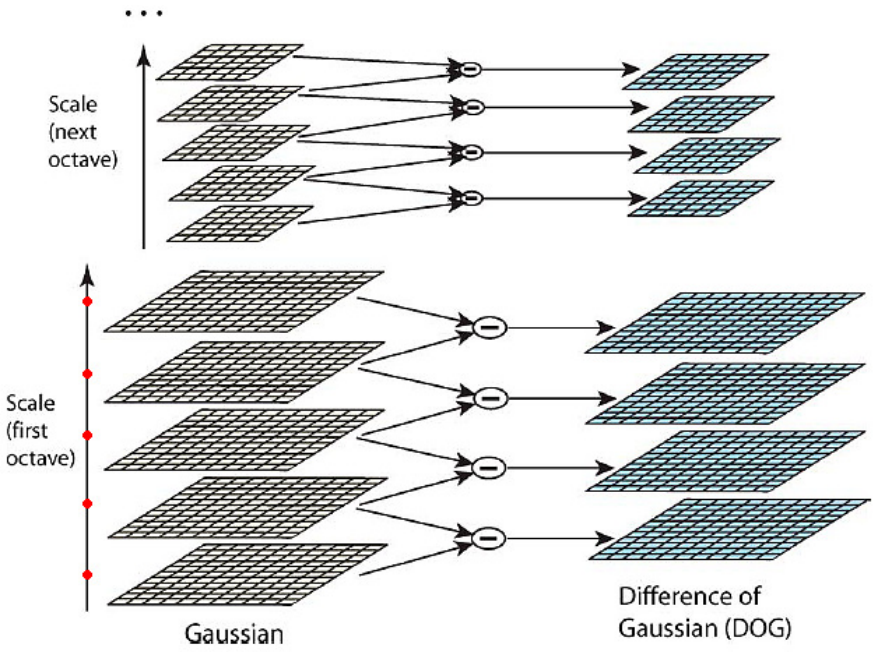

그림 1. Two octaves of a Gaussian scale-space image pyramid with S=2 intervals & DoG Pyramid.

Local Extrema Detection in DoG:

각각의 sample point(pixel)는 그를 중심으로 한 DoG의 $3\times3\times3$ 영역의 26개의 pixel과 비교하여

extrema인가를 결정 (이는 scale 공간 (σ) 과 2D 영상 공간 (x, y) 두 곳 모두에서 extrema를 검출함을 보증함), sample point가 local extrema일 때 sample point는 후보 keypoint로 됨.

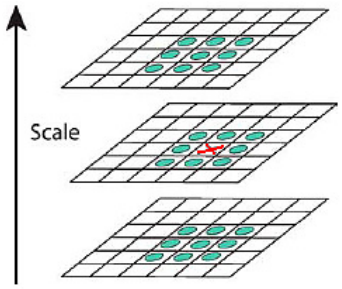

그림 2. Local extrema detection in DoG.

[Lowe]의 논문에서 사용된 parameters:

$O: ???, \quad S: 3 \text{scale/octave}, \quad \sigma_0 = 1.6 \cdot 2^{1/S}, \quad \sigma_n = 0.5, \quad O_{\min} = -1.$

Accurate Keypoint Localization

3D quadratic함수를 적용하여 후보 keypoint의 (x, y, σ) 를 결정함.

Contrast가 작은 후보 keypoint 제거:

3D quadratic함수를 적용한 후 얻어진 후보 keypoint의 새로운 좌표 $\hat{\mathbf{x}} = (\hat{x}, \hat{y}, \hat{\sigma})^T$ 가 $D(\hat{\mathbf{x}}) < 0.03$ 을 만족하면 제거 (assume: pixel value range is $[0\ \ 1]$).

Edge Responses 제거:

Poorly defined된 DoG의 extrema는 edge를 across하는 방향에서 큰 principal curvature를 갖고 있음, 반면에 edge에 수직인 방향에는 작은 principal curvature를 갖고 있음.
후보 keypoint의 위치에서 구한 Hessian 매트릭스 $\mathbf{H}$ 의 eigenvalue는 principal curvature에 정비례함.

