Scale Invariant Feature Transform

Tony Lindeberg (2012), Scholarpedia, 7(5):10491. doi:10.4249/scholarpedia.10491 revision #153939 [link to/cite this article]

- **Prof. Tony Lindeberg**, KTH Royal Institute of Technology, Stockholm, Sweden

Scale Invariant Feature Transform (SIFT) is an image descriptor for image-based matching and recognition developed by David Lowe (1999, 2004). This descriptor as well as related image descriptors are used for a large number of purposes in computer vision related to point matching between different views of a 3-D scene and view-based object recognition. The SIFT descriptor is invariant to translations, rotations and scaling transformations in the image domain and robust to moderate perspective transformations and illumination variations. Experimentally, the SIFT descriptor has been proven to be very useful in practice for image matching and object recognition under real-world conditions.

In its original formulation, the SIFT descriptor comprised a method for detecting interest points from a grey-level image at which statistics of local gradient directions of image intensities were accumulated to give a summarizing description of the local image structures in a local neighbourhood around each interest point, with the intention that this descriptor should be used for matching corresponding interest points between different images. Later, the SIFT descriptor has also been applied at dense grids (dense SIFT) which have been shown to lead to better performance for tasks such as object categorization, texture classification, image alignment and biometrics . The SIFT descriptor has also been extended from grey-level to colour images and from 2-D spatial images to 2+1-D spatio-temporal video.

Contents

1 Interest point detection
- 1.1 Scale-invariant interest points from scale-space extrema
- 1.2 Interpolation
- 1.3 Suppression of interest point responses along edges

2 Image descriptor
- 2.1 Scale and orientation normalization
- 2.2 Weighted position-dependent histogram of local gradient directions
- 2.3 Contrast normalization
- 2.4 Theoretical explanation

3 Matching of local image descriptors
- 3.1 Nearest neighbour matching of local image descriptors
- 3.2 Best-bin-first approximation for selecting point matches
- 3.3 Affine Hough transform based evidence accumulation for object models

4 Extensions
- 4.1 PCA SIFT
- 4.2 Colour SIFT
- 4.3 SIFT-like image descriptors for spatio-temporal recognition
- 4.4 Dense SIFT

5 Related image descriptors
- 5.1 Receptive field histograms
- 5.2 Histograms of oriented gradients (HOG)
- 5.3 Gradient location and orientation histogram (GLOH)

5.4 Speeded up robust features (SURF)
5.5 Gauss-SIFT
6 Application areas
6.1 Multi-view matching
6.2 Object recognition
6.3 Object category classification
6.4 Robotics
7 Implementations
8 References
9 See also
10 Further reading
11 External links

Interest point detection

Scale-invariant interest points from scale-space extrema

The original SIFT descriptor (Lowe 1999, 2004) was computed from the image intensities around interesting locations in the image domain which can be referred to as interest points, alternatively key points. These interest points are obtained from scale-space extrema of differences-of-Gaussians (DoG) within a difference-of-Gaussians pyramid. The concept of difference-of-Gaussian bandpass pyramids was originally proposed by Burt and Adelson (1983) and by Crowley and Stern (1984).

A Gaussian pyramid is constructed from the input image by repeated smoothing and subsampling, and a difference-of-Gaussians pyramid is computed from the differences between the adjacent levels in the Gaussian pyramid. Then, interest points are obtained from the points at which the difference-of-Gaussians values assume extrema with respect to both the spatial coordinates in the image domain and the scale level in the pyramid.

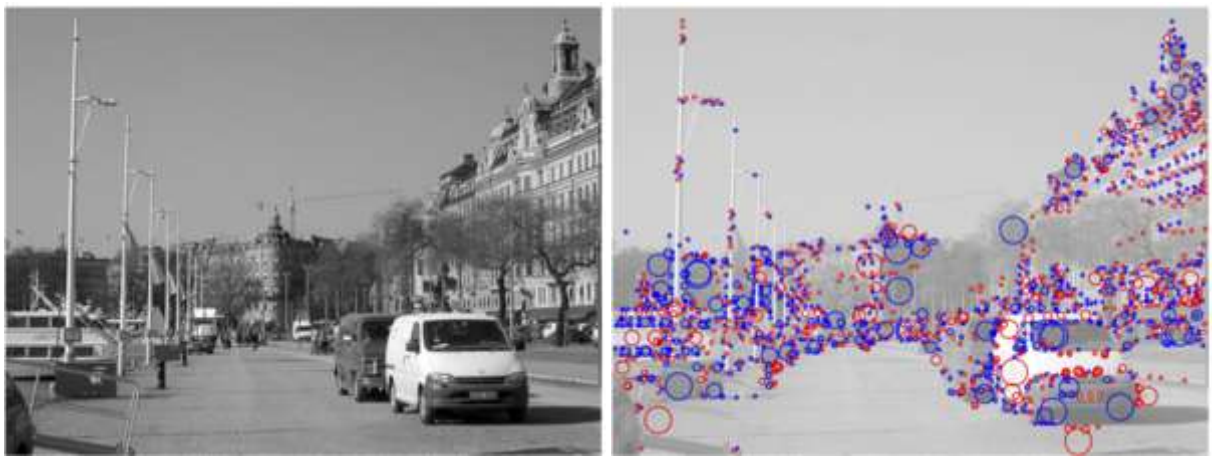

Figure 1: Scale-invariant interest points detected from a grey-level image using scale-space extrema of the Laplacian. The radii of the circles illustrate the selected detection scales of the interest points. Red circles indicate bright image features with $\nabla^2 L < 0$, whereas blue circles indicate dark image features with $\nabla^2 L > 0$.

