In: *Encyclopedia of Computer Science and Engineering* (Benjamin Wah, ed), John Wiley and Sons, Volume IV, pages 2495–2504, Hoboken, New Jersey, 2009. dx.doi.org/10.1002/9780470050118.ecse609 (Sep 2008)

Scale-space

Tony Lindeberg
School of Computer Science and Communication,
KTH (Royal Institute of Technology), SE-100 44 Stockholm, Sweden

Abstract

Scale-space theory is a framework for multiscale image representation, which has been developed by the computer vision community with complementary motivations from physics and biologic vision. The idea is to handle the multiscale nature of real-world objects, which implies that objects may be perceived in different ways depending on the scale of observation. If one aims to develop automatic algorithms for interpreting images of unknown scenes, there is no way to know a priori what scales are relevant. Hence, the only reasonable approach is to consider representations at all scales simultaneously. From axiomatic derivations is has been shown that given the requirement that coarse-scale representations should correspond to true simplifications of fine scale structures, convolution with Gaussian kernels and Gaussian derivatives is singled out as a canonical class of image operators for the earliest stages of visual processing. These image operators can be used as basis to solve a large variety of visual tasks, including feature detection, feature classification, stereo matching, motion descriptors, shape cues, and image-based recognition. By complementing scale-space representation with a module for automatic scale selection based on the maximization of normalized derivatives over scales, early visual modules can be made scale invariant. In this way, visual modules can adapt automatically to the unknown scale variations that may occur because of objects and substructures of varying physical size as well as objects with varying distances to the camera. An interesting similarity to biologic vision is that the scale-space operators resemble closely receptive field profiles registered in neurophysiologic studies of the mammalian retina and visual cortex.

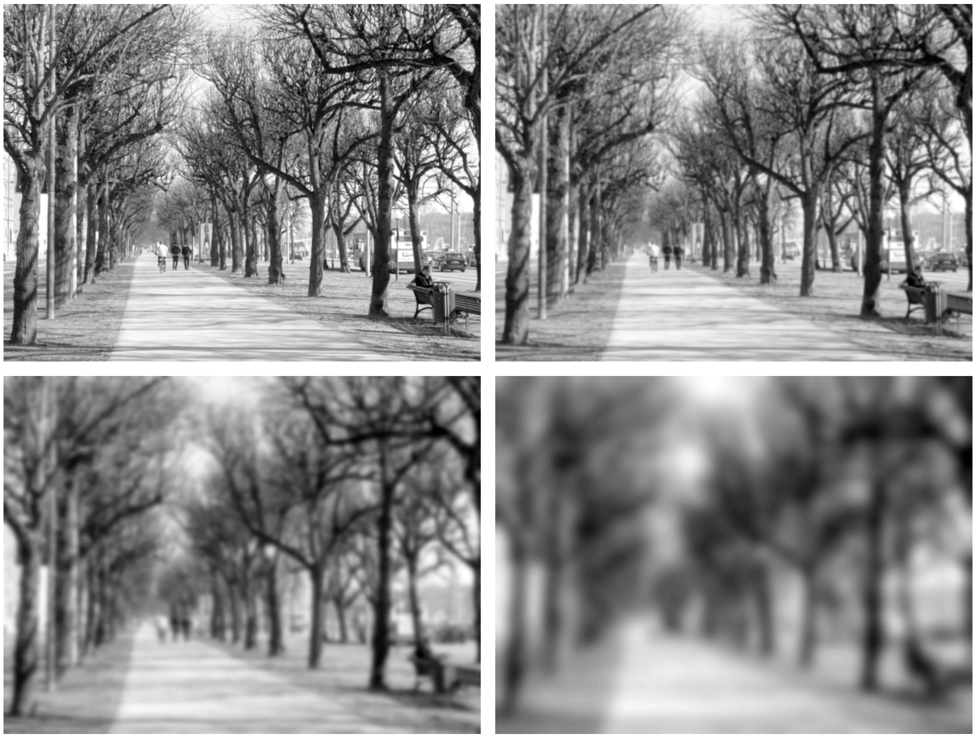

Figure 1: (top left) A grey-level image of size 560 × 420 pixels. (top right)–(bottom right) Scale-space representations computed at scale levels $t = 1$, 8 and 64 (in pixel units).

The need for multi-scale representation of image data

An inherent property of real-world objects is that they only exist as meaningful entities over certain ranges of scale. A simple example is the concept of a branch of a tree, which makes sense only at a scale from, say, a few centimetres to at most a few metres, It is meaningless to discuss the tree concept at the nanometre or kilometre level. At those scales, it is more relevant to talk about the molecules that form the leaves of the tree or the forest in which the tree grows. When observing such real-world objects with a camera or an eye, there is an addition scale problem due to perspective effects. A nearby object will appear larger in the image space than a distant object, although the two objects may have the same size in the world. These facts, that objects in the world appear in different ways depending on the scale of observation and in addition may undergo scale changes during an imaging process, have important implications if one aims at describing them. It shows that the notion of *scale* is fundamental to the understanding of both natural and artificial perception.

In computer vision and image analysis, the notion of scale is essential to designing methods for deriving information from images and multi-dimensional signals. To be able to extract any information from image data, one obviously has to interact with the data in some way, using some operator or measurement probe. The type of information that can be obtained is largely determined by the relationship between the size of the actual structures in the data and the size (resolution) of the operators (probes). Some of the very fundamental problems in computer vision and image processing concern *what* operators to use, *where* to apply them and *how large* they should be. If these problems are not appropriately addressed, then the task of interpreting the operator response can be very hard. Notably, the scale information required to view the image data at an appropriate scale may in many cases be *a priori* unknown.

The idea behind a scale-space representation of image data is that in the absence of any prior information about what scales are appropriate for a given visual task, the only reasonable approach is to represent the data at multiple scales. Taken to the limit, a scale-space representation furthermore considers representations at *all* scales simultaneously. Thus, given any input image, this image embedded into a one-parameter family of derived signals, in which fine-scale structures are progressively suppressed. When constructing such a multi-scale representation, a crucial requirement is that the coarse-scale representations should constitute simplifications of corresponding structures at finer scales—they should not be accidental phenomena created by the smoothing method intended to suppress fine scale structures. This idea has been formalized in a variety of ways by different authors, and a noteworthy coincidence is that similar conclusions can be obtained from several different starting points. A fundamental result of scale-space theory is that if rather general conditions are imposed on the types of computations that are to be performed in the earliest stages of visual processing, then convolution by the Gaussian kernel and its derivatives provide a canonical class of image operators with unique properties. The requirements (scale-space axioms; see below) that specify the uniqueness are essentially linearity and spatial shift invariance, combined with different ways of formalizing the notion that new structures should not be created in the transformation from fine to coarse scales.

