SIFT

Scale Invariant Feature Transform
by David Lowe

Short Explanation of the Approach
By Michela Lecca
What is SIFT?

- SIFT is an algorithm developed by David Lowe in 2004 for the extraction of interest points from gray-level images.
- The algorithm is described in
- A C++ implementation is available on the net
  http://www.vlfeat.org/~vedaldi/code/siftpp.html
What is SIFT?

- The input is a gray-level image. The output is a list of 2D points on the image each associated to a vector of low-level descriptors. These points are said keypoints and their descriptors are invariant by rescaling, in-plane rotating, noise addition and in some cases by changes of illuminant.

- Keypoints provide a local image description.

- They are used to find visual correspondences between images for different applications, like image alignment or object recognition.
Example: SIFT Image Description

813 Keypoints
SIFT: Application

• Image Alignment Example
SIFT: Application

- Image Correspondences
SIFT: Application

- Object Recognition
SIFT: Application

- Object Recognition
Work Flow

- **IMAGE**
  - SCALE-SPACE IMAGE REPRESENTATION
  - KEYPOINTS COMPUTATION BY DoG
  - CONTRAST-BASED EDGE FILTER
  - KEYPOINTS ORIENTATION
  - SIFT DESCRIPTOR
Scale-Space Representation

- SIFT describes an image or a portion of it by interest points (corners) whose detection requires a multi-scale approach:

Classic Multi-Scale Representation:

\[
P(f)_{n+1} = S(G_{\sigma} \ast P(f)_n) \\
P(f)_1 = f
\]

At each level of the pyramid the image is rescaled (sub-sampled) and smoothed by a Gaussian
Scale-Space Representation

• The SIFT scale-space image representation consists of a set of $N$ octaves $\{\Theta_1, \ldots, \Theta_N\}$ defined by two parameters $s$ and $\sigma$.

• Let $f$ be the input image. Each octave is an ordered set of $s + 3$ images such that

$$L(x, y, k^m \sigma) = G(x, y, k^m \sigma) \ast f_i(x, y), \quad k = \sqrt{2}$$

with $f_i$ i-th sub-sample of $f$ and $m = 0, 1, \ldots, s + 2$ and $i = 1, \ldots, N$. 
Suppose $s = 2$. Then each octave contains $s + 3$ images.
DoG for Corner Detection

- The keypoints extracted by SIFT are corners, i.e. discontinuity points of the gradient function:

- These are extracted by a DoG (difference of Gaussians).
DoG for Corner Detection

- The computation of the DoG in each octave is very fast and efficient.
- In fact the DoG is obtained by subtraction of subsequent images in the considered octave.

\[
\text{DoG}(x, y, \sigma, \sigma') = [G(x, y, \sigma) - G(x, y, \sigma')] * f(x, y)
\]
Keypoints Computation

• The keypoints are the extrema of the DoG functions, i.e. they are maximum or minimum of the function

\[ \text{DoG}(x, y, \sigma) \]

• These are computed by analyzing for each point a neighborhood 3 x 3 at the superior and inferior scale in the considered octave:
Keypoints Computation

• The location of the extrema is refined by considering a parabolic fit.
• Due to the re-iterated Gaussian filtering, many extrema exhibit small values of the contrast. These keypoints are not robust to noise and they are generally not relevant for the description of the image.
• Two filters are used to discard the keypoints with small contrast and the edges, that are not discriminative for the image.
• This step is achieved by considering the approximation of the DoG gradient by the Taylor polynom truncated at the first order.
SIFT descriptors

- Each keypoint is now codified as a triplet $(x, y, \sigma)$ whose gradient has magnitude and orientation given by

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

$$\theta(x, y, \sigma) = \arctan \frac{L(x, y + 1, \sigma) - L(x, y - 1, \sigma)}{L(x + 1, y, \sigma) - L(x - 1, y, \sigma)}$$

- A neighborhood $N$ around each keypoint is considered. The orientation of the gradient of the points in $N$ is represented by an histogram $H$ with 36 bins. The peak of $H$ is assigned to $(x, y, \sigma)$, so that the keypoint is described now by a vector $(x, y, \sigma, \theta)$, where $\theta$ is the orientation of the peak of $H$. If there are more peaks $\theta_1, \ldots, \theta_n$ more keypoints $(x, y, \sigma, \theta_1), \ldots, (x, y, \sigma, \theta_n)$ are generated.
SIFT descriptors

- Each keypoint is now codified as a triplet \((x, y, \sigma)\) whose gradient has magnitude and orientation given by

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

\[
\theta(x, y, \sigma) = \arctan \frac{L(x, y + 1, \sigma) - L(x, y - 1, \sigma)}{L(x + 1, y, \sigma) - L(x - 1, y, \sigma)}
\]

- A neighborhood \(N\) around each keypoint is considered. The orientation of the gradient of the points in \(N\) is represented by an histogram \(H\) with 36 bins. The peak of \(H\) is assigned to \((x, y, \sigma)\), so that the keypoint is described now by a vector \((x, y, \sigma, \theta)\), where \(\theta\) is the orientation of the peak of \(H\). If there are more peaks \(\theta_1, \ldots, \theta_n\) more keypoints \((x, y, \sigma, \theta_1), \ldots, (x, y, \sigma, \theta_n)\) are generated.
SIFT descriptors

- For each keypoint $P$ a squared region $R$ around $P$ is considered and partitioned in $4 \times 4$ parts. An histogram with 8 bins is used for representing the orientation of the points in each of the sub-regions of $R$.
- The final descriptor associated to $P$ is a vector that concatenate the histograms of the sub-regions of $R$.
- The descriptor vector has $(4 \times 4) \times 8 = 128$ entries.
Example: Image Description

Image Size: 640 x 480
[colums x rows]

981 Keypoints
Matching

• Lowe proposes a method for matching the keypoints.
• Let $R, Q$ be the lists with the keypoints of two images $I_1, I_2$. A keypoint $r$ of $R$ matches the keypoint $q$ of $Q$ if

\[
\| D(r) - D(q) \|^2 = \min_{z \in Q} \| D(r) - D(z) \|^2
\]

\[
\frac{\| D(r) - D(q) \|^2}{\min_{z \in Q - \{q\}} \| D(r) - D(z) \|^2} \geq T
\]

$T = 0.49$
References