In: *Advances in Imaging and Electron Physics*, Elsevier, volume 178, pages 1-96, 2013.

Generalized axiomatic scale-space theory[*]

Tony Lindeberg

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

Abstract

A fundamental problem in vision concerns what types of image operations should be used at the first stages of visual processing. This paper presents a principled approach to this problem by describing a generalized axiomatic scale-space theory that makes it possible to derive the notions of linear scale-space, affine Gaussian scale-space and linear spatio-temporal scale-space using similar sets of assumptions (scale-space axioms).

Based on a requirement that new image structures should not be created with increasing scale formalized into a condition of non-enhancement of local extrema, a complete classification is given of the linear (Gaussian) scale-space concepts that satisfy these conditions on isotropic spatial, non-isotropic spatial and spatio-temporal domains, which results in a general taxonomy of Gaussian scale-spaces for continuous image data. The resulting theory allows filter shapes to be tuned from specific context information and provides a theoretical foundation for the recently exploited mechanisms of affine shape adaptation and Galilean velocity adaptation, with highly useful applications in computer vision. It is also shown how time-causal and time-recursive spatio-temporal scale-space concepts can be derived from similar or closely related assumptions.

The receptive fields arising from the spatial, spatio-chromatic and spatio-temporal derivatives resulting from these scale-space concepts can be used as a general basis for expressing image operations for a large class of computer vision or image analysis methods. The receptive field profiles generated by necessity from these theories do also have close similarities to receptive fields measured by cell recordings in biological vision, specifically regarding space-time separable cells in the retina and the lateral geniculate nucleus (LGN) as well as both space-time separable and non-separable cells in the striate cortex (V1) of higher mammals.

Keywords: scale-space, multi-scale representation, scale-space axioms, non-enhancement of local extrema, scale invariance, Gaussian kernel, Gaussian derivative, spatial, affine, spatio-chromatic, spatio-temporal, time-causal, time-recursive, receptive field, diffusion, computer vision, image analysis, image processing.

Digitally published with doi:10.1016/B978-0-12-407701-0.00001-7.

[*] The support from the Swedish Research Council, Vetenskapsrådet (contract 2010-4766), the Royal Swedish Academy of Sciences as well as the Knut and Alice Wallenberg Foundation is gratefully acknowledged.

Contents

1 Introduction

For us humans as well as many other living organisms, vision is a main source of information about the surrounding world, which allows us to gather information about objects in our environment at a distance and without interacting with them physically. In science, medicine and our daily life, digital information in the form of images and video is becoming ubiquitous by the developments of visual sensors and information technology.

While we humans seemingly effortlessly make use of visual perception to recognize and categorize familiar as well as unfamiliar objects and activities in complex environments, there are several reasons why visual tasks are hard to automate. One main source of difficulty originates from the fact that visual input does not provide direct, but indirect, information about the objects that the light is reflected from. The image formation process is associated with a set of natural image transformations, including the perspective mapping, which implies scale variations due to depth as well as perspective distortions due to variations in the viewing direction. For time-dependent image data, objects usually move relative to the observer. Moreover, image measurements are strongly influenced by external illumination, occlusion and interference with neighbouring objects in the environment. Therefore, individual measurements of intensity and/or colour at single image points do hardly ever provide sufficient information about an object, since beyond material properties, any pointwise measurement is strongly influenced by the external illumination as well as the orientation of the surface of the object relative to the viewing direction of the observer. The relevant information for visual perception is instead contained in the *relative* relations between image measurements at adjacent points. A key issue in computer vision and image processing is therefore to design image features and image descriptors that are sufficiently descriptive and discriminatory to be able to distinguish between different types of visual patterns, while also being robust (invariant) to the types of transformations that are present in the image formation process, as well as robust to noise and other perturbations. Borrowing terminology originally developed for biological vision (Hubel & Wiesel 2005), an image operator that computes image features or image descriptor from the local image information around a specific point in space or space-time can be referred to as a receptive field.

The problem of computing or designing appropriate image features from image data, or alternatively stated modelling the function of visual receptive fields, is closely related to the fact that real-world objects are composed of different types of structures at different scales. A general methodology that has been developed to handle this issue is by imposing structural constraints on the image operators, which reflect symmetry properties in the world. From such arguments, theoretical results have been presented showing that Gaussian kernels and Gaussian derivatives constitute a canonical class of image operations, and can be regarded as an idealized model for linear receptive fields over a spatial (time-independent) image domain (Iijima 1962, Koenderink 1984, Koenderink & van Doorn 1992, Lindeberg 1994*b*, Lindeberg 1994*a*, Florack 1997, ter Haar Romeny 2003, Lindeberg 2011). Empirical evidence have shown that this is a highly fruitful approach, and *scale-space theory* developed from these principles has established itself as a promising paradigm for early vision with a large number of successful applications, including feature detection, stereo matching, computation of optic flow, tracking, estimation of shape cues and view-based object recognition (Lindeberg 2008).

Whereas scale-space theory was originally developed to handle (i) image structures at different scales in spatial data to make it possible to handle image phenomena caused by objects having substructures of different size and of different distances to the observer, scale-space theory has later evolved into a general theory of early visual operations, to handle other types of image variations including (ii) locally linearized image deformations caused by variations in the viewing direction relative to the object, (ii) locally linearized relative motions between the object and the observer in spatial-temporal image data (video) and (iv) the effect of local illumination transformations on receptive field responses (Lindeberg 2013*b*). Computational mechanisms developed from these premises have also been shown to be consistent with properties of receptive fields measured by cell recordings in biological vision (Young 1987, Young et al. 2001, Lindeberg 2011, Lindeberg 2013*a*, Lindeberg 2013*b*).

During the last decades, a large number of interesting developments have been made regarding computational methods for automated interpretation of visual information; several of these either based on or with close relations to scale-space theory and image measurements in terms of receptive fields. Prior to the establishment of scale-space theory it was indeed very hard to construct computer vision algorithms that work robustly on real-world image data acquired under natural imaging conditions. Specifically, mechanisms for scale invariance as obtained as one of the consequences of this theory (Lindeberg 1998, Lindeberg 2013*c*) have allowed for breakthroughs regarding methods for image matching and object recognition with important follow-up consequences for constructing computer vision systems in several domains.