In summary, for any two-dimensional signal $f\colon \mathbb{R}^2 \to \mathbb{R}$, its *scale-space representation* $L\colon \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$ is defined by [37, 14, 22, 34, 7, 30]

$$L(x, y;\ t) = \int_{(\xi,\eta)\in\mathbb{R}^2} f(x-\xi, y-\eta)\, g(\xi, \eta;\ t)\, d\xi\, d\eta \tag{1}$$

where $g\colon \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$ denotes the Gaussian kernel

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

and the variance $t = \sigma^2$ of this kernel is referred to as the *scale parameter*. Equivalently, the scale-space family can be obtained as the solution of the (linear) diffusion equation

$$\partial_t L = \frac{1}{2}\nabla^2 L \tag{3}$$

with initial condition $L(\cdot, \cdot;\ t) = f$. Then, based on this representation, *scale-space derivatives* at any scale t can be computed either by differentiating the scale-space directly or by convolving the original image with Gaussian derivative kernels:

$$L_{x^\alpha y^\beta}(\cdot, \cdot;\ t) = \partial_{x^\alpha y^\beta} L(\cdot, \cdot;\ t) = (\partial_{x^\alpha y^\beta} g(\cdot, \cdot;\ t)) * f(\cdot, \cdot). \tag{4}$$

Since scale-space derivatives can also be computed by convolving the original image with Gaussian derivative operators $g_{x^\alpha y^\beta}(\cdot, \cdot;\ t)$, they are also referred to as *Gaussian derivatives*. This way of defining derivatives in scale-space makes the inherently ill-posed problem of computing image derivatives well-posed, with a close connection to generalized functions.

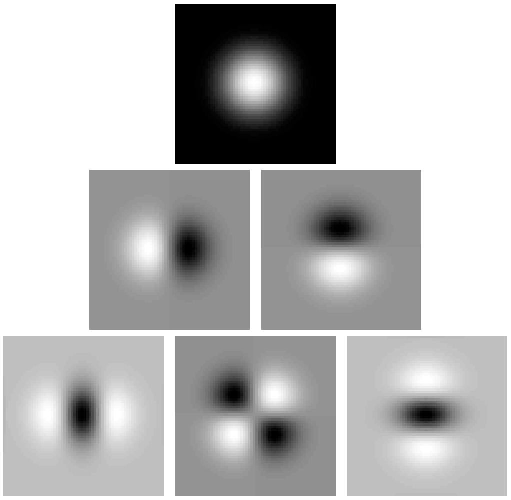

Figure 2: The Gaussian kernel and its derivatives up to order two in the 2-D case.

For simplicity, we shall here restrict ourselves to 2-D images. With appropriate generalizations or restrictions, however, most of these concepts apply in arbitrary dimensions.

Feature detection at a given scale in scale-space

The set of scale-space derivatives up to order N at a given image point and a given scale is referred to as the *N-jet* [15, 16] and corresponds to a truncated Taylor expansion of a locally smoothed image patch. These derivatives together constitutes a basic type of feature within the scale-space framework and provide a compact characterization of the local image structure around the image point at that scale. For $N = 2$, the 2-jet at a single scale contains the partial derivatives

$$(L_x, L_y, L_{xx}, L_{xy}, L_{yy}) \tag{5}$$

and directional filters in any direction $(\cos\varphi, \sin\varphi)$ can be obtained from

$$\partial_\varphi L = \cos\varphi L_x + \sin\varphi L_y \quad \text{and} \quad \partial_{\varphi\varphi} L = \cos^2\varphi L_{xx} + 2\cos\varphi\sin\varphi L_{xy} + \sin^2\varphi L_{yy}. \tag{6}$$

From the five components in the 2-jet, four differential invariants can be constructed, which are invariant to local rotations; the gradient magnitude $|\nabla L|$, the Laplacian $\nabla^2 L$, the determinant of the Hessian $\det \mathcal{H}L$ and the rescaled level curve curvature $\tilde{\kappa}(L)$:

$$\begin{cases} |\nabla L|^2 = L_x^2 + L_y^2, \\ \nabla^2 L = L_{xx} + L_{yy}, \\ \det \mathcal{H}L = L_{xx}L_{yy} - L_{xy}^2, \\ \tilde{\kappa}(L) = L_x^2 L_{yy} + L_y^2 L_{xx} - 2L_x L_y L_{xy}. \end{cases} \tag{7}$$

A theoretically well-founded approach to feature detection is to use rotationally variant descriptors such as the N-jet, directional filter banks or rotationally invariant differential invariants as primitives for expressing visual modules. For example, with v denoting the gradient direction $(L_x, L_y)^T$, a differential geometric formulation of *edge detection* at a given scale can be expressed from the image points for which the second-order directional derivative in the gradient direction L_{vv} is zero and the third-order directional derivative L_{vvv} is negative:

$$\begin{cases} \tilde{L}_{vv} = L_x^2 \, L_{xx} + 2 \, L_x \, L_y \, L_{xy} + L_y^2 \, L_{yy} = 0, \\ \tilde{L}_{vvv} = L_x^3 \, L_{xxx} + 3 \, L_x^2 \, L_y \, L_{xxy} + 3 \, L_x \, L_y^2 \, L_{xyy} + L_y^3 \, L_{yyy} < 0. \end{cases} \tag{8}$$

A single-scale *blob detector* that responds to bright and dark blobs can be expressed from the minima and the maxima of the Laplacian response $\nabla^2 L$. An affine covariant blob detector that also responds to saddles can be expressed from the maxima and the minima of the determinant of the Hessian $\det \mathcal{H}L$. A straight-forward and affine covariant *corner detector* can be expressed from the maxima and minima of the rescaled level curve curvature $\tilde{\kappa}(L)$. With p denoting the main eigendirection of the Hessian matrix, which is parallel to the vector

$$(\cos\varphi, \sin\varphi) \sim \left(\sqrt{1 + \frac{L_{xx} - L_{yy}}{\sqrt{(L_{xx} - L_{yy})^2 + 4L_{xy}^2}}}, \operatorname{sign}(L_{xy}) \sqrt{1 - \frac{L_{xx} - L_{yy}}{\sqrt{(L_{xx} - L_{yy})^2 + 4L_{xy}^2}}} \right), \tag{9}$$

a differential geometric *ridge detector* at a fixed scale can be expressed from the zero-crossings of the first derivative L_p in this direction for which the second-order directional derivative L_{pp} is negative and in addition $|L_{pp}| \geq |L_{qq}|$. Similarly, valleys can be extracted from the zero-crossings of L_q that satisfy $L_{qq} \geq 0$ and $|L_{qq}| \geq |L_{pp}|$.