This method for detecting interest points in the SIFT operator can be seen as a variation of a scale-adaptive blob detection method proposed by Lindeberg (1994, 1998), where blobs with associated scale levels are detected from *scale-space extrema* of the scale-normalized Laplacian. The scale-normalized Laplacian is normalized with respect to the scale level in scale-space and is defined as

$$\nabla^2_{norm} L(x, y; s) = s\,(L_{xx} + L_{yy}) = s\left(\frac{\partial^2 L}{\partial x^2} + \frac{\partial^2 L}{\partial y^2}\right) = s\,\nabla^2 (G(x, y; s) * f(x, y))$$

from smoothed image values $L(x, y; s)$ computed from the input image $f(x, y)$ by convolution with Gaussian kernels

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

of different widths $s = \sigma^2$, where σ denotes the standard deviation and s the variance of the Gaussian kernel. Then, the scale-space extrema are detected from the points $(x, y; s)$ in scale-space at which the scale-normalized Laplacian assumes local extrema with respect to space and scale. In a discrete setting, such comparisons are usually made in relation to all neighbours of a point in a $3 \times 3 \times 3$ neighbourhood over space and scale. The difference-of-Gaussians operator constitutes an approximation of the Laplacian operator

$$DOG(x, y; s) = L(x, y; s + \Delta s) - L(x, y; s) \approx \frac{\Delta s}{2}\, \nabla^2 L(x, y; s)$$

which by the implicit normalization of the differences-of-Gaussian responses, as obtained by a self-similar distribution of scale levels $\sigma_{i+1} = k\,\sigma_i$ used by Lowe, also constitutes an approximation of the scale-normalized Laplacian with $\Delta s\,\nabla^2 L = (k^2 - 1)\,t\,\nabla^2 L = (k^2 - 1)\,\nabla^2_{norm} L$, thus implying

$$DOG(x, y; s) \approx \frac{(k^2-1)}{2}\ \nabla^2_{norm} L(x, y; s).$$

It can be shown that this method for detecting interest points leads to *scale-invariance* in the sense that (i) the interest points are preserved under scaling transformations and (ii) the selected scale levels are transformed in accordance with the amount of scaling (Lindeberg 1998). Hence, the scale values obtained from these interest points can be used for normalizing local neighbourhoods with respect to scaling variations (Lindeberg 2013a, 2014) which is essential for the scale-invariant properties of the SIFT descriptor; see also (Lindeberg 2008) for an overview of the scale-space theory on which these image operations are based. The Laplacian operation is rotationally invariant. Therefore, (iii) these interest points will also be rotationally invariant.

The difference-of-Gaussians approach proposed by Lowe constitutes a computationally efficient way to compute approximations of such Laplacian interest points. Another way of detecting scale-space extrema of the Laplacian efficiently for real-time implementation has been presented by Lindeberg and Bretzner (2003) based on a hybrid pyramid. A closely related method for real-time scale selection has been developed by Crowley and Riff (2003).

Interpolation

Both the difference-of-Gaussians approach by Lowe and the Laplacian approach by Lindeberg and Bretzner involve the fitting of a quadratic polynomial to the magnitude values around each scale-space extremum to localize the scale-space extremum with a resolution higher than the sampling density over space and scale. This post-processing stage is in particular important to increase the accuracy of the scale estimates for the purpose of scale normalization.

Suppression of interest point responses along edges

In addition to responding to blob-like and corner-like image structures, the Laplacian operator may also lead to strong responses along edges. To suppress such points, which will be less useful for matching, Lowe (1999, 2004) formulated a criterion in terms of the ratio between the eigenvalues of the Hessian matrix

$$\mathcal{H}L = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix}$$

computed at the position and the scale of the interest point, which can be reformulated in terms of the trace and the determinant of the Hessian matrix to allow for more efficient computations

$$\frac{\det \mathcal{H}L}{\mathrm{trace}^2 \mathcal{H}L} = \frac{L_{xx} L_{yy} - L_{xy}^2}{(L_{xx} + L_{yy})^2} \geq \frac{r}{(r+1)^2}$$

where $r \geq 1$ denotes an upper limit on the permitted ratio between the larger and the smaller eigenvalues. (To avoid possible division by the determinant of the Hessian, which may approach zero along edges, the original statement by Lowe has been reformulated here to give a more well-conditioned numerical criterion.)

To suppress image features with low contrast, the interest points are usually also thresholded on the magnitude of the response.

Image descriptor

At each interest point as obtained above, an image descriptor is computed. The SIFT descriptor proposed by Lowe (1999, 2004) can be seen as a position-dependent histogram of local gradient directions around the interest point. To obtain scale invariance of the descriptor, the size of this local neighbourhood needs to be normalized in a scale-invariant manner. To obtain rotational invariance of the descriptor, a dominant orientation in this neighbourhood is determined from the orientations of the gradient vectors in this neighbourhood and is used for orienting the grid over which the position-dependent histogram is computed with respect to this dominant orientation to achieve rotational invariance.

Scale and orientation normalization

In the SIFT descriptor, the size estimate of an area around the interest point is determined as a constant times the detection scale s of the interest point, which can be motivated by the property of the scale selection mechanism in the interest point detector of returning a characteristic size estimate associated with each interest point (Lindeberg 1998).

To determine a preferred orientation estimate for the interest point, a local histogram of gradient directions is accumulated over a neighbourhood around the interest point with (i) the gradient directions computed from gradient vectors $\nabla L(x, y; s)$ at the detection scale s of the interest point and (ii) the area of the accumulation window proportional to the detection scale s. To find the dominant orientation, peaks are detected in this orientation histogram. To handle situations where there may be more than one dominant orientation around the interest point, multiple peaks are accepted if the height of secondary peaks is above 80 % of the height of the highest peak. In the case of multiple peaks, each peak is used for computing a new image descriptor for the corresponding orientation estimate.

When computing the orientation histogram, the increments are weighted by the gradient magnitude and also weighted by a Gaussian window function centered at the interest point and with its size proportional to the detection scale. To increase the accuracy of the orientation estimate, a rather dense sampling of the orientations is used, with 36 bins in the histogram. Moreover, the position of the peak is localized by local parabolic interpolation around the maximum point in the histogram.

Weighted position-dependent histogram of local gradient directions

Given these scale and orientation estimate for an interest point, a rectangular grid is laid out in the image domain, centered at the interest point, with its orientation determined by the main peak(s) in the histogram and with the spacing proportional to the detection scale of the interest point. From experiments, Lowe (1999, 2004) found that a 4×4 grid is often a good choice.

For each point on this grid, a local histogram of local gradient directions at the scale of the interest point

$$\arg\nabla L = \mathrm{atan2}(Ly, Lx)$$

is computed over a local neighbourhood around this grid point with the gradient directions quantized into 8 discrete directions. During the accumulation of the histograms, the increments in the histogram bins are weighted by the gradient magnitude

$$|\nabla L| = \sqrt{L_x^2 + L_y^2}$$

at each grid point to give stronger weights to image points where the gradient estimates can be expected to be more reliable. To give stronger weights to gradient orientations near the interest point, the entries in the histogram are also weighed by a Gaussian window function centered at the interest point and with its size proportional to the detection scale of the interest point. Taken together, the local histograms computed at all the 4×4 grid points and with 8 quantized directions lead to an image descriptor with $4 \times 4 \times 8 = 128$ dimensions for each interest point. This resulting image descriptor is referred to as the SIFT descriptor.