The subject of this article is to give an overview of some of the theoretical foundations of scale-space theory by showing how classes of natural image operations can be singled out in an axiomatic manner by imposing structural constraints on permissible classes of image operations. The approach will bear close resemblance to approaches in theoretical physics, where symmetry properties of the world are used for constraining physical theories.

1.1 Organization of the presentation

The presentation begins in section 2 with a general treatment of how measurements of signals from the real world, such as image data, are intimately related to the notion of scale. Section 3 gives an overview of some of the basic structural assumptions of scale-space theory as they can be motivated by the requirement of enabling consistent measurements of image observations for a camera or a visual agent that observes objects in the real world under the variability of natural image transformations. Section 4 explains how these structural assumptions can be formalized into a set of scale-space axioms for image data that are defined over a purely spatial (time-independent) image domain. Section 5 then describes the spatial Gaussian scale-space concepts that arise by necessity from the assumptions. Section 6 develops a corresponding set of scale-space axioms for (time-varying) spatio-temporal image data, and Section 7 shows how three different types of spatio-temporal concepts can be obtained from these assumptions depending on how the special nature of time is treated, and corresponding to different temporal smoothing kernels over time in the respective cases as developed in Section 8. Section 9 gives an overview of the history of previous axiomatic scale-space formulations. Finally, Section 10 concludes with a summary and discussion.

2 Image measurements and the notion of scale

The process of image measurements implies that the incoming light that falls on the visual sensor must be integrated over some non-infinitesimal region over space for some non-infinitesimal amount of time. For a two-dimensional camera, we can model the image intensity I that is sampled from the incoming image irradiance E at an image point $x = (x_1, x_2)^T$

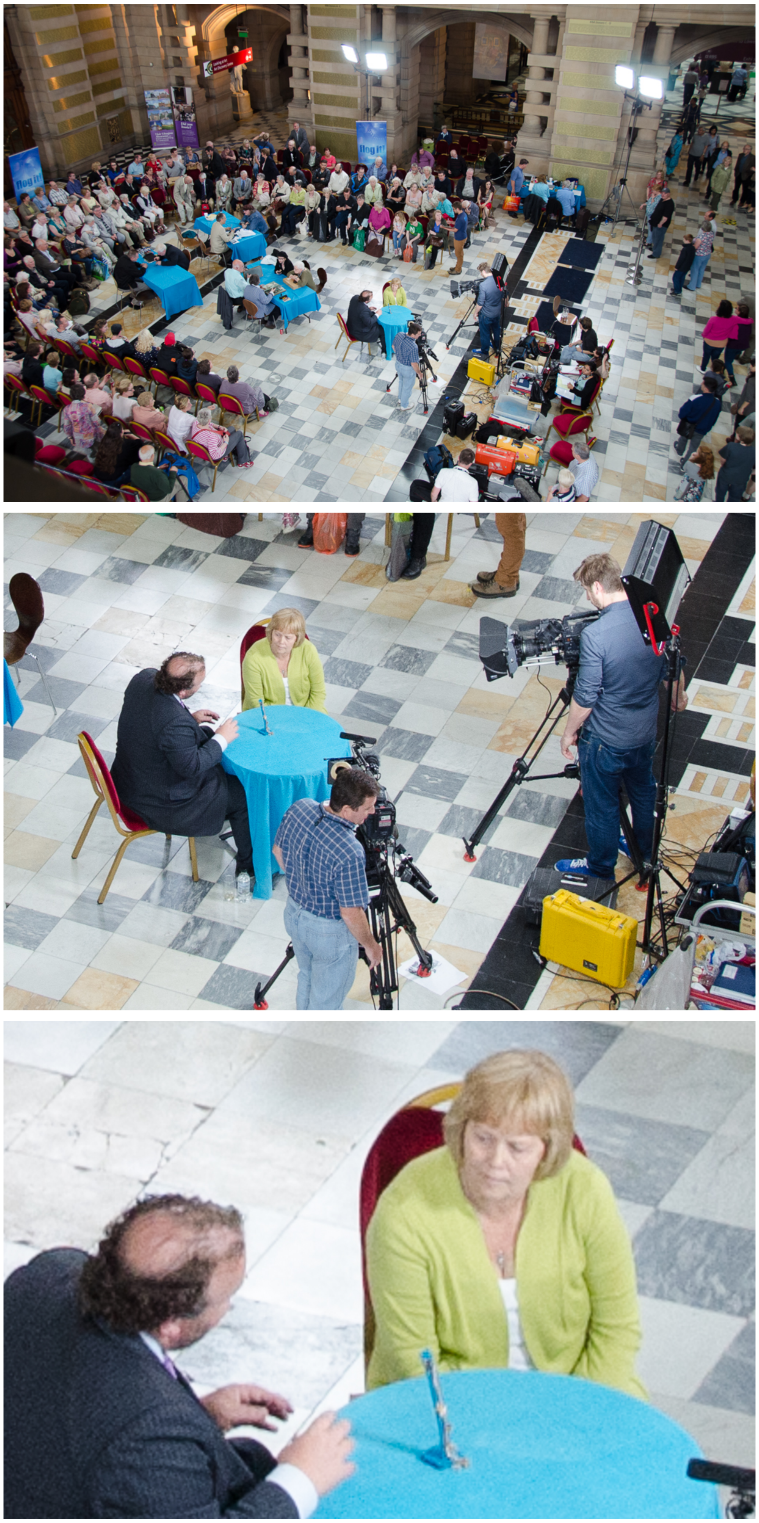


Figure 1: Illustration of qualitatively different types of image structures that may appear in image data depending on the scale of observation. This figure simulates this phenomenon by gradually zooming in to finer scale image structures in a high resolution photograph. At a coarse level of scale, the crowd may be perceived as a kind of texture, whereas finer scale structures become visible as we zoom in to finer scales, such as individual faces and substructures in these. In a corresponding manner, a corresponding manifestation of qualitatively different types of image structures depending on the scale of observation will arise in many other imaging, image analysis or computer vision applications.

and time moment t according to

$$I(x,t) = \int_{\xi \in \mathbb{R}^2} \int_{\eta \in \mathbb{R}} \int_{\lambda \in \mathbb{R}} E(x - \xi, t - \eta, \lambda) \, g(\xi) \, h(\eta) \, \Lambda(\lambda) \, d\xi \, d\eta \, d\lambda \tag{1}$$

where

- $g(\xi)$ is a spatial window function over which spatial integration is performed,
- $h(\eta)$ is a temporal window function over which temporal integration is performed and
- $\Lambda(\lambda)$ is the wavelength sensitivity function of the sensor.