Feature classification and image matching from the multi-scale N-jet

By combining N-jet representations at multiple scales, usually with the scale levels distributed by ratios of two when the scale parameter is measured in units of the standard deviation $\sigma = \sqrt{t}$ of the Gaussian, we obtain a *multi-scale N-jet vector*. This descriptor is useful for a variety of different tasks. For example, the task of *texture classification* can be expressed as a classification and/or clustering problem on the multi-scale N-jet over regions in the image [21]. Methods for *stereo matching* can be formulated in terms of comparisons of local N-jets [13], either in terms of explicit search or coarse-to-fine schemes based on differential corrections within the support region of the multi-scale N-jet. Moreover, straightforward (rotationally dependent) methods for *image-based object recognition* can expressed in terms of vectors or (global or regional) histograms of multi-scale N-jets [32, 31, 27, 2]. Methods for rotationally invariant image-based recognition can formulated in terms of histograms of multi-scale vectors of differential invariants.

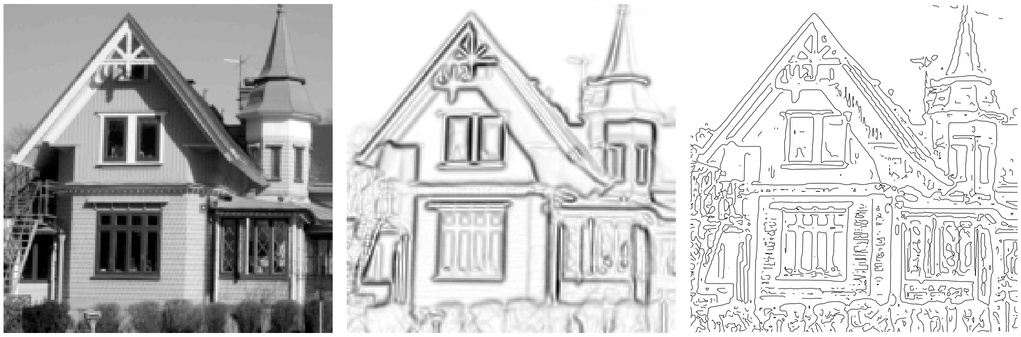

Figure 3: *Edge detection:* (left) A grey-level image of size 180 × 180 pixels. (middle) The negative value of the gradient magnitude $|\nabla L|$ computed at $t = 1$. (right) Differential geometric edges at $t = 1$ with a complementary low threshold $\sqrt{t}|\nabla L| \geq 1$ on the gradient magnitude.

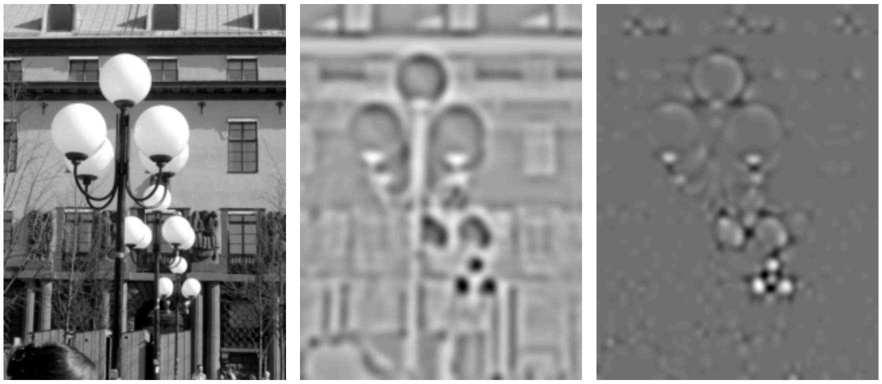

Figure 4: *Differential descriptors for blob detection/interest point detection:* (left) A grey-level image of size 210×280 pixels. (middle) The Laplacian $\nabla^2 L$ computed at $t = 16$. (right) The determinant of the Hessian $\det \mathcal{H} L$ computed at $t = 16$.

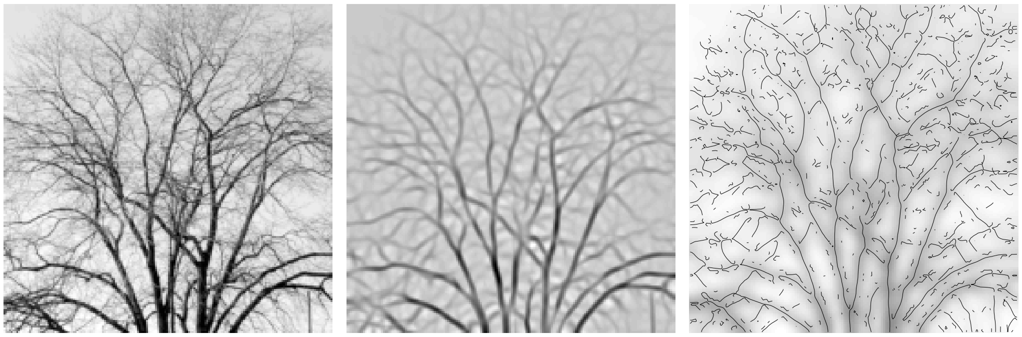

Figure 5: *Fixed scale valley detection:* (left) A grey-level image of size 180×180 pixels. (middle) The negative value of the valley strength measure L_{qq} computed at $t = 4$. (right) Differential geometric valleys detected at $t = 4$ using a complementary low threshold on $t|L_{qq}| \geq 1$ and then overlayed on a bright copy of the original grey-level image.

Windowed image descriptors with two scale parameters

The image descriptors considered so far are all dependent on a single scale parameter t. For certain problems, it is useful to introduce image descriptors that depend on two scale parameters. One such descriptor is the *second-moment matrix* (structure tensor) defined as

$$\mu(x, y;\ t, s) = \int_{(\xi,\eta)\in\mathbb{R}^2} \begin{pmatrix} L_x^2(\xi,\eta;\ t) & L_x(\xi,\eta;\ t)\, L_y(\xi,\eta;\ t) \\ L_x(\xi,\eta;\ t)\, L_y(\xi,\eta;\ t) & L_y^2(\xi,\eta;\ t) \end{pmatrix} g(x-\xi, y-\eta;\ s)\, d\xi\, d\eta \tag{10}$$

where t is a local scale parameter describing the scale of differentiation and s is an integration scale parameter describing the extent over which local statistics of derivatives is accumulated. In principle, the formulation of this descriptor implies a two-parameter variation. In many practical applications, however, it is common practice to couple the two scale parameters by a constant factor C such that $s = Ct$ where $C > 1$.