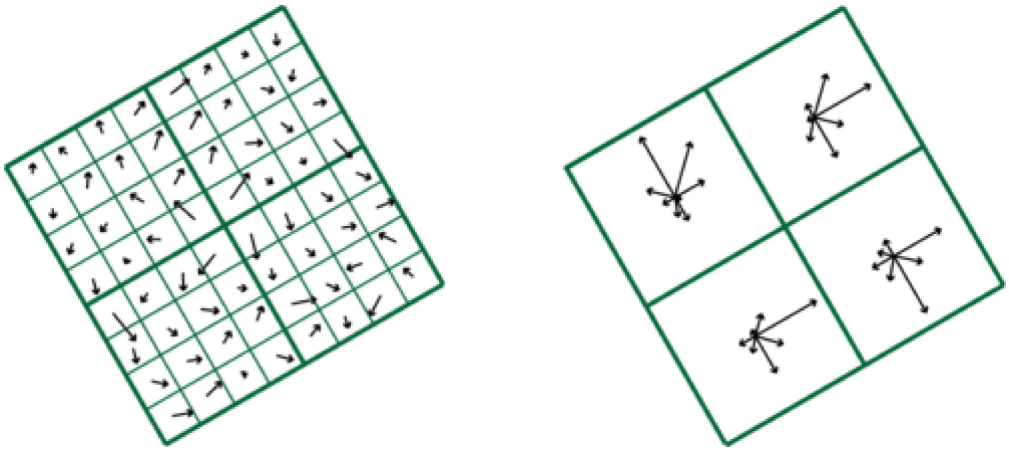

Figure 2: Illustration of how the SIFT descriptor is computed from sampled values of the gradient orientation and the gradient magnitude over a locally adapted grid around each interest point, with the scale factor determined from the detection scales of the interest point and the orientation determined from the dominant peak in a gradient orientation histogram around the interest point. This figure shows an image descriptor computed over a 2×2 whereas the SIFT descriptor is usually computed over a 4×4 grid.

To increase the accuracy of the local histograms, trilinear interpolation is used for distributing the weighted increments for the sampled image measurements into adjacent histogram bins. In other words, each entry in the bin is multiplied by an additional weight of $1 - d$, where d is the distance between the sample and the central position of the bin, expressed in units of the bin spacing in the histogram.

A closely related notion of orientation histograms ("zoning") has also been previously used for optical character recognition (Trier et al. 1996), although in that context formulated for locally binarized image patterns.

Contrast normalization

To obtain contrast invariance, the SIFT descriptor is normalized to unit sum. In this way, the weighted entries in the histogram will be invariant under local affine transformations of the image intensities around the interest point, which improves the robustness of the image descriptor under illumination variations.

To avoid local high contrast measurements from being given too excessive emphasis in the image descriptor, Lowe (1999, 2004) proposed a two-stage normalization, where the entries after a first-stage unit sum normalization are limited to not exceed 0.2, whereafter the modified image descriptor is normalized to unit sum again.

Theoretical explanation

The use of local position-dependent histograms of gradient directions for matching and recognition in SIFT constitutes a specific example of using image descriptors based on image measurements in terms of *receptive fields*. More generally, receptive fields in terms of Gaussian derivatives have been proposed as a canonical model for linear receptive fields in computer vision by Koenderink and van Doorn (1987, 1992) and Lindeberg (1994, 2011, 2013b). The pyramid representation previously proposed by Burt and Adelson (1983) and Crowley and Stern (1984) and used by Lowe can be seen as a numerical approximation of such Gaussian receptive fields. By the theoretical analysis in (Lindeberg 2013b) it can be shown that such receptive fields capture inherent characteristics of the reflectance patterns of surfaces of objects and do thus enable visual recognition.

The use of scale selection in the interest point detection step ensures that the interest points will be invariant under scaling transformations (Lindeberg 1998, 2013c, 2015). Specifically, the scale normalization of the image descriptor establishes a *local scale-invariant reference frame* which implies that also the image descriptors and the matching schemes based on those will be invariant under scaling transformations (Lindeberg 2013a, 2014). Thereby, image matching and object recognition based on such image features will have the ability to handle objects of different sizes as well as objects seen from different distances to the camera.

A more general set of scale-space interest point detectors for image-based matching and recognition and with better properties than Laplacian or difference-of-Gaussians interest points is presented in (Lindeberg 2015).

Matching of local image descriptors

Nearest neighbour matching of local image descriptors

Given a set of image descriptors computed from two different images, these image descriptors can be mutually matched by for each point finding the point in the other image domain that minimizes the Euclidean distance between the descriptors represented as 128-dimensional vectors. To suppress matches that could be regarded as possibly ambiguous, Lowe only accepted matches for which the ratio between the distances to the nearest and the next nearest points is less than 0.8.