$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}$, α, β 를 각각 $\mathbf{H}$ 의 큰 eigenvalue와 작은 eigenvalue라고 하면:

$Tr(\mathbf{H}) = \alpha + \beta, \quad Det(\mathbf{H}) = \alpha\beta.$

$r = \alpha / \beta$ 라 가정하면, $Tr(\mathbf{H})^2 / Det(\mathbf{H}) = (r+1)^2 / r.$

$(r+1)^2 / r$ 은 α, β 가 같은 값일 때 가장 작음, r 가 증가함에 따라 $(r+1)^2 / r$ 도 증가.

$Tr(\mathbf{H})^2 / Det(\mathbf{H}) < (r+1)^2 / r$ (where $r = 10$)을 만족하면 edge response라 결정하며 후보 keypoint에서 제거됨.

Orientation Assignment

Keypoint 영역 pixel들의 gradient 방향분포특성을 이용하여 각각의 keypoint에 방향을 할당.

Keypoint를 중심으로 하는 영역에서 pixel들을 선택하여 각각의 gradient 방향을 계산한 후 histogram을 그린다. Histogram에 포함되는 pixel들은 각각의 gradient magnitude와 Gaussian circular window에 의해 weight됨 ($\sigma_{\text{Gaussian circular window}} = 1.5 \cdot \sigma_{\text{keypoint}}$). Histogram의 범위는 $[0\ \ 360^{\circ}]$이며 10°를 하나의 bin으로 표현하여 모두 36개의 bin을 포함한다. Histogram의 가장 큰 peak값을 keypoint의 주 방향으로 정한다. Histogram에 가장 큰 peak값의 80%에 해당하는 값들이 존재한다면 그 방향을 keypoint의 보조 방향으로 정한다.

위의 3개의 과정을 거쳐 keypoint 검출이 완료된다.
각각의 keypoint는 3개의 정보를 포함한다: location (x, y), scale σ, orientation θ.

The Local Image Descriptor

우선, 좌표축을 keypoint의 방향으로 회전하여 rotation invariant를 보증.
다음, keypoint를 중심으로 8×8 pixel들을 선택함, pixel의 gradient magnitude와 orientation은 Gaussian circular window에 의해 weight됨. 각각의 4×4 pixel에서 매개 pixel 위치의 gradient 방향은 축적(accumulate)되어 8개의 bin을 포함하는 gradient 방향 histogram을 그린다. 이때 4×4 pixel은 하나의 seed로 되며, 이 seed에는 8개의 방향정보가 있다. keypoint는 $2\times2=4$개의 seed로 구성된다.
[Lowe]의 논문에서는 각각의 keypoint는 $4\times4=16$개의 seed로 describe할 것을 건의.

최종적으로 하나의 keypoint로부터 $4\times4\times8=128$개의 SIFT Feature Vector가 생성됨.
이때, SIFT Feature Vector는 scale, rotation 등 기하변환의 영향을 제거한 것임, 나아가서 SIFT Feature Vector를 normalize하여 illumination의 영향을 제거할 수 있음.

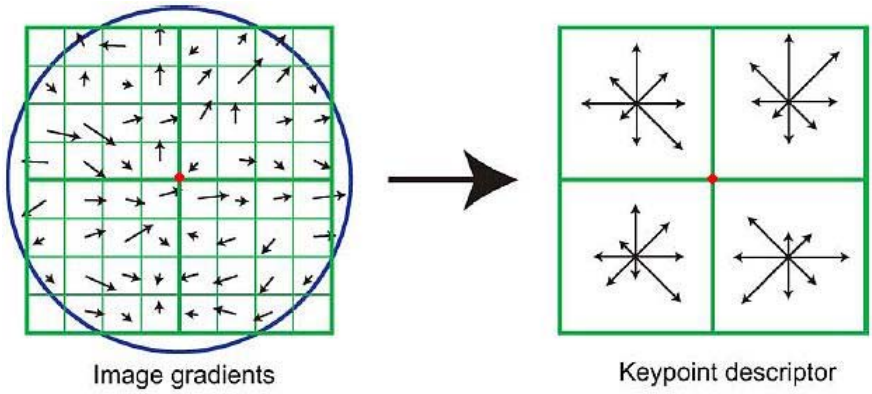

그림 3. Create keypoint descriptor.

Matching:

영상1의 i-th keypoint와 영상2의 keypoint들과의 Feature Vector의 Euclidian 거리를 계산하여 가장 가까운 두 개의 keypoint를 선택. Euclidian 거리가 각각 d_1, d_2이고 $d_1 < d_2$일 때, $(d_1 / d_2 < threshold\ value)$이면 keypoint가 매칭됨.