One common application of this descriptor is for a multi-scale version of the *Harris corner detector* [12], by detecting positive spatial maxima of the entity

$$H = \det(\mu) - k\, \mathrm{trace}^2(\mu) \tag{11}$$

where $k \approx 0.04$ is a constant. Another common application is for affine normalization/shape adaptation and will be developed below. The eigenvalues $\lambda_{1,2} = \frac{1}{2}(\mu_{11} + \mu_{22} \pm \sqrt{(\mu_{11} - \mu_{22})^2 + 4\mu_{12}^2})$

and the orientation $\arg(\mu_{11} - \mu_{22}, 2\mu_{12})$ of μ are also useful for *texture segmentation* and texture classification.

Other commonly used image descriptors that depend an additional integration scale parameter include regional histograms obtained using Gaussian window functions as weights [18].

Scale-space representation of colour images

The input images f considered so far have all been assumed to be scalar grey-level images. For a vision system, however, colour images are often available, and the use of colour cues can increase the robustness and discriminatory power of image operators. For the purpose of scale-space representation of colour images, an initial red-green blue-yellow colour opponent transformation is often advantageous [11]. While the computation of a scale-space representation of the RGB channels independently does not necessarily constitute the best way of defining a colour scale-space, by performing the following pre-transformation prior to scale-space smoothing

$$\begin{cases} I = (R+G+B)/3 \\ U = R - G \\ V = B - (R+G)/2 \end{cases} \tag{12}$$

Gaussian scale-space smoothing, Gaussian derivatives and differential invariants can then be defined from the IUV colour channels separately. The luminance channel I will then mainly reflect the interaction between reflectance, illuminance direction and intensity, while the chromatic channels, U and V, will largely make the interaction between illumination colour and surface pigmentation more explicit. This approach has been successfully applied to the tasks of image feature detection and image-based recognition based on the N-jet. For example, red-green and blue-yellow colour opponent receptive fields of center-surround type can be obtained by applying the Laplacian $\nabla^2 L$ to the chromatic U and V channels.

An alternative approach to handling colours in scale-space is provided by the *Gaussian colour model*, in which the spectral energy distribution $E(\lambda)$ over all wavelengths λ is approximated by the sum of a Gaussian function and first- and second-order derivatives of Gaussians [17]. In this way, a three-dimensional colour space is obtained, where the channels $\hat{E}$, $\hat{E}_\lambda$ and $\hat{E}_{\lambda\lambda}$ correspond to a second-order Taylor expansion of a Gaussian-weighted spectral energy distribution around a specific wavelength λ_0 and smoothed to a fixed spectral scale t_{λ_0}. In practice, this model can be implemented in different ways. For example, with $\lambda_0 = 520$ nm and $t_{\lambda_0} = 55$ nm, the Gaussian colour model has been approximated by the following colour space transformation [9]:

$$\begin{pmatrix} \hat{E} \\ \hat{E}_\lambda \\ \hat{E}_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} -0.019 & 0.048 & 0.011 \\ 0.019 & 0 & -0.016 \\ 0.047 & -0.052 & 0 \end{pmatrix} \begin{pmatrix} 0.621 & 0.113 & 0.194 \\ 0.297 & 0.563 & 0.049 \\ -0.009 & 0.027 & 1.105 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{13}$$

using the CIE 1931 XYZ colour basis as an intermediate representation. In analogy with the IUV colour space, spatio-spectral Gaussian derivatives and differential invariants can then be defined by applying Gaussian smoothing and Gaussian derivatives to the channels in this representation. This approach has been applied for constructing approximations of colour invariants assuming specific models for illumination and/or surface properties [10].

Affine scale-space, affine image deformations and affine covariance

The regular linear scale-space representation obtained by smoothing with the rotationally symmetric Gaussian kernel is closed under translations, rotations and rescalings. This means that image transformations within this group can be captured perfectly by regular scale-space operators. To obtain a scale-space representation that is closed under affine transformations, a natural generalization is to consider an *affine scale-space* [22] obtained by the convolution with Gaussian kernels with their shapes determined by positive definite covariance matrices

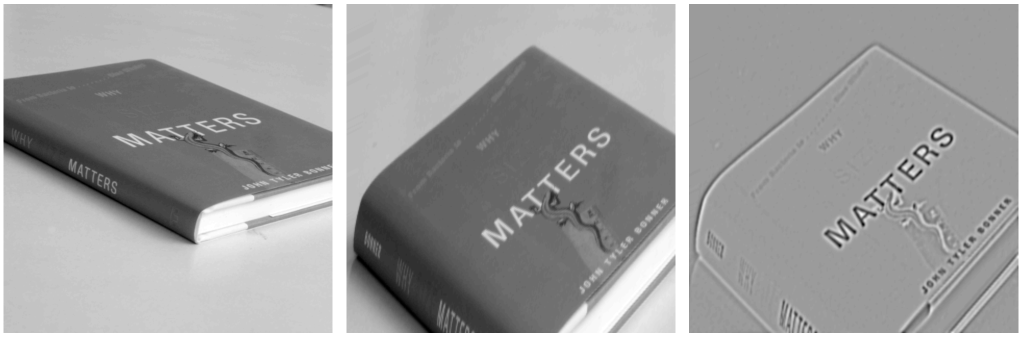

Figure 6: *Affine normalization by shape adaptation in affine scale-space:* (left) A grey-level image with an oblique view of a book cover. (middle) The result of affine normalization of a central image patch using iterative shape adaptation with affine transformations proportional to $A = \mu^{1/2}$. (right) An example of computing differential geometric descriptors, here the Laplacian $\nabla^2 L$ at scale $t = 2$, in the affinely normalized frame.