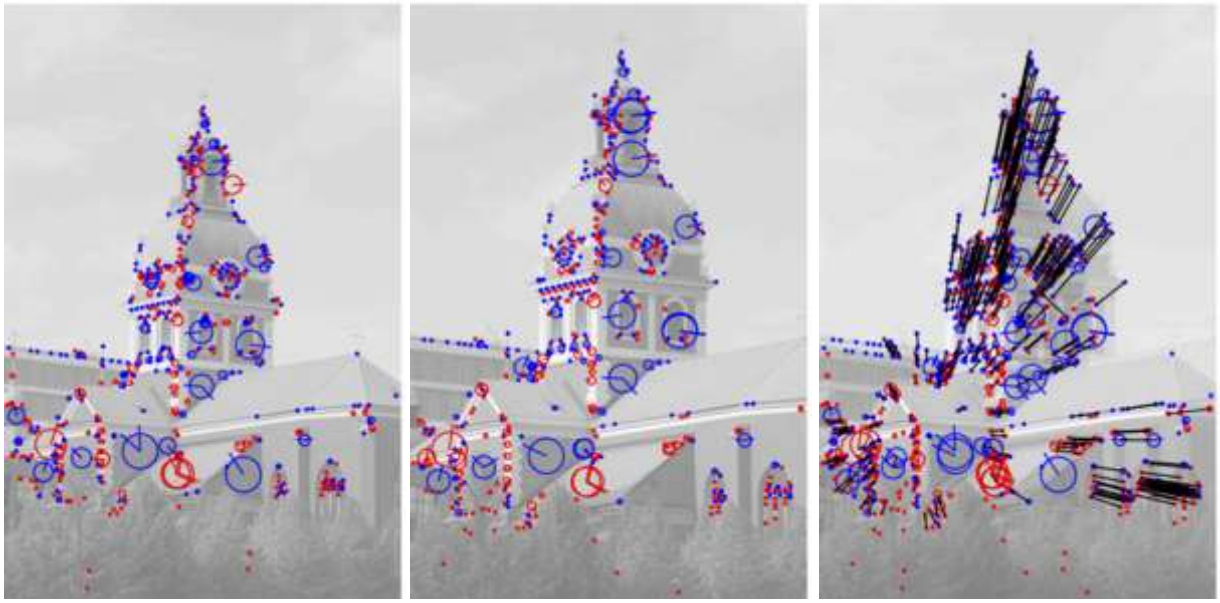

Figure 3: Interest points detected from two images of the same scene with the computed image matches drawn as black lines between corresponding interest points. The blue and red arrows at the centers of the circles illustrate the orientation estimates obtained from peaks in local orientation histograms around the interest points.

Best-bin-first approximation for selecting point matches

If we would apply the above mentioned nearest neighbour matching approach for recognizing an object against a large collection of objects in a database, such nearest neighbour matching would imply comparisons to all the image descriptors stored in the database. To speed up the resulting nearest-neighbour matching for larger data sets, Lowe (2004) applied an approximate best-bin-first (BBF) algorithm (Beis and Lowe 1997) that scales better with increasing numbers of image features. In later work (Muja and Lowe 2009), this approach has been furthered to hierarchical k-means trees and randomized k-d trees.

Affine Hough transform based evidence accumulation for object models

When applying the SIFT descriptor for object recognition, Lowe (2004) developed a Hough transform approach based on triples of image matches to accumulate evidence for objects as represented by sets of interest points with associated image descriptors.

When integrating the different components together, matching based on the SIFT descriptor quickly established itself as a state-of-the-art method for image-based matching and object recognition. In an experimental evaluation of the robustness of different image descriptors performed by Mikolajczyk and Schmid (2005), the SIFT descriptor was found to be more robust to image deformations than steerable filters, differential invariants, moment invariants, complex filters and cross-correlation of different types of interest points.

Extensions

PCA SIFT

Ke and Sukthankar (2004) proposed an alternative approach for defining local image descriptors, similar to the SIFT descriptor in the sense of detecting interest points with associated scale estimates from scale-space extrema and performing orientation normalization from peaks in a local orientation histogram, but different in terms of the actual image measurements underlying the image descriptors. Instead of computing gradient orientations, they first compute local maps of the *gradient magnitude*

$$|\nabla L| = \sqrt{L_x^2 + L_y^2}$$

over local patches around the interest points. To achieve scale invariance, the local patch for each interest point is warped to a scale normalized 39×39 reference frame common to all interest points. These local patches are then oriented with respect to a dominant image orientation to achieve rotational invariance. A normalization to unit sum is also performed to achieve local contrast invariance.

Then, these local gradient maps are projected to a lower-dimensional subspace (with 20 dimensions) using principal component analysis (PCA). Thus, given a specific interest point, the corresponding gradient map is computed, and after contrast normalization projected to the lower-dimensional subspace. Then, these local image descriptors are matched by minimizing the Euclidean distance. From experimental results, Ke and Sukthankar argued that PCA-SIFT is both faster and more distinctive that the regular SIFT descriptor.

Colour SIFT

Different ways of extending the SIFT descriptor from grey-level to colour images have been proposed by different authors. Bosch et al. (2006) computed SIFT descriptors over all three channels in the HSV colour space, resulting in a 3×128-dimensional HSV-SIFT image descriptor. Van de Weijer and Schmid (2006) concatenated the SIFT descriptor with either weighted hue or opponent angle histograms and evaluated the performance of the resulting composed image descriptors for computing point matches on different data sets.

Burghouts and Geusebroek (2009) defined a set of image descriptors that were based on a set of colour invariants given an illumination model. These colour invariants were in turn expressed in terms of the Gaussian colour model proposed by Koenderink. Specifically, they constructed a set of colour SIFT descriptors by replacing the grey-level gradient in the regular SIFT operator by different colour gradients that are invariant to different combinations of local intensity level, shadows, shading and highlights and evaluated these image descriptors on benchmark data sets. It was shown that one of the descriptors in this evaluation, referred to as C-colour-SIFT, performs better than the regular grey-level SIFT operator as well as better than the above mentioned colour SIFT descriptors based on either the HSV colour space or hue with regard to the problems of point matching and image category classification.

Van de Sande et al. (2010) performed a related study of the invariance properties of different colour representations under different types of illumination transformations, encompassing light intensity changes, light intensity shifts, light colour changes and light colour shifts. Specifically, the authors considered colour representations in terms of colour histograms, colour moments and colour invariants as well as different types of SIFT-like colour descriptors. Experimentally, they found that an OpponentSIFT descriptor based on colour-opponent channels lead to the best performance with regard to the problem of object category classification.

SIFT-like image descriptors for spatio-temporal recognition

The SIFT descriptor has been generalized from 2-D spatial images to 2+1-D spatio-temporal video by Laptev and Lindeberg (2004), by computing position-dependent histograms over local spatio-temporal neighbourhoods of either *spatio-temporal gradient vectors* (where t denotes time)

$$\nabla L = (L_x, L_y, L_t)$$

or *optic flow* (local image velocities)

$$(u, v)$$

computed at each position in the 2+1-D spatio-temporal domain.

Specifically, the image descriptors were computed at local spatio-temporal interest points detected using a spatio-temporal scale selection mechanism to allow for local adaptation and thus scale invariance with respect to both spatial scales and temporal scales. It was shown that this approach makes it possible to recognize human actions based on local spatio-temporal image descriptors in an analogous way as the local spatial SIFT descriptors allow for object recognition and object category classification.

To obtain invariance with respect to possibly unknown relative motions between the objects in the world and the observer, this approach was also combined with a *velocity adaptation* mechanism to adapt the spatio-temporal smoothing operations to local motions and was demonstrated to allow for recognition of spatio-temporal events in cluttered scenes (Laptev et al. 2007).