The spatial extent of this support region defined by the window function g and the temporal integration time defined by the temporal window h do therefore define natural inner spatial and temporal scales of the visual observation beyond which further information is not accessible (see figure 2 for an illustration). When analyzing the image data by a computerized vision or image analysis system, alternatively in biological perception, it is, however, not at all evident that these scale levels would be the best scales for computing image features from the image data. Therefore, a mechanism is needed for changing the scale of observation when processing image data by automated analysis methods.

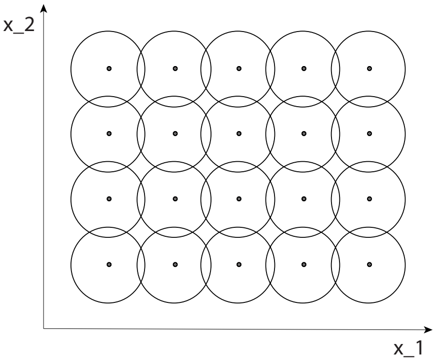


Figure 2: When sampling image data from the real world, the distribution of continuous image intensities must be integrated over non-infinitesimal regions over space and some non-infinitesimal amount of time. This figure gives a schematic illustration of the spatial support regions corresponding to the application of similar spatial window functions over a uniform rectangular grid in space, for which the spatial extent of these window functions determines the spatial inner scale of observation. In a corresponding manner, the image intensities do also have to be integrated for some non-infinitesimal amount of time, thus defining the temporal inner scale of the image measurement. These two inner scale levels determined by the image sampling process are, however, usually not the best scales for computing image features from the data as a basis for analyzing the image contents by automated computer vision or image analysis methods. For this reason, a principled theoretical framework is needed for changing the level of observation in real-world image data.

The world around us consists of different types of image structures at different scales. For example, for a crowd of people, we may at a coarse scale perceive the crowd as a type of texture, where different parts of the persons in the crowd constitute the texture elements. If we then look at some individual at a finer scale, we can expect that finer details will become visible, such as the eyes, the nose, the mouth of a face or even finer scale substructures of these (see figure 1).

Therefore, qualitatively different types of descriptions of image data may be warranted depending on the scale of observation and the types of image structures we are analyzing. This need is well understood, for example, in cartography, where maps are produced at different degrees of abstraction. A map of the world contains the largest countries and islands, and possibly, some of the major cities, whereas towns and smaller islands appear at first in a map of a country. In a city guide, the level of abstraction is changed considerably to include streets and buildings, etc. An atlas can be seen as a symbolic multi-scale representation of the world around us, constructed manually, and with very specific purposes in mind.

In physics, phenomena are modelled in different ways depending on the scale of the phenomena that are of interest, ranging from particle physics and quantum mechanics at the finest scales, through thermodynamics and solid mechanics dealing with every-day phenomena, to astronomy and relativity theory at scales much larger than those we are usually dealing with. Notably, a physical description may depend strongly upon the scale at which the world is modelled. This is in clear contrast to certain idealized mathematical entities, such as a "point" or a "line", which appear in the same way independent upon the scale of observation.

Specifically, the need for multi-scale representation of the data arises when designing methods for automatically analyzing and deriving information from images or signals that are the result of real-world measurements. It is clear that to be able to extract any type of information from data it is necessary to interact with it using some operators. The type of information that can be obtained is to a large extent 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 image processing and computer vision 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 data can be very hard.

In certain controlled situations, appropriate scales for analysis may be known *a priori*. For example, a characteristic property of a good physicist is his intuitive ability to select proper scales for modelling a problem. Under other circumstances, for example, in applications dealing with automated image processing, however, it may not at all be obvious to determine what are the proper scales. One such example is a vision system with the task of analyzing unknown scenes (see figure 3 for an illustration). Therefore, a theoretically well-founded methodology is needed for modelling the notion of scale in image data, and if needed, also performing the analysis at a different scale than the data was sampled at.

For images of *a priori* unknown objects there is usually no way to know in advance what scale levels are suitable or best for analyzing the image data. For images defined from two-dimensional spatial projections of the surrounding three-dimensional world, the perspective mapping implies that the same object may appear at different scales depending on the distance between the camera and the object. Therefore, an uncommitted approach to this problem consists of allowing for the image data to be analyzed at any level of scale, or alternatively stated at all scales simultaneously. In computer vision and image processing, this problem has been addressed by the notion of multi-scale representations such as pyramids or scale-space and in signal processing or numerical analysis by wavelets.

When constructing a multi-scale representation one may ask formally what types of image operations would be suitable for computing an image representation at a coarser scale from the original image data: Would any type of smoothing operation be permissible? Of crucial importance when constructing a multi-scale representation is that the smoothing operation does not introduce "new" spurious image structures in the image representations at coarse scales that do not correspond to simplifications of corresponding image structures at finer scales. How should such operations be formalized in a theoretically well-founded

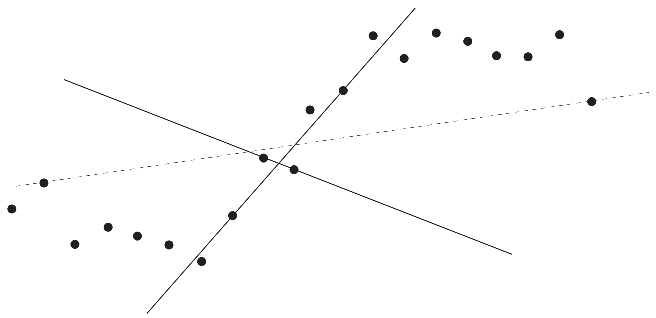


Figure 3: Illustration of the basic scale problem when computing derivative approximations as a basis for *e.g.* edge detection in computer vision or image processing. Assume that the dots represent (noisy) grey-level values along an imagined cross-section of an object boundary, and that the task is to find the boundary of the object. The lines show the effect of computing derivative approximations using a central difference operator with varying step size. There do also exist more refined approaches to gradient estimation, but they will also face similar problems. Clearly, only a certain interval of step sizes is appropriate for extracting the major slope of the signal corresponding to the object boundary. Of course, this slope may also be interpreted as due to noise (or some other phenomena that should be neglected) if it is a part superimposed onto some coarser-scale structure (not visible here). Therefore, even an as apparently simple problem as of finding a slope in measurement data is scale dependent. More generally, a similar type of scale problem will be present for any model for visual operations that is expressed in terms of derivatives of visual data.

manner?