Σ:

$$g(\bar{x};\ \Sigma_t) = \frac{1}{2\pi\sqrt{\det \Sigma_t}} e^{-\bar{x}^T \Sigma_t^{-1} \bar{x}/2}, \tag{14}$$

where $\bar{x} = (x, y)^T$. This affine scale-space combined with directional derivatives can serve as a model for *oriented elongated filter banks*. Consider two input images f and f' that are related by an affine transformation $\bar{x}' = A\bar{x}$ such that $f'(A\bar{x}) = f(\bar{x})$. Then, the affine scale-space representations L and L' of f and f' are related according to $L'(x';\ \Sigma') = L(x;\ \Sigma)$ where $\Sigma' = A \Sigma A^T$. A second-moment matrix (10) defined from an affine scale-space with covariance matrices Σ_t and Σ_s transforms according to

$$\mu'(A\bar{x};\ A\Sigma_t A^T, A\Sigma_s A^T) = A^{-T} \mu(x;\ \Sigma_t, \Sigma_s) A^{-1}. \tag{15}$$

If we can determine covariance matrices Σ_t and Σ_s such that $\mu(x;\ \Sigma_t, \Sigma_s) = c_1 \Sigma_t^{-1} = c_2 \Sigma_s^{-1}$ for some constants c_1 and c_2, we obtain a fixed-point that is preserved under affine transformations. This property has been used for expressing *affine invariant interest point operators*, *affine invariant stereo matching* as well as *affine invariant texture segmentation and texture recognition methods* [26, 1, 28, 33, 20]. In practice, affine invariance at a given image point can (up to an unknown scale factor and a free rotation angle) be accomplished by *shape adaptation*, that is by estimating the second-moment matrix μ using a rotationally symmetric scale-space and then iteratively either choosing the covariance matrices Σ_t and Σ_s proportional to μ^{-1} until the fixed-point has been reached or equivalently by warping the input image by linear transformations proportional to $A = \mu^{1/2}$ until the second-moment matrix is sufficiently close to a constant times the unit matrix.

Automatic scale selection and scale invariant image descriptors

Although the scale-space theory presented so far provides a well-founded framework for expressing visual operations at multiple scales, it does not address the problem of how to *select* locally appropriate scales for further analysis. Whereas the problem of finding "the best scales" for handling a given data set may be regarded as intractable unless further information is available, there are many situations in which a mechanism is required for generating hypotheses about interesting scales for further analysis. Specifically, since the size of and the distance to objects may vary in real-life applications, there is a need for defining scale invariant image descriptors. A general methodology [23] for generating hypotheses about interesting scale levels is by studying the evolution properties over scales of (possibly non-linear) combinations of *γ-normalized derivatives* defined by

$$\partial_\xi = t^{\gamma/2}\, \partial_x \quad \text{and} \quad \partial_\eta = t^{\gamma/2}\, \partial_y \tag{16}$$

where γ is a free parameter to be determined for the task at hand. Specifically, *scale levels can be selected from the scales at which γ-normalized derivative expressions assume local extrema with respect to scale*. A general rationale for this statement is that under a scaling transformation $(x', y') = (sx, sy)$ for some scaling factor s with $f'(sx, sy) = f(x, y)$, if follows that for matching scale levels $t' = s^2 t$ the m:th order normalized derivatives at corresponding scales $t' = s^2 t$ in scale-space transform according to

$$L'_{\xi'^m}(x', y';\ t') = s^{m(\gamma-1)} L_{\xi^m}(x, y;\ t). \tag{17}$$

Hence, for any differential expression that can be expressed as a homogeneous polynomial in terms of γ-normalized derivatives, it follows that local extrema over scales will be preserved under scaling transformations. In other words, if a γ-normalized differential entity $\mathcal{D}_{norm} L$ assumes an extremum over scales at the point $(x_0, y_0;\ t_0)$ in scale-space, then under a rescaling transformation an extremum in the transformed differential invariant is assumed at $(sx_0, sy_0;\ s^2 t_0)$ in the scale-space L' of the transformed image. This general property means that if we can find an expression based on γ-normalized Gaussian derivatives that assumes a local extremum over scales for a suitable set of image structures, then we can define a *scale invariant feature detector and/or image descriptor* by computing it at the scale at which the local extremum over scales is assumed.

The scale estimate $\hat{t}$ obtained from the scale selection step can therefore be used for tuning or guiding other early visual processes to be truly scale invariant [22, 23]. With the incorporation of a mechanism for automatic scale selection, *scale-tuned visual modules* in a vision system will have the ability to handle objects of different sizes as well as objects with different distances to the camera in the same manner. These are essential requirement for any vision system intended to function robustly in a complex dynamic world.

By studying the γ-normalized derivative response to a one-dimensional sine wave $f(x) = \sin(\omega x)$, for which the maximum over scales in the m:th order derivative is assumed at scale $\sigma_{max} = \sqrt{\gamma m} \lambda / 2\pi$ (measured in units of the standard deviation of the Gaussian) is proportional to the wavelength $\lambda = 2\pi/\omega$ of the signal, one can see that there is a qualitative similarity between this construction and a peak in a local Fourier transform. There are however also two major differences: (i) no window size is needed for computing the Fourier transform, and (ii) this approach applies also to non-linear differential expressions.

Specifically, if we choose $\gamma = 1$, then under scaling transformations the magnitudes of normalized scale-space derivatives in equation (17) are *equal* at corresponding points in scale-space. For $\gamma \neq 1$ they are related according to a (known) power of the scaling factor (see also [6] for earlier work on receptive field responses under similarity transformations).

Scale invariant feature detectors with integrated scale selection mechanism

The most commonly used entity for automatic scale selection is the *scale normalized Laplacian* [22, 23]

$$\nabla^2_{norm} L = t(L_{xx} + L_{yy}) \tag{18}$$

with $\gamma = 1$. A general motivation for the usefulness of this descriptor for general purpose scale selection can be obtained from the fact that the scale-space representation at any point can be decomposed into an integral of Laplacian responses over scales:

$$L(x, y;\ t_0) = -(L(x, y;\ \infty) - L(x, y;\ t_0)) = -\int_{t=t_0}^{\infty} \partial_t L(x, y;\ t)\, dt = -\frac{1}{2} \int_{t=t_0}^{\infty} \nabla^2 L(x, y;\ t)\, dt. \tag{19}$$

After a reparameterization of the scale parameter into *effective scale* $\tau = \log t$, we obtain:

$$L(x, y;\ t_0) = -\frac{1}{2} \int_{t=t_0}^{\infty} t\, \nabla^2 L(x, y;\ t)\, \frac{dt}{t} = -\int_{\tau=\tau_0}^{\infty} \nabla^2_{norm} L(x, y;\ \tau)\, d\tau. \tag{20}$$

By detecting the scale at which the normalized Laplacian assumes its positive maximum or negative minimum over scales, we therefore determine the scale for which the image in a scale-normalized bandpass sense contains the maximum amount of information.