Dense SIFT

When applying the SIFT descriptor to tasks such as object category classification or scene classification, experimental evaluations show that better classification results are often obtained by computing the SIFT descriptor over dense grids in the image domain as opposed to at sparse interest points as obtained by an interest operator. A basic explanation for this is that a larger set of local image descriptors computed over a dense grid usually provide more information than corresponding descriptors evaluated at a much sparser set of image points.

This direction of development was initiated by Bosch et al. (2006, 2007) and has now established itself as a state-of-the-art approach for visual object category classification. When applied to object categorization tasks in practice, the computation of dense SIFT descriptors is usually accompanied with a clustering stage, where the individual SIFT

descriptors are reduced to a smaller vocabulary of visual words, which can then be combined with a bag-of-words model or related methods (Csurka et al. 2004, Lazebnik et al. 2006).

For the task of establishing image correspondences between initially unrelated different images of a 3-D object or a 3-D scene, the detection of sparse interest points is, however, still important an important pre-processing step to keep down the complexity when establishing image correspondences.

Related image descriptors

The SIFT descriptor can be seen as a special case of a more general class of image descriptors that are based on histograms of local receptive field responses.

Receptive field histograms

Swain and Ballard (1991) initiated a direction of research on histogram-based image descriptors by showing that reasonable performance of object recognition could be obtained by comparing RGB histograms of images of objects, thereby disregarding any spatial relationships between image features at different points. Schiele and Crowley (2000) extended this idea to histograms of either first-order partial derivatives or combinations of gradient magnitudes and Laplacian responses computed at multiple scales.

Linde and Lindeberg (2004, 2012) generalized this approach further to more general composed receptive field histograms constructed from different combinations of Gaussian derivatives or differential invariants computed from grey-level and colour-opponent cues up to order two and performed an extensive evaluation of such histogram descriptors with regard to object instance recognition and object category classification. Specifically, they proposed an efficient way of handling higher-dimensional histograms and introduced a set of composed complex cue histograms that lead to better performance than previously used primitive receptive field histograms of lower dimensionality.

Histograms of oriented gradients (HOG)

Inspired by the highly discriminatory property of local position-dependent gradient orientation histograms as used in the SIFT descriptor, Dalal and Triggs (2005) developed a closely related image descriptor defined from a set of gradient orientation histograms

$$\arg \nabla L = \operatorname{atan2}(Ly, Lx)$$

computed over a grid in the image domain. In contrast to SIFT descriptor, which is a local image descriptor, the resulting *histograms of oriented gradients* (HOG) descriptor is a *regional* image descriptor. In this sense, the HOG descriptor is closely related to the regional receptive field histograms that are defined over subregions in the image domain, with the differences that (i) the HOG operator includes a dependency on image positions by being composed of a set of smaller histograms defined over subregions and (ii) by being defined from gradient orientations instead of partial derivatives or differential invariants. In contrast to the SIFT descriptor, the HOG descriptor is, however, not normalized with respect to orientation. Therefore, the HOG descriptor is not rotationally invariant. The histograms in the HOG operator are, however, normalized with respect to image contrast.

Dalal and Triggs developed two versions of the HOG operator: one where the local histograms are computed over a rectangular grid (R-HOG) and one where the histograms are accumulated over a circular grid (C-HOG). Experimentally, the authors showed that the HOG descriptor allows for robust detection of humans in natural environments.

Gradient location and orientation histogram (GLOH)

Mikolajczyk and Schmid (2005) proposed an image descriptor referred to as GLOH, which is closely related to the original SIFT descriptor in the sense of also being a local position-dependent histogram of gradient orientations around an interest point. The GLOH descriptor does, however, differ in the respects of (i) being computed over a log-polar grid as opposed to a rectangular grid, (ii) using a larger number of 16 bins for quantizing the gradient directions as opposed to 8 bins as used in the regular SIFT descriptor, and (iii) using principal component analysis to reduce the dimensionality of the image descriptor. From their experimental results, the authors argued that the GLOH descriptor lead to better performance for point matching on structured scenes whereas the SIFT descriptor performed better on textured scenes.

Speeded up robust features (SURF)

The SURF descriptor proposed by Bay et al. (2006, 2008) is closely related to the SIFT descriptor in the sense that it is also a feature vector derived from receptive-field-like responses in a neighbourhood of an interest point. The SURF descriptor does, however, differ in the following respects:

- it is based on Haar wavelets instead of derivative approximations in an image pyramid,
- the interest points constitute approximations of scale-space extrema of the determinant of the Hessian instead of the Laplacian operator,
- the entries in the feature vector are computed as sums and absolute sums of first-order derivatives $\sum L_x, \sum |L_x|, \sum L_y, \sum |L_y|$ instead of histograms of coarsely quantized gradient directions.

Experimentally, the SURF operator leads to performance comparable to the SIFT operator. Due to the implementation in terms of Haar wavelets, the SURF operator is, however, faster.

Gauss-SIFT

Gauss-SIFT (Lindeberg 2015) is a pure image descriptor defined by performing all image measurements underlying the pure image descriptor in SIFT by Gaussian derivative responses as opposed to derivative approximations in an image pyramid as done in regular SIFT. In this way, discretization effects over space and scale can be reduced to a minimum allowing for potentially more accurate image descriptors.

In (Lindeberg 2015) such pure Gauss-SIFT image descriptors were combined with a set of generalized scale-space interest points including Laplacian of the Gaussian and determinant of the Hessian interest points. In an extensive experimental evaluation on a poster dataset comprising multiple views of 12 posters over scaling transformations up to a factor of 6 and viewing direction variations up to a slant angle of 45 degrees, it was shown that substantial increase in performance of image matching (higher efficiency scores and lower 1-precision scores) could be obtained by replacing Laplacian of Gaussian interest points by determinant of the Hessian interest points. Since difference-of-Gaussians interest points constitute a numerical approximation of Laplacian of the Gaussian interest points, this shows that a substantial increase in matching performance is possible by replacing the difference-of-Gaussians interest points in SIFT by determinant of the Hessian interest points.

A quantitative comparison between the Gauss-SIFT descriptor and a corresponding Gauss-SURF descriptor did also show that Gauss-SIFT does generally perform significantly better than Gauss-SURF for a large number of different scale-space interest point detectors. This study therefore shows that disregarding discretization effects the pure image descriptor in SIFT is significantly better than the pure image descriptor in SURF, whereas the underlying interest point detector in SURF, which can be seen as numerical approximation to scale-space extrema of the determinant of the Hessian, is significantly better than the underlying interest point detector is SIFT.