An approach that has been taken in the area of scale-space theory is by imposing structural requirements on types of processing stages that are allowed and restricting the class of image operations in an axiomatic manner (Iijima 1962, Witkin 1983, Koenderink 1984, Lindeberg 1990, Koenderink & van Doorn 1992, Lindeberg 1994*b*, Lindeberg 1994*a*, Lindeberg 1996, Sporring et al. 1996, Florack 1997, Weickert et al. 1999, ter Haar Romeny 2003, Duits et al. 2004). The subject of this article is to describe a recent generalization of this theory (Lindeberg 2011) that encompasses scale-space representations for spatial and spatio-temporal image data in a unified manner and gives rise to image operations that are qualitatively very similar to receptive fields measured by cell recordings in biological vision.

3 Structural assumptions of scale-space theory

The notion of a *visual front-end* refers to a set of processes at the first stages of visual processing, which are assumed to be of a general nature and whose output can be used as input to different later stage processes, without being too specifically adapted to a particular task that would limit the applicability to other tasks. Major arguments for the definition of a visual front-end are that the first stages of visual processing should be as *uncommitted* as possible and allow initial processing steps to be *shared* between different later-stage visual modules, thus implying a *uniform structure* on the first stages of computations (Koenderink et al. 1992) (Lindeberg 1994*b*, section 1.1).

Basic assumptions underlying the formulation of scale-space theory are that:

- The image data arise from projections of a structured 3-D world, with basic symmetry properties under:

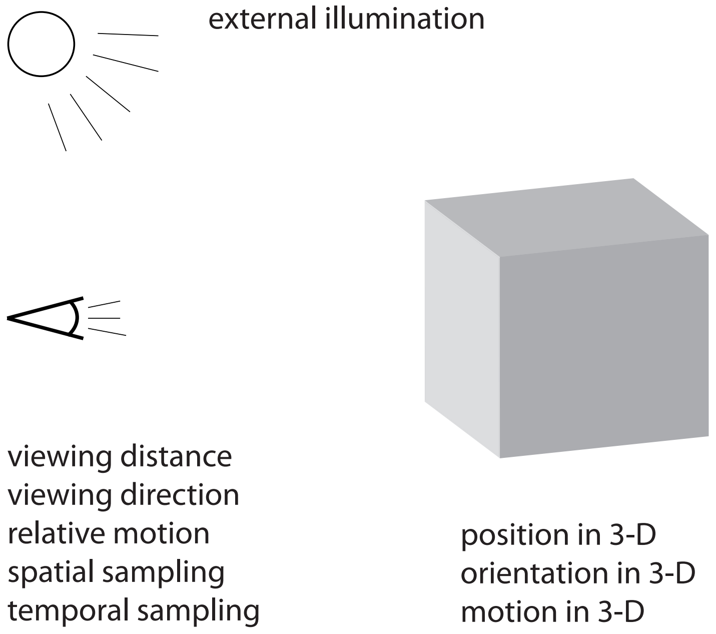


Figure 4: Basic factors that influence the formation of images for a two-dimensional camera that observes objects in the three-dimensional world. In addition to the position, orientation and motion of the object in 3-D, the perspective projection onto the image plane is affected by the viewing distance, viewing direction and relative motion of the camera in relation to the object, the spatial and temporal sampling characteristics of the image sensor as well the usually unknown external illumination field in relation to the geometry of the scene and the camera.. A main goal of the generalized scale-space theory presented in this article is to provide a theoretical framework for handling the interaction between these inherent variabilities of the image formation process and the image operators that are to be used for computing image features from the measured image data. (The effect of illumination variations on receptive field measurements is, however, not explicitly treated in this article; see (Lindeberg 2013*a*, Lindeberg 2013*b*) for a general theoretical analysis regarding this matter.)

  - translations and rotations of objects in the 3-D environment,
  - different distances to the camera,
  - relative motion velocities between the camera and the observer and
  - illumination variations.

- Visual observations are performed with a non-infinitesimal aperture function (probe) which must be taken into explicit account in subsequent analysis of the image data. Specifically, different scale levels than used for sampling the image data are usually needed when analyzing the data.
- Real-world objects may appear in different ways depending on the scale of observation.
- Image representations at coarser scales should correspond to simplifications of image representations at finer scales.

For a computerized vision system or a biological vision to behave in a stable manner when exposed to natural image data, we would like objects to be perceived or described in a similar way under such basic image transformations (see figure 4 for an illustration).

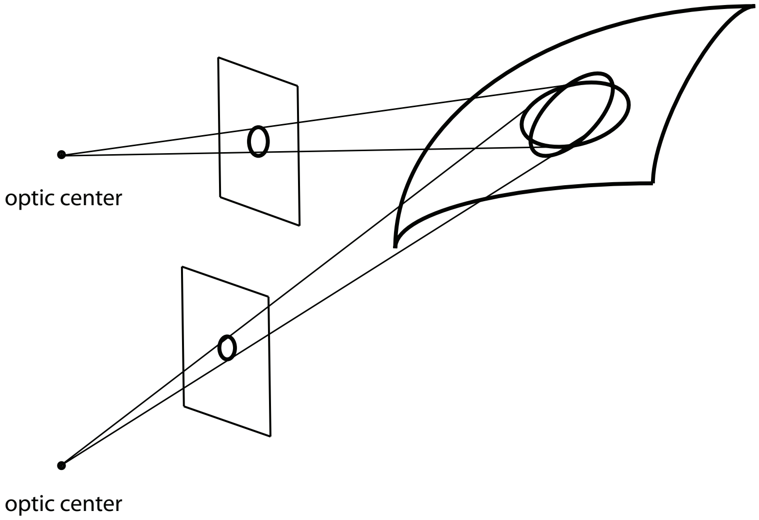


Figure 5: Consider a method for computing image features from image data that is based on rotationally symmetric image operations over the spatial image domain. If such a method is applied to two different images of a three-dimensional object that have been taken from different viewing directions, then the backprojections of the image operations onto the surface of the object will in general correspond to different regions in physical space over which corresponding information will be weighed differently. If such image features are to be used for *e.g.* image matching or object recognition, then there will be a systematic error caused by the mismatch between the backprojections of the receptive fields from the image domain onto the world. By requiring the family of receptive fields to be covariant under local affine image deformations, it is possible to reduce this amount of mismatch such that the backprojected receptive fields can be made similar when projected onto the tangent plane of the surface by local linearizations of the perspective mapping. Corresponding effects occur when analyzing spatio-temporal image data (video) based on receptive fields that are restricted to being space-time separable only. If an object is observed over time by two cameras having different relative motions between the camera and the observer, then the corresponding receptive fields cannot be matched unless the family of receptive fields possesses sufficient covariance properties under local Galilean transformations.