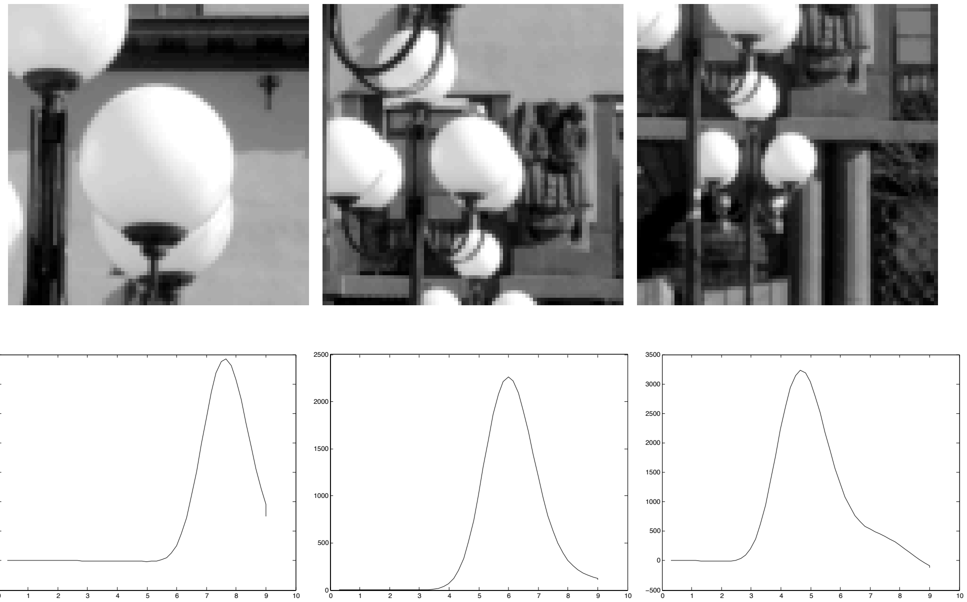

Figure 7: *Automatic scale selection from local extrema over scales of normalized derivatives:* (top row) Subwindows showing different details from figure 4 with image structures of different size. (bottom row) *Scale-space signatures* of the scale normalized determinant of the Hessian $\det \mathcal{H}_{norm} L$ accumulated at the centre of each window. The essential property of the scale dependency of these scale normalized differential entities is that *the scale at which a local extremum over scales is assumed is proportional to the size of the corresponding image structure in the image domain.* The horizontal axis on these graphs represent effective scale, roughly corresponding to the logarithm of the scale parameter: $\tau \approx \log_2 t$.

Another motivation for using the scale-normalized Laplacian operator for early scale selection is that it serves as an excellent blob detector. For any (possibly non-linear) γ-normalized differential expression $\mathcal{D}_{norm}L$, let us first define a *scale-space maximum* (minimum) as a point for which the γ-normalized differential expression assumes a maximum (minimum) over *both* space and scale. Then, we obtain a straightforward *blob detector with automatic scale selection that responds to dark and bright blobs*, from the scale-space maxima and the scale-space minima of $\nabla^2_{norm} L$. Another *blob detector with automatic scale selection that also responds to saddles* can be defined from the scale-space maxima and minima of the scale-normalized determinant of the Hessian [23]:

$$\det \mathcal{H}_{norm} L = t^2 (L_{xx} L_{yy} - L_{xy}^2). \tag{21}$$

For both of these scale invariant blob detectors, *the selected scale reflects the size of the blob*. For the purpose of scale invariance, it is, however, not necessary to use the same entity for determining spatial interest points as for determining interesting scales for those. An alternative approach to *scale invariant interest point detection* is to use the Harris operator in equation (11) for determining spatial interest points and the scale normalized Laplacian for determining the scales at these points [28]. *Affine covariant interest points* can in turn be obtained by combining any of these three interest point operators with subsequent affine shape adaptation following equation (15) combined with a determination of the remaining free rotation angle using, for example, the orientation of the image gradient.

The free parameter γ in the γ-normalized derivative concept can be related to the dimensionality of the type of image features to be detected or to a normalization of the Gaussian derivative kernels to constant L_p-norm over scales. For blob-like image descriptors, such as the interest points described above, $\gamma = 1$ is a good choice and corresponds to L_1-normalization. For other types of image structures, such as thin edges, elongated ridges or rounded corners, other values will be preferred. For example, for the problem of edge detection, $\gamma = 1/2$ is a useful choice for capturing the width of diffuse edges, while for the purpose of ridge detection, $\gamma = 3/4$ is preferable for tuning the scale levels to the width of an elongated ridge [24, 8, 19].

Scale-space axioms

Besides its practical usefulness for computer vision problems, the scale-space representation satisfies a number of theoretical properties that define it as a unique form of multi-scale image representation: The linear scale-space representation is obtained from *linear* and *shift-invariant* transformations. Moreover, the Gaussian kernels are *positive* and satisfy the *semi-group property*

$$g(\cdot, \cdot;\ t_1) * g(\cdot, \cdot;\ t_2) = g(\cdot, \cdot;\ t_1 + t_2). \tag{22}$$

which implies that any coarse scale representation can be computed from any fine-scale representation using a similar transformation as in the transformation from the original image

$$L(\cdot, \cdot;\ t_2) = g(\cdot;\ t_2 - t_1) * L(\cdot, \cdot;\ t_1). \tag{23}$$

In one dimension, Gaussian smoothing implies that *new local extrema or new zero-crossings cannot be created with increasing scales* [37, 22]. In two and higher dimension, the scale-space representation obeys *non-enhancement of local extrema* (causality), which implies that the value at a local maximum is guaranteed not to increase while the value at a local minimum is guaranteed to not decrease [14, 22, 25]. The regular linear scale-space is *closed* under