Application areas

The scale invariant feature transform (SIFT) with its related image descriptors in terms of histograms of receptive field-like image operations have opened up an area of research on image-based matching and recognition with numerous application areas. Being based on theoretically well-founded scale-space operations or approximations thereof, these approaches have been demonstrated to allow for robust computation of image features and image descriptors from real-world image data.

Multi-view matching

The SIFT descriptor with its associated matching methods can be used for establishing point matches between different views of a 3-D object or a scene. By combining such correspondences with *multi-view geometry* (Hartley and Zisserman 2004), 3-D models of objects and scenes can be constructed.

Similar methods for establishing multi-view correspondences can also be used for synthesizing novel views of a 3-D object/scene given a set of other views of the same object/scene (view interpolation) (Chen and Williams 1993, Zitnick et al. 2004, Liu et al. 2011) or for combining multiple partially overlapping images of the same scene into wider panoramas (Brown and Lowe 2007).

Object recognition

In his pioneering work on object recognition using the SIFT operator, Lowe demonstrated that robust and efficient recognition of objects in natural scenes can be performed based on collections of local image features. In close relation to this, a growing area of research has been developed concerning so-called bag of words methods and related methods for recognizing objects in real-world scenarios.

Besides the specific area of object recognition, these types of methods can also be used for related tasks such as visual search in image databases (Lew et al. 2006, Datta et al. 2008), human computer interaction based on visual input (Porta 2002, Jaimesa and Sebe 2008) or biometrics (Bicego et al. 2006, Li 2009, Wang et al. 2010).

Object category classification

Whereas the task of recognizing a previously seen object in a scene can be effectively addressed using the SIFT descriptor or the other closely related image descriptors described in this survey, the task of classifying previously unseen objects into object categories has turned out to be a harder problem. In the research to develop such methods, object categorization in terms of dense SIFT features (Bosch et al. 2007, Mutch and Lowe 2008) is as of 2012 still one of the better approaches.

Robotics

For a robot that moves in a natural environment, image correspondences in terms of SIFT features or related image descriptors can be used for tasks such as (i) localizing the robot with respect to a set of known references, (ii) mapping the surrounding from image data that are acquired as the robot moves around (See et al. 2005, Saeedi et al. 2006) or (iii) recognizing and establishing geometric relations to objects in the environment for robot manipulation (Siciliano and Khatib 2008).

Implementations

For efficient real-time processing, parallel implementations of SIFT have been developed for graphical processor units (GPUs) (Heymann et al. 2007) and field-programmable gate arrays (FPGAs) (Se et al. 2001, Se et al. 2004). For off-line processing, there are publically available implementations, such as VLFeat (open source) and David Lowe's SIFT demo program (Linux and Windows binaries) (see "External links" below).

References

- Lowe, David G. (1999). Object recognition from local scale-invariant features. *Proc. 7th International Conference on Computer Vision (ICCV'99)* (Corfu, Greece): 1150-1157. doi:10.1109/ICCV.1999.790410 (http://dx.doi.org/10.1109/ICCV.1999.790410) .
- Lowe, David G. (2004). Distinctive image features from scale-invariant key points. *International Journal of Computer Vision* 60(2): 91-110. doi:10.1023/B:VISI.0000029664.99615.94 (http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94) .
- Burt, Peter and Adelson, Ted (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications* 9(4): 532–540. doi:10.1109/tcom.1983.1095851 (http://dx.doi.org/10.1109/tcom.1983.1095851) .
- Crowley, James L. and Stern, Richard M. (1984). Fast computation of the difference of low pass transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6(2): 212-222. doi:10.1109/tpami.1984.4767504 (http://dx.doi.org/10.1109/tpami.1984.4767504) .
- Lindeberg, Tony (1994). Scale-Space Theory in Computer Vision. Kluwer/Springer, Boston.
- Lindeberg, Tony (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision* 30(2): 77-116. doi:10.1023/A:1008045108935 (http://dx.doi.org/10.1023/A:1008045108935) .
- Lindeberg, Tony (2013a). Invariance of visual operations at the level of receptive fields. *PLOS ONE* 8(7): e66990:1-33. doi:10.1371/journal.pone.0066990 (http://dx.doi.org/10.1371/journal.pone.0066990) .
- Lindeberg, Tony (2014). Scale selection. *Computer Vision: A Reference guide* (K. Ikeuchi, ed.), Springer: 701-713. doi:10.1007/978-0-387-31439-6_242 (http://dx.doi.org/10.1007/978-0-387-31439-6_242) .
- Lindeberg, Tony (2008). Scale-space. *Encyclopedia of Computer Science and Engineering* John Wiley and Sons: IV:2495--2504. doi:10.1002/9780470050118.ecse609 (http://dx.doi.org/10.1002/9780470050118.ecse609) .
- Lindeberg, Tony and Bretzner, Lars (2003). Real-time scale selection in hybrid multi-scale representations. *Proc. Scale-Space'03* Springer Lecture Notes in Computer Science 2695: 148-163. doi:10.1007/3-540-44935-3_11 (http://dx.doi.org/10.1007/3-540-44935-3_11) .
- Crowley, James L. and Riff, Olivier (2003). Fast computation of scale normalised Gaussian receptive fields. *Proc. Scale-Space'03* Springer Lecture Notes in Computer Science 2695: 584-598. doi:10.1007/3-540-44935-3_41 (http://dx.doi.org/10.1007/3-540-44935-3_41) .
- Trier, Øivind Due; Jain, Anil K. and Taxt, Torfinn (1996). Feature extraction methods for character recognition - A survey. *Pattern Recognition* 29(4): 641-662. doi:10.1016/0031-3203(95)00118-2 (http://dx.doi.org/10.1016/0031-3203(95)00118-2) .
- Koenderink, Jan and van Doorn, Andrea (1987). Representation of local geometry in the visual system. *Biological Cybernetics* 53: 383-396. doi:10.1007/BF00318371 (http://dx.doi.org/10.1007/BF00318371) .
- Koenderink, Jan and van Doorn, Andrea (1992). Generic neighbourhood operations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14(6): 597-605. doi:10.1109/34.141551 (http://dx.doi.org/10.1109/34.141551) .
- Lindeberg, Tony (2011). Generalized Gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space. *Journal of Mathematical Imaging and Vision* 40(1): 36-81. doi:10.1007/s10851-010-0242-2 (http://dx.doi.org/10.1007/s10851-010-0242-2) .
- Lindeberg, Tony (2013b). A computational theory of visual receptive fields. *Biological Cybernetics* 107(6): 589-645. doi:10.1007/s00422-013-0569-z (http://dx.doi.org/10.1007/s00422-013-0569-z) .
- Lindeberg, Tony (2013c). Scale selection properties of generalized scale-space interest points. *Journal of Mathematical Imaging and Vision* 46(2): 177-210. doi:10.1007/s10851-012-0378-3 (http://dx.doi.org/10.1007/s10851-012-0378-3) .