At a high level of abstraction, one may be interested in constructing *invariant* image features that can be used *e.g.* for recognizing objects from different viewpoints and whose numerical values will be equal or only moderately affected by basic image transformations. At a lower level of abstraction, a weaker condition is to require the image operations to be *covariant* under basic image transformations, implying that the image features are not truly invariant but nevertheless transform in a predictable and well-understood manner under basic image transformations. If the underlying image operations would not be covariant, then there would be a systematic bias in the visual operations, corresponding to the amount of mismatch between the backprojections of the image operations to the world corresponding to *e.g.* two images of the same physical object taken from different views or an object that moves with different velocity relative to the observer for two observations having significant extent over both space and time (see figure 5 for an illustration).

Regarding the types of image transformations, the non-linear perspective mapping implies that rigid translations and rotations, which correspond to linear operations in 3-D, give rise to non-linear transformations in the image plane. To simplify the analysis, we shall therefore approximate these transformations by local linearizations at any image point, implying that the perspective mapping will be approximated by local affine transformations and that relative motions between the camera and the observer will be approximated by local Galilean transformations.

A main subject of scale-space theory is to provide a formal and theoretically well-founded framework for handling image structures at different scales that is consistent with such structural requirements corresponding to symmetry properties of the environment in the cases of purely spatial, spatio-chromatic or spatio-temporal image data, respectively.

4 Scale-space axioms for spatial image domains

In the following, we will describe a set of structural requirements that can be stated on early visual operations regarding (i) spatial geometry, (ii) the image measurement process with its close relationship to the notion of scale and (iii) internal representations of image data that are to be computed by a general purpose vision system. In section 6 this treatment will then be extended with complementary requirements that arise due to (iv) the special nature of time and structural requirements concerning (v) spatio-temporal geometry.

4.1 Structural scale-space axioms

Let us initially restrict ourselves to static (time-independent) data and focus on the spatial aspects: If we regard the incoming image intensity f as defined on an N-dimensional continuous image plane $f \colon \mathbb{R}^N \to \mathbb{R}$ with Cartesian image coordinates denoted by $x = (x_1, \dots, x_N)^T$, then the problem of defining a set of early visual operations can be formulated as finding a family of operators $\mathcal{T}_s$ that are to act on f to produce a family of new intermediate image representations[1]

$$L(\cdot;\ s) = \mathcal{T}_s\, f(\cdot) \tag{2}$$

which are also to defined as functions on $\mathbb{R}^N$, *i.e.*, $L(\cdot;\ s) \colon \mathbb{R}^N \to \mathbb{R}$. These intermediate representation may be dependent on some parameter $s \in \mathbb{R}^M$, which in the simplest case may be one-dimensional or under more general circumstances multi-dimensional.

On a spatial domain where the smoothing operation is required to be rotationally symmetric, a one-dimensional parameter s with $M = 1$ may be regarded as sufficient, whereas a higher dimensionality of the parameter s is needed to account for different amounts of smoothing along different directions in space as will be needed in the presence of general affine image transformations.

Linearity. If we want these initial visual processing stages to make as few irreversible decisions as possible, it is natural to initially require $\mathcal{T}_s$ to be a *linear operator* such that

$$\mathcal{T}_s(a_1 f_1 + a_2 f_2) = a_1 \mathcal{T}_s f_1 + a_2 \mathcal{T}_s f_2 \tag{3}$$

holds for all functions $f_1, f_2 \colon \mathbb{R}^N \to \mathbb{R}$ and all scalar constants $a_1, a_2 \in \mathbb{R}$.

Linearity also implies that a number of special properties of receptive fields (to be developed below) will transfer to spatial derivatives of these and do therefore imply that different types of image structures will be treated in a similar manner irrespective of what types of linear filters they are captured by.

[1] In equation (2), the symbol "·" at the position of the first argument of L is a place holder to emphasize that in this relation, L is regarded as a function and not evaluated with respect to its first argument x. The following semi-colon emphasizes the different natures of the image coordinates x and the filter parameters s.

Derivative operators are essential for modelling visual operations, since they respond to relative differences between image intensities in a local neighbourhood and are therefore less sensitive to illumination variations than zero-order (undifferentiated) image intensities (see (Lindeberg 2013*a*, section 2.3) for a precise statement).

Translational invariance. Let us also require $\mathcal{T}_s$ to be a *shift-invariant operator* in the sense that it commutes with the shift operator $\mathcal{S}_{\Delta x}$ defined by $(\mathcal{S}_{\Delta x} f)(x) = f(x - \Delta x)$, such that

$$\mathcal{T}_s \left(\mathcal{S}_{\Delta x} f \right) = \mathcal{S}_{\Delta x} \left(\mathcal{T}_s f \right) \tag{4}$$

holds for all $\Delta x \in \mathbb{R}^N$. The motivation behind this assumption is the basic requirement that the representation of a visual object should be similar irrespective of its position in the image plane.[2] Alternatively stated, the operator $\mathcal{T}_s$ can be said to be *homogeneous across space*.

Convolution structure. From a general result in linear systems theory it follows from the assumptions of linearity and shift-invariance that the internal representations $L(\cdot;\ s)$ are given by *convolution transformations* (Hirschmann & Widder 1955)

$$L(x;\ s) = (T(\cdot;\ s) * f)(x) = \int_{\xi \in \mathbb{R}^N} T(\xi;\ s)\, f(x - \xi)\, d\xi \tag{5}$$

where $T(\cdot;\ s)$ denotes some family of convolution kernels. These convolution kernels and their spatial derivatives can also be referred to as (spatial) receptive fields.