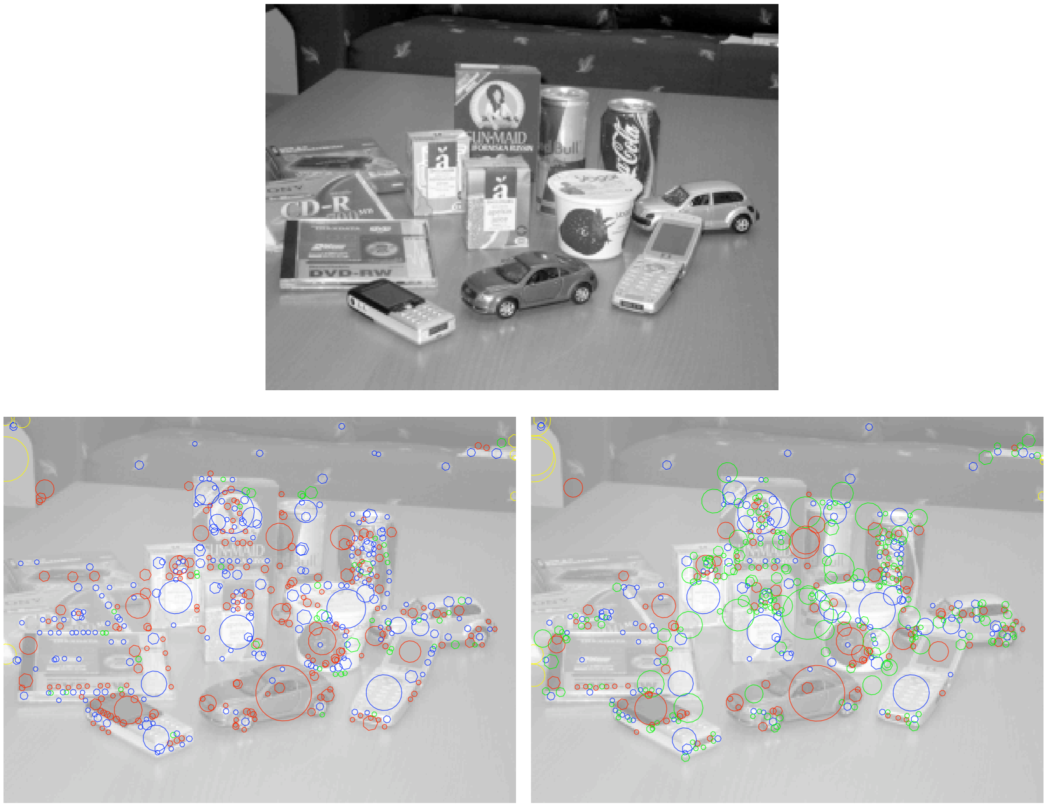

Figure 8: *Scale-invariant feature detection:* (top) Original grey-level image. (bottom left) The 1000 strongest scale-space extrema of the scale normalized Laplacian ∇^2_{norm}. (bottom right) The 1000 strongest scale-space extrema of the scale normalized determinant of the Hessian $\det \mathcal{H}_{norm} L$. Each feature is displayed by a circle with its size proportional to the detection scale. In addition, the colour of the circle indicates the type of image feature; red for dark features, blue for bright features and green for saddle-like features.

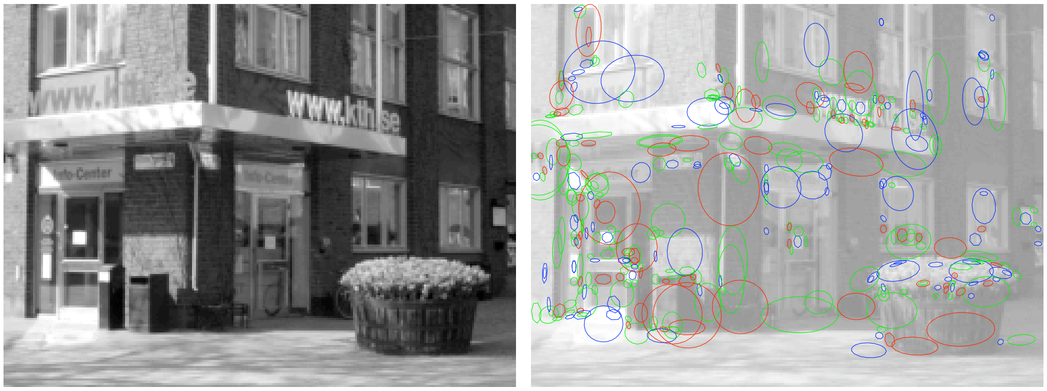

Figure 9: *Affine covariant image features:* (left) Original grey-level image. (right) The result of applying affine shape adaptation to the 500 strongest scale-space extrema of the scale normalized determinant of the Hessian $\det \mathcal{H}_{norm} L$ (resulting in 384 features for which the iterative scheme converged). Each feature is displayed by an ellipse with its size proportional to the detection scale and the shape determined from a linear transformation A determined from a second-moment matrix μ. In addition, the colour of the ellipse indicates the type of image feature; red for dark features, blue for bright features and green for saddle-like features.

translations, rotations and scaling transformations. In fact, it can be shown that Gaussian smoothing arises *uniquely* for different subsets of combinations of these special and highly useful properties. For partial views of the history of scale-space axiomatics, please refer to [25, 36] and the references therein.

Concerning the topic of automatic scale selection, it can be shown that the notion of γ-normalized derivatives in equation (16) arises by necessity from the requirement that local extrema over scales should be preserved under scaling transformations [23].

Relations to biological vision

Interestingly, the results of this computationally motivated analysis of early visual operations are in qualitative agreement with current knowledge about biological vision. Neurophysiological studies have shown that there are receptive field profiles in the mammalian retina and visual cortex that can be well modelled by Gaussian derivative operators [38, 5].

Summary and outlook

Scale-space theory provides a well-founded framework for modelling image structures at multiple scales, and the output from the scale-space representation can be used as input to a large variety of visual modules. Visual operations such as feature detection, feature classification, stereo matching, motion estimation, shape cues and image-based recognition can be expressed directly in terms of (possibly non-linear) combinations of Gaussian derivatives at multiple scales. In this sense, scale-space representation can serve as a basis for early vision.

The set of early uncommitted operations in a vision system, which perform scale-space smoothing, compute Gaussian derivatives at multiple scales and combine these into differential invariants or other types of general purpose features to be used as input to later stage visual processes, is often referred to as a *visual front-end*.

Pyramid representation is a predecessor to scale-space representation, constructed by simultaneously smoothing and subsampling a given signal [3, 4]. In this way, computationally highly efficient algorithms can be obtained. A problem with pyramid representations, however, is that it is algorithmically harder to relate structures at different scales, due to the discrete nature of the scale levels. In a scale-space representation, the existence of a continuous scale parameter makes it conceptually much easier to express this *deep structure*.

For features defined as zero-crossings of differential invariants, the implicit function theorem directly defines trajectories across scales, and at those scales where bifurcations occur, the local behaviour can be modelled by singularity theory. Nevertheless, pyramids are frequently used for expressing computationally more efficient approximations to different scale-space algorithms.

Extensions of linear scale-space theory concern the formulation of non-linear scale-space concepts more committed to specific purposes [29, 35]. There are strong relations between scale-space theory and *wavelets*, although these two notions of multi-scale representations have been developed from somewhat different premises.

References

[1] Baumberg, A. (2000) ‘Reliable feature matching across widely separated views’, *Proc. Comp. Vision Patt. Recogn.*, I:1774–1781.

[2] Bay, H., Tuytelaars, T. and van Gool, Luc "SURF: Speeded Up Robust Features", In: *Proc. European Conf. on Computer Vision*, Springer LNCS 3951, I:404–417, 2006.