- Lindeberg, Tony (2015). Image matching using generalized scale-space interest points. *Journal of Mathematical Imaging and Vision* 52(1): 3-36. doi:10.1007/s10851-014-0541-0 (http://dx.doi.org/10.1007/s10851-014-0541-0) .
- Beis, Jeffrey S. and Lowe, David G. (1997). Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. *Proc. Conference on Computer Vision and Pattern Recognition (CVPR'97)* (Puerto Rico): 1000-1006.
- Muja, Marius and Lowe, David G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. *Proc. International Conference on Computer Vision Theory and Applications (VISAPP'09)* (Lisbon, Portugal): 331--340.
- Mikolajczyk, Krystian and Schmid, Cordelia (2005). A performance evaluation of local descriptors. *International IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(19): 1615--1630. doi:10.1109/tpami.2005.188 (http://dx.doi.org/10.1109/tpami.2005.188) .
- Ke, Yan and Sukthankar, Rahul (2004). PCA-SIFT: A more distinctive representation for local image descriptors. *Proc. Computer Vision and Pattern Recognition (CVPR'04)* (Pittsburgh, PA): II:506-513.
- van de Weijer, Joost and Schmid, Cordelia (2006). Coloring local feature extraction. *Proc. 9th European Conference on Computer Vision (ECCV'06)* Springer Lecture Notes in Computer Science 3952: 334-348. doi:10.1007/11744047_26 (http://dx.doi.org/10.1007/11744047_26) .
- Burghouts, Gertjan J. and Geusebroek, Jan-Mark (2009). Performance evaluation of local colour invariants. *Computer Vision and Image Understanding* 113: 48-62. doi:10.1016/j.cviu.2008.07.003 (http://dx.doi.org/10.1016/j.cviu.2008.07.003) .
- van de Sande, Koen; Gevers, Theo and Jan-Snoek, Cees G. M. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9): 1582-1596. doi:10.1109/tpami.2009.154 (http://dx.doi.org/10.1109/tpami.2009.154) .
- Laptev, Ivan and Lindeberg, Tony (2004). Local descriptors for spatio-temporal recognition. *ECCV'04 Workshop on Spatial Coherence for Visual Motion Analysis, (Prague, Czech Republic), May 2004* Springer Lecture Notes in Computer Science 3667: 91-103. doi:10.1007/11676959_8 (http://dx.doi.org/10.1007/11676959_8) .
- Laptev, Ivan; Caputo, Barbara; Schuldt, Christian and Lindeberg, Tony (2007). Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding* 108: 207-229. doi:10.1016/j.cviu.2006.11.023 (http://dx.doi.org/10.1016/j.cviu.2006.11.023) .
- Bosch, Anna; Zisserman, Andrew and Munoz, Xavier (2006). Scene classification via pLSA. *Proc. 9th European Conference on Computer Vision (ECCV'06)* Springer Lecture Notes in Computer Science 3954: 517~530.
- Bosch, Anna; Zisserman, Andrew and Munoz, Xavier (2007). Image classification using random forests and ferns. *Proc. 11th International Conference on Computer Vision (ICCV'07)* (Rio de Janeiro, Brazil): 1-8.
- Csurka, Gabriella; Dance, Christopher R.; Fan, Lixin; Willamowski, Jutta and Bray, Cédric (2004). Visual categorization with bags of keypoints. *Proc. ECCV'04 International Workshop on Statistical Learning in Computer Vision* (Prague, Czech Republic): 1-22.
- Lazebnik, Svetlana; Schmid, Cordelia and Ponce, Jean (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories *Proc. IEEE Conference on Computer Vision & Pattern Recognition (CVPR'06)* (New York): 2169-2178.
- Swain, Michael J. and Ballard, Dana H. (1991). Color indexing. *International Journal of Computer Vision* 7(1): 11-32. doi:10.1007/bf00130487 (http://dx.doi.org/10.1007/bf00130487) .
- Schiele, Bernt and Crowley, James L. (2000). Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision* 26(1): 31-50. doi:10.1007/bfb0015571 (http://dx.doi.org/10.1007/bfb0015571) .

- Linde, Oskar and Lindeberg, Tony (2004). Object recognition using composed receptive field histograms of higher dimensionality. *Proc 17th International Conference on Pattern Recognition (ICPR'04)* (Cambridge, U.K.): I:1-6. doi:10.1109/ICPR.2004.1333965 (http://dx.doi.org/10.1109/ICPR.2004.1333965) .
- Linde, Oskar and Lindeberg, Tony (2012). Composed complex-cue histograms: An investigation of the information content in receptive field based image descriptors for object recognition. *Computer Vision and Image Understanding* 116: 538-560. doi:10.1016/j.cviu.2011.12.003 (http://dx.doi.org/10.1016/j.cviu.2011.12.003) .
- Dalal, Nadal and Triggs, Bill (2005). Histograms of oriented gradients for human detection. *Proc. Computer Vision and Pattern Recognition (CVPR'05)* (San Diego, CA): I:886-893.
- Bay, Herbert; Tuytelaars, Tinne and van Gool, Luc (2006). SURF: Speeded up robust features. *Proc. 9th European Conference on Computer Vision (ECCV'06)* Springer Lecture Notes in Computer Science 3951: 404-417. doi:10.1007/11744023_32 (http://dx.doi.org/10.1007/11744023_32) .
- Bay, Herbert; Ess, Andreas; Tuytelaars, Tinne and van Gool, Luc (2008). SURF: Speeded up robust features. *Computer Vision and Image Understanding* 110(3): 346-359. doi:10.1016/j.cviu.2007.09.014 (http://dx.doi.org/10.1016/j.cviu.2007.09.014) .
- Hartley, Richard and Zisserman, Andrew (2004). Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge U.K..
- Chen, Shenchang Eric and Williams, Lance (1993). View interpolation for image synthesis. *Proc. ACM SIGGRAPH 1993* (Anahem, CA): 279--288.
- Zitnick, C. Lawrence; Kang, Sing Bing; Uyttendaele, Matthew; Winder, Simon and Szeliski, Richard (2004). High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics (TOG): Proc. ACM SIGGRAPH 2004* 23(3)(1): 600-608. doi:10.1145/1015706.1015766 (http://dx.doi.org/10.1145/1015706.1015766) .
- Ce, Liu; Jenny, Yuen and Antonio, Torralba (2011). SIFT flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(5): 978-994. doi:10.1007/978-3-540-88690-7_3 (http://dx.doi.org/10.1007/978-3-540-88690-7_3) .
- Brown, Matthew and Lowe, David G. (2007). Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision* 74(1): 1-19. doi:10.1007/s11263-006-0002-3 (http://dx.doi.org/10.1007/s11263-006-0002-3) .
- Lew, Michael S.; Sebe, Nicu; Djeraba, Chabane and Jain, Ramesh (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)* 2(1): 1-19. doi:10.1145/1126004.1126005 (http://dx.doi.org/10.1145/1126004.1126005) .
- Datta, Ritendra; Joshi, Dhiraj; Li, Jia and Wang, James Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys* 40(2): 1-60. doi:10.1145/1348246.1348248 (http://dx.doi.org/10.1145/1348246.1348248) .
- Porta, Marcus (2002). Vision-based user interfaces: Methods and applications. *International Journal of Human-Computer Studies* 57(1): 27–73. doi:10.1006/ijhc.2002.1012 (http://dx.doi.org/10.1006/ijhc.2002.1012) .
- Jaimesa, Alejandro and Sebe, Nicuj (2007). Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding* 108(1-2): 116-134. doi:10.1016/j.cviu.2006.10.019 (http://dx.doi.org/10.1016/j.cviu.2006.10.019) .
- Manuele, Bicego; Andrea, Lagorio; Enrico, Grosso and Massimo, Tistarelli (2006). On the use of SIFT Features for face authentication. *Proc. Computer Vision and Pattern Recognition Workshop (CVPRW'06)* (New