Regularity. To be able to use tools from functional analysis, we will initially assume that both the original signal f and the family of convolution kernels $T(\cdot;\ s)$ are in the Banach space $L^2(\mathbb{R}^N)$, *i.e.*, that $f \in L^2(\mathbb{R}^N)$ and $T(\cdot;\ s) \in L^2(\mathbb{R}^N)$ with the norm

$$\|f\|_2^2 = \int_{x \in \mathbb{R}^N} |f(x)|^2\, dx. \tag{6}$$

Then, also the intermediate representations $L(\cdot;\ s)$ will be in the same Banach space and the operators $\mathcal{T}_s$ can be regarded as well-defined.

Positivity (non-negativity). Concerning the convolution kernels, one may require these to be non-negative in order to constitute smoothing transformations.

$$T(x;\ s) \geq 0. \tag{7}$$

Normalization. Furthermore, it may be natural to require the convolution kernels to be normalized to unit mass

$$\int_{x \in \mathbb{R}^N} T(x;\ s)\, dx = 1. \tag{8}$$

to leave a constant signal unaffected by the smoothing transformation.

[2] For a two-dimensional camera that observes objects in a three-dimensional world, translational invariance may be more natural to express with respect to a spherical camera geometry, since then the image representation will be independent of the viewing direction of the camera. In this presentation, we will, however, restrict ourselves to a planar camera geometry, since the algebraic modelling is simpler with such a model, which can also be regarded as a reasonable approximation in the central field of view.

Quantitative measurement of the spatial extent and the spatial offset of non-negative scale-space kernels. For a non-negative convolution kernel, we can measure its spatial offset by the mean operator

$$m = \bar{x} = M(T(\cdot;\ s)) = \frac{\int_{x \in \mathbb{R}^N} x\, T(x;\ s)\, dx}{\int_{x \in \mathbb{R}^N} T(x;\ s)\, dx} \tag{9}$$

and its spatial extent by the spatial covariance matrix

$$\Sigma = C(T(\cdot;\ s)) = \frac{\int_{x \in \mathbb{R}^N} ((x - \bar{x})\,(x - \bar{x})^T\, T(x;\ s)\, dx}{\int_{x \in \mathbb{R}^N} T(x;\ s)\, dx}. \tag{10}$$

Using the additive properties of mean values and covariance matrices under convolution, which hold for non-negative distributions, it follows that

$$m = M(T(\cdot;\ s_1) * T(\cdot;\ s_2)) = M(T(\cdot;\ s_1)) + M(T(\cdot;\ s_2)) = m_1 + m_2, \tag{11}$$

$$\Sigma = C(T(\cdot;\ s_1) * T(\cdot;\ s_2)) = C(T(\cdot;\ s_1)) + C(T(\cdot;\ s_2)) = \Sigma_1 + \Sigma_2 \tag{12}$$

4.2 Simplifying image structures over scale

The reduction of the first stage of visual processing to a set of convolution transformations raises the question of what types of convolution kernels $T(\cdot;\ s)$ should be regarded as natural?

The issue of scale. A fundamental property of the convolution operation is that it may reflect different types of image structures depending on the spatial extent (the width) of the convolution kernel:

- Convolution with a *large support* kernel will have the ability to respond to phenomena at *coarse scales.*
- A kernel with *small support* will on the other hand be necessary capture phenomena at *fine scales*.

From this viewpoint, it is natural to associate an interpretation of *scale* with every image measurement. Let us therefore assume that the parameter s represents such a scale attribute and let us assume that for a one-dimensional scale parameter the scale parameter should always be non-negative $s \in \mathbb{R}_+$ whereas for a multi-dimensional scale parameter there should exist some mapping from the real-valued multi-dimensional scale parameter $s \in \mathbb{R}^M$ to some non-negative quantification of the notion of scale.

Identity operation with continuity. To guarantee that the limit case of the internal scale-space representations when the scale parameter s tends to zero should correspond to the original image data f, we will assume that

$$\lim_{s \downarrow 0} L(\cdot;\ s) = \lim_{s \downarrow 0} \mathcal{T}_s f = f. \tag{13}$$

Hence, the intermediate image representations $L(\cdot;\ s)$ can be regarded as a family of derived representations parameterized by a scale parameter s. With $s = (s_1, \dots, s_M)$ representing a multi-dimensional scale parameter $s \in \mathbb{R}^M$, equation (13) should be interpreted as $\lim_{|s| \downarrow 0} L(\cdot;\ s) = \lim_{|s| \downarrow 0} \mathcal{T}_s f = f$ with $|s| = \sqrt{s_1^2 + \dots + s_M^2}$.

Semi-group structure. For such image measurements to be properly related *between* different scales, it is natural to require the operators $\mathcal{T}_s$ with their associated convolution kernels $T(\cdot;\ s)$ to form a *semi-group*[3] over s

$$\mathcal{T}_{s_1}\,\mathcal{T}_{s_2} = \mathcal{T}_{s_1+s_2} \tag{14}$$

with a corresponding semi-group structure for the convolution kernels

$$T(\cdot;\ s_1) * T(\cdot;\ s_2) = T(\cdot;\ s_1+s_2). \tag{15}$$

Then, the transformation between any different and ordered[4] scale levels s_1 and s_2 with $s_2 \geq s_1$ will obey the *cascade property*

$$L(\cdot;\ s_2) = T(\cdot;\ s_2-s_1) * T(\cdot;\ s_1) * f = T(\cdot;\ s_2-s_1) * L(\cdot;\ s_1) \tag{16}$$

i.e. a similar type of transformation as from the original data f. An image representation having these properties is referred to as a (spatial) *multi-scale representation*.

Note the close similarity between the additive structure of scale parameters obtained in this way with the additive structure of mean values (11) and covariance matrices (12) under convolution of non-negative convolution kernels.