[3] Burt, P. J. and Adelson, E. H. (1983) ‘The Laplacian pyramid as a compact image code’, *IEEE Trans. Comm.* 9:4, 532–540.

[4] Crowley, J. and Parker, A. C. (1984) ‘A representation for shape based on peaks and ridges in the Difference of Low-Pass Transform’, *IEEE Trans. Pattern Anal. Machine Intell.* 6(2), 156–170.

[5] DeAngelis, G. C., Ohzawa, I. and Freeman, R. D. (1995) ‘Receptive field dynamics in the central visual pathways’, *Trends in Neuroscience* 18(10), 451–457.

[6] Field, D. J. (1987) ‘Relations between the statistics of natural images and the response properties of cortical cells’, *J. Opt. Soc. Am.* 4, 2379–2394.

[7] Florack, L. M. J. (1997) *Image Structure*, Kluwer/Springer.

[8] Frangi, A. F., Niessen W. J., Hoogeveen, R. M., van Walsum, T., Viergever, M. A. (2000) ‘Model-based detection of tubular structures in 3D images’, *IEEE Trans. Med. Imaging*, 18:10, 946-56.

[9] Geusebroek, J. M., van den Boomgaard, R., Smeulders, A. W. M. and Dev, A. (2000) ‘Color and scale: The spatial structure of color images’. In: *Proc. European Conf. on Computer Vision*, Springer LNCS 1842, I:331–341.

[10] Geusebroek, J. M., van den Boomgaard, R., Smeulders, A. W. M. and Geerts, H. (2001) ‘Color invariance’, *IEEE Patt. Anal. Mach. Intell*, 23:12, 1338-1346.

[11] Hall, D., de Verdiere, V. and Crowley, J. (2000) ‘Object recognition using coloured receptive fields, In: *Proc. European Conf. on Computer Vision*, Springer LNCS 1842, I:164–177.

[12] Harris, C. and Stevens, M. (1988) ‘A combined corner and edge detector’, In: *Proc. Alvey Workshop on Visual Motion*, 156–162.

[13] Jones, D. G. and Malik, J. (1992) ‘A computational framework for determining stereo correspondences from a set of linear spatial filters, In: *Proc. Eur. Conf. Comp. Vis.*, 395–410.

[14] Koenderink, J. J. (1984) ‘The structure of images’, *Biological Cybernetics* 50, 363–370.

[15] Koenderink, J. J. and van Doorn, A. J. (1987) ‘Representation of local geometry in the visual system’, *Biological Cybernetics* 55, 367–375.

[16] Koenderink, J. J. and van Doorn, A. J. (1992) ‘Generic neighborhood operators’, *IEEE Trans. Pattern Anal. Machine Intell.* 14(6), 597–605.

[17] Koenderink, J. J. and Kappers, A. (1998) ‘Colour space’, unpublished lecture notes, Utrecht University, The Netherlands.

[18] Koenderink, J. J. and van Doorn, A. J. (1999) ‘The structure of locally orderless images’, *Int. J. of Computer Vision* 31(2), 159–168.

[19] Krissian, K., Malandain, G., Ayache N., Vaillant R., and Trousset Y. (2000) ‘Model-based detection of tubular structures in 3D images’, *Comput. Vis. Image Underst.*, 80:2, 130–171.

[20] Lazebnik, S., Schmid, C. and Ponce, J. (2003) ‘Affine-invariant local descriptors and neighbourhood statistics for texture recognition’, In: *Proc. Int. Conf. Comp. Vis.*, I:649–655.

[21] Leung, T. and Malik, J. (2001) ‘Representing and recognizing the visual appearance of materials using three-dimensional textons’, *Int. J. of Computer Vision*, 43(1), 29–44.

[22] Lindeberg, T. (1994) *Scale-Space Theory in Computer Vision*, Kluwer/Springer.

[23] Lindeberg, T. (1998*a*) ‘Feature detection with automatic scale selection’, *Int. J. of Computer Vision* 30(2), 77–116.

[24] Lindeberg, T. (1998*b*) ‘Edge detection and ridge detection with automatic scale selection’, *Int. J. of Computer Vision*, 30(2), 117–154.

[25] Lindeberg, T. (1997) ‘On the axiomatic foundations of linear scale-space: Combining semi-group structure with causailty vs. scale invariance’, In: J. Sporring et al (eds.) *Gaussian Scale-Space Theory*, pp. 75–98, Kluwer/Springer.

[26] Lindeberg, T. and Gårding, J. (1997) ‘Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D structure’, *Image and Vision Computing* 15, 415–434.

[27] Lowe, D. (2004) ‘Distinctive image features from scale-invariant keypoints’, *Int. J. of Computer Vision*, 60:2, 91-110

[28] Mikolajczyk, K. and Schmid, C. (2004) ‘Scale and affine invariant interest point detectors’, *Int. J. of Computer Vision*, 60:1, 63 - 86.

[29] Romeny, B. t. H. ed. (1997) *Geometry-Driven Diffusion in Computer Vision*, Kluwer/Springer.

[30] Romeny, B. t. H. (2003) *Front-End Vision and Multi-Scale Image Analysis*, Kluwer/Springer.

[31] Schiele, B. and Crowley, J. (2000) ‘Recognition without correspondence using multidimensional receptive field histograms’, *Int. J. of Computer Vision* 36(1), 31–50.

[32] Schmid, C. and Mohr, R. (1997) ‘Local grayvalue invariants for image retrieval’, *IEEE Trans. Pattern Anal. Machine Intell.* 19(5), 530–535.

[33] Schaffalitzky, F. and Zisserman, A. (2001) ‘Viewpoint invariant texture matching and wide baseline stereo’, In: *Proc. Int. Conf. Comp. Vis.*, 2:636–644.

[34] Sporring, J., Nielsen, M., Florack, L. and Johansen, P., eds. (1996) *Gaussian Scale-Space Theory: Proc. PhD School on Scale-Space Theory*, Kluwer/Springer.

[35] Weickert, J. (1998) *Anisotropic Diffusion in Image Processing*, Teubner-Verlag, Germany.

[36] Weickert, J. ‘Linear scale space has first been proposed in Japan’, *J. Math. Imaging and Vision*, 10(3):237–252, 1999.

[37] Witkin, A. P. (1983) Scale-space filtering, In: *Proc. 8th Int. Joint Conf. Art. Intell.*, 1019–1022.

[38] Young, R. A. (1987) ‘The Gaussian derivative model for spatial vision’, *Spatial Vision* 2, 273–293.