York,NY): 35-35.

- Stan, Li (2009). Encyclopedia of Biometrics. Springer, Boston.
- J.-G., Wang; J., Li; W.-Y., Yau and E., Sung (2010). Boosting dense SIFT descriptors and shape contexts of face images for gender recognition. *Proc. Computer Vision and Pattern Recognition Workshop (CVPRW'10)* (San Francisco,CA): 1-8.
- Mutch, Jim and Lowe, David G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision* 80(1): 45-57. doi:10.1007/s11263-007-0118-0 (http://dx.doi.org/10.1007/s11263-007-0118-0) .
- Se, Stephen; Lowe, David G. and Little, James J. (2005). Vision-based global localization and mapping for mobile robots. *IEEE Transaction on Robotics* 21(3): 364-375. doi:10.1109/tro.2004.839228 (http://dx.doi.org/10.1109/tro.2004.839228) .
- Saeedi, Parvaneh; Lawrence, Peter D. and Lowe, David G. (2006). Vision-based 3D trajectory tracking for unknown environments. *IEEE Transaction on Robotics* 22(1): 119-136. doi:10.1109/tro.2005.858856 (http://dx.doi.org/10.1109/tro.2005.858856) .
- Siciliano, Bruno and Khatib, Oussama (2008). Springer Handbook of Robotics. Springer, Boston.
- Heymann, S.; M"uller, K.; Smolic, A.; Fröhlich, B. and Wiegand, T. (2007). SIFT implementation and optimization for general-purpose GPU. *Proc. 15th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG'07)* Plzen, Czech Republic: 317-322.
- Se, Stephen; Lowe, David and Little, Jim (2001). Vision-based mobile robot localization and mapping using scale-invariant features. *Proc. IEEE International Conference on Robotics and Automation (ICRA'01)* volume 2: 2051-2058.
- Se, Stephen; Ng, Ho-kong; Jasiobedzki, Piotr and Moyung, Tai-jing (2004). Vision-based modeling and localization for planetary exploration rovers. *Proc 55th International Astronautical Congress (IAC'04)* Vancouver: Canada.

See also

- Receptive field

Further reading

- Lindeberg, Tony (1994). Scale-space theory: A basic tool for analysing structures at different scales. *J. of Applied Statistics* 21(2): 224-270. doi:10.1080/757582976 (http://dx.doi.org/10.1080/757582976) .
- Lindeberg, Tony (2008). Scale-space. *Encyclopedia of Computer Science and Engineering* John Wiley and Sons: IV:2495-2504. doi:/10.1002/9780470050118.ecse609 (http://dx.doi.org//10.1002/9780470050118.ecse609) .
- Tuytelaars, Tinne and Mikolajczyk, Krystian (2008). Local invariant feature detectors: A survey. Foundations and Trends® in Computer Graphics and Vision volume 3, number 3, Now Publishers Inc.
- Lindeberg, Tony (2013a). Invariance of visual operations at the level of receptive fields. *PLOS ONE* 8(7): e66990:1-33. doi:10.1371/journal.pone.0066990 (http://dx.doi.org/10.1371/journal.pone.0066990) .
- Lindeberg, Tony (2013b). A computational theory of visual receptive fields. *Biological Cybernetics* 107(6): 589-645. doi:10.1007/s00422-013-0569-z (http://dx.doi.org/10.1007/s00422-013-0569-z) .

External links

- Wikipedia: Scale-invariant feature transform (http://en.wikipedia.org/wiki/Scale-invariant_feature_transform)

- Wikipedia: Scale-space (http://en.wikipedia.org/wiki/Scale_space)
- VLFeat open source SIFT implementation at http://www.vlfeat.org/ (http://www.vlfeat.org)
- David Lowe's demo program at http://www.cs.ubc.ca/~lowe/keypoints/ (http://www.cs.ubc.ca/~lowe/keypoints/)

Sponsored by: Jian-Gang Wang, Institute for Infocomm Research, Singapore

Reviewed by (http://www.scholarpedia.org/w/index.php?title=Scale_Invariant_Feature_Transform&oldid=125063) : Xiaoyang Tan, Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics

Reviewed by (http://www.scholarpedia.org/w/index.php?title=Scale_Invariant_Feature_Transform&oldid=0) : Jian-Gang Wang, Institute for Infocomm Research, Singapore

Accepted on: 2012-05-08 03:39:16 GMT (http://www.scholarpedia.org/w/index.php?title=Scale_Invariant_Feature_Transform&oldid=125156)

Categories: Vision | Pattern Recognition | Computational Intelligence | Machine learning | Robotics | Biometrics