Self-similarity over scale. Regarding the family of convolution kernels used for computing a multi-scale representation, it is also natural to require them to *self-similar over scale*, such that if $s \in \mathbb{R}$ is a one-dimensional scale parameter then all the kernels correspond to rescaled copies

$$T(x;\ s) = \frac{1}{\varphi(s)}\bar{T}\left(\frac{x}{\varphi(s)}\right) \tag{17}$$

of some prototype kernel $\bar{T}$ for some transformation[5] $\varphi(s)$ of the scale parameter. If $s \in \mathbb{R}_+^M$ is a multi-dimensional scale parameter, the requirement of self-similarity over scale can be

[3] Concerning the parameterization of this semi-group, we will in the specific case of a one-dimensional (scalar) scale parameter assume the parameter $s \in \mathbb{R}$ to have a direct interpretation of scale, whereas in the case of a multi-dimensional parameter $s = (s_1, \dots, s_M) \in \mathbb{R}^M$, these parameters could also encode for other properties of the convolution kernels in terms of the orientation θ in image space or the degree of elongation $e = \sigma_1/\sigma_2$, where σ_1 and σ_2 denote the spatial extents in different directions. The convolution kernels will, however, not be be required to form a semi-group over any type of parameterization, such as the parameters θ or e. Instead, we will assume that there exists some parameterization s for which an additive linear semi-group structure can be defined and from which the latter types of parameters can then be derived.

[4] With $s_1 = (s_{1,1}, \dots, s_{1,M})$ and $s_2 = (s_{2,1}, \dots, s_{2,M})$ denoting two M-dimensional scale parameters, the inequality $s_2 \geq s_1$ should be interpreted as a requirement that the scale levels s_1 and s_2 have to be *ordered* in the sense that the increment $u = s_2 - s_1$ should correspond to a *positive direction* in parameter space that can be interpreted as increasing levels of scale. For example, for the affine Gaussian scale-space concept $L(x;\ \Sigma)$ to be considered later in section 5.3, for which the scale parameter over a two-dimensional spatial image domain can be parameterized by positive semi-definite 2×2 covariance matrices Σ, the requirement of an ordered and positive scale direction u between the scale-space representations computed for two different covariance matrices Σ_1 and Σ_2 means that the difference between these covariance matrices $\Sigma_u = \Sigma_2 - \Sigma_1$ must be positive semi-definite. With the corresponding multi-dimensional scale parameters s_1 and s_2 expressed as vectors $s_1 = (\Sigma_{1,11}, \Sigma_{1,12}, \Sigma_{1,22})$ and $s_2 = (\Sigma_{2,11}, \Sigma_{2,12}, \Sigma_{2,22})$ where $\Sigma_{k,ij}$ denote the elements of Σ_k for $k = 1$ and 2, the condition for $u = (u_1, u_2, u_3) = s_2 - s_1$ to correspond to a positive direction in parameter space can therefore be expressed as $u_1 u_3 - u_2^2 \geq 0$ and $u_1 + u_3 \geq 0$.

[5] The reason for introducing a function φ for transforming the scale parameter s into a scaling factor $\varphi(s)$ in image space, is that the requirement of a semi-group structure (14) does not imply any restriction on how the parameter s should be related to image measurements in dimensions of length — the semi-group structure only implies an abstract ordering relation between coarser and finer scales $s_2 > s_1$ that could also be satisfied for any monotonously increasing transformation of the parameter s. For the Gaussian scale-space concept having a scalar

generalized into

$$T(x;\ s) = \frac{1}{|\det \varphi(s)|} \bar{T}\left(\varphi(s)^{-1} x\right) \tag{18}$$

where $\varphi(s)$ now denotes a non-singular $N \times N$-dimensional matrix regarding an N-dimensional image domain and $\varphi(s)^{-1}$ its inverse. With this definition, a multi-scale representation having a scalar scale parameter $s \in \mathbb{R}_+$ will be based on uniform rescalings of the prototype kernel, whereas a multi-scale representation based on a multi-dimensional scale parameter might also allow for rotations as well as non-uniform affine deformations of the prototype kernel.

Together, the requirements of a semi-group structure and self-similarity over scales imply that the parameter s gets both a (i) *qualitative* interpretation of the notion of scale in terms of an abstract *ordering relation* due to the cascade property in equation (16), and (ii) a *quantitative* interpretation of scale in terms of the *scale-dependent spatial transformations* in equations (17) and (18). When these conditions are simultaneously satisfied, we say that the intermediate representation $L(\cdot;\ s)$ constitutes a candidate for being regarded as a *weak scale-space representation*.

Infinitesimal generator. For theoretical analysis it preferably if we can treat the scale parameter s as a continuous parameter and if image representations at adjacent scales can be related by partial differential equations. Such relations can be expressed if the semi-group possesses an *infinitesimal generator* (Hille & Phillips 1957, page 308) (Pazy 1983, page 5)

$$\mathcal{B}L = \lim_{h \downarrow 0} \frac{T(\cdot;\ h) * f - f}{h} \tag{19}$$

and imply that the image representations at adjacent scales can be related by an evolution equation of the form

$$\partial_s L(x;\ s) = (\mathcal{B}L)(x;\ s) \tag{20}$$

where we would preferably like the operator $\mathcal{B}$ to be a partial differential operator. The set of elements $f \in L^2(\mathbb{R}^N)$ for which $\mathcal{B}$ exists is denoted $D(\mathcal{B})$. This set is not empty and never reduces to the zero element. Actually, $D(\mathcal{B})$ is even dense in $L^2(\mathbb{R}^N)$ (Hille & Phillips 1957, page 308) (Pazy 1983, page 5).

In equation (20), we have for simplicity assumed the scale parameter s to be a scalar (one-dimensional) parameter. For a multi-parameter scale-space having a scale parameter of the form $s = (s_1, \dots, s_M)$, an analogous concept can be defined in terms of the *directional derivative of the semi-group* along any *positive direction* $u = (u_1, \dots, u_M)$ in the parameter space

$$(\mathcal{D}_u L)(x;\ s) = (\mathcal{B}(u)\, L)(x;\ s) = (u_1 \mathcal{B}_1 + \dots + u_M \mathcal{B}_M)\ L(x;\ s) \tag{21}$$

where each $\mathcal{B}_k$ $(k = 1 \dots M)$ constitutes the infinitesimal generator for the parameter s_k along the unit direction e_k in the M-dimensional parameter space

$$\mathcal{B}_k L = \lim_{h \downarrow 0} \frac{T(\cdot;\ h\, e_k) * f - f}{h} \tag{22}$$

scale parameter according to equations (36)–(37) this transformation is given by $\sigma = \varphi(s) = \sqrt{s}$, whereas for the affine Gaussian scale-space concept according to equation (50) it is given by the matrix square root function $\varphi(s) = \Sigma^{1/2}$, where Σ denotes the covariance matrix that describes the spatial extent and the orientation of the affine Gaussian kernel.

and with the notion of a "positive direction" in parameter space defined in a similar way as in footnote 4.

In (Lindeberg 2011) it is shown how such differential relationships can be ensured given a proper selection of functional spaces and sufficient regularity requirements over space x and scale s in terms of Sobolev norms. We shall therefore henceforth regard the internal representations $L(\cdot;\ s)$ as differentiable with respect to the scale parameter(s).

4.2.1 Non-creation of new image structures with increasing scale

A further requirement on a scale-space representation is that convolution with a scale-space kernel $T(\cdot;\ s)$ should correspond to *smoothing transformation* in the sense that coarser scale representations should be guaranteed to constitute *simplifications* of corresponding finer scale representations. This means that new image structures must not be created at coarser scales $L(\cdot;\ s)$ that do not correspond to simplifications of corresponding structures in the original data f.

Non-creation of local extrema (zero-crossings). For one-dimensional signals $f\colon \mathbb{R} \to \mathbb{R}$, such a condition can be formalized as the requirement that the number of local extrema in the data must not increase with scale for any signal and is referred to as *non-creation of local extrema*. Formally, a one-dimensional kernel T is a scale-space kernel if for any signal f, the number of local extrema in $T * f$ is guaranteed to not exceed the number of local extrema in f (Lindeberg 1990). It can be shown that for a one-dimensional signal, this condition can also be equivalently expressed in terms of zero-crossings.

For higher-dimensional signals, however, this condition cannot be applied, since it can be shown that there are no non-trivial linear transformations that are guaranteed to never increase the number of local extrema in an image (Lifshitz & Pizer 1990) (Lindeberg 1990) (Lindeberg 1994*b*, Chapter 4, pages 101–103).

Non-enhancement of local extrema. A particularly useful way of formalizing the requirement of non-creation of new image structures with increasing scale is that *local extrema must not be enhanced with increasing scale*. In other words, if a point $(x_0;\ s_0)$ is a local (spatial) maximum of the mapping $x \mapsto L(x;\ s_0)$ then the value must not increase with scale. Similarly, if a point $(x_0;\ s_0)$ is a local (spatial) minimum of the mapping $x \mapsto L(x;\ s_0)$, then the value must not decrease with scale. Given the above mentioned differentiability property with respect to scale, we say that the multi-scale representation constitutes a *scale-space representation* if it for a scalar scale parameter satisfies the following conditions (Lindeberg 1996):

$$\partial_s L(x_0;\ s_0) \leq 0 \qquad \text{at any non-degenerate local maximum,} \tag{23}$$

$$\partial_s L(x_0;\ s_0) \geq 0 \qquad \text{at any non-degenerate local minimum,} \tag{24}$$

or for a multi-parameter scale-space

$$(\mathcal{D}_u L)(x_0;\ s_0) \leq 0 \qquad \text{at any non-degenerate local maximum,} \tag{25}$$

$$(\mathcal{D}_u L)(x_0;\ s_0) \geq 0 \qquad \text{at any non-degenerate local minimum,} \tag{26}$$

for any positive direction $u = (u_1, \ldots, u_M)$ in the parameter space (see figure 6).

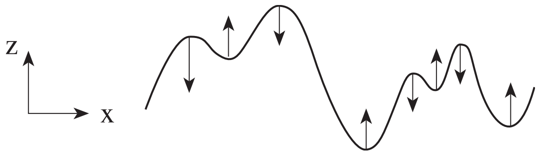


Figure 6: The requirement of non-enhancement of local extrema is a way of restricting the class of possible image operations by formalizing the notion that new image structures must not be created with increasing scale, by requiring that the value at a local maximum must not increase and that the value at a local minimum must not decrease.

Basic implications of the requirements of non-creation of structure. For one-dimensional signals, it can be shown that the requirement of non-creation of local extrema implies that a scale-space kernel must be positive and unimodal, both in the spatial domain and the Fourier domain (Lindeberg 1990). By considering the response to a constant signal, it furthermore follows from the requirement of non-enhancement of local extrema that a scale-space kernel should be normalized to constant L_1-norm (equation (8)).

4.3 Covariance requirements regarding spatial domains

Scale covariance. A basic requirement on a scale-space representation is that it should be able to handle rescalings in the image domain in a consistent manner. If the input image is transformed by a uniform scaling factor

$$f' = \mathcal{S} \, f \quad \text{corresponding to} \quad f'(x') = f(x) \quad \text{with} \quad x' = S \, x, \tag{27}$$

then there should exist some transformation of the spatial scale parameter $s' = S(s)$ such that the corresponding scale-space representations are equal (see figure 7):

$$L'(x';\; s') = L(x;\; s) \quad \text{corresponding to} \quad \mathcal{T}_{S(s)} \, \mathcal{S} \, f = \mathcal{S} \, \mathcal{T}_s \, f. \tag{28}$$

$$\begin{tikzcd}
\mathcal{T}_s \, f \arrow[r, "\mathcal{S}"] & L \\
f \arrow[u, "\mathcal{T}_s"'] \arrow[r, "\mathcal{S}"] & \mathcal{S} \, f \arrow[u, "\mathcal{T}_{S(s)}"']
\end{tikzcd}$$

Figure 7: Commutative diagram for scale-space representations computed under uniform scalings of the spatial domain. Such a spatial rescaling transformation may, for example, represent image data that have been acquired using visual sensors that sample the image data with different resolution or an object that is observed with different distances between the camera and the object.

Rotational covariance. If we restrict ourselves to a scale-space representation based on a scalar (one-dimensional) scale parameter $s \in \mathbb{R}_+$, then it is natural to require the scale-space kernels to be *rotationally symmetric*

$$T(x;\; s) = h(\sqrt{x_1^2 + \dots + x_N^2};\; s) \tag{29}$$

for some one-dimensional function $h(\cdot;\; s) \colon \mathbb{R} \to \mathbb{R}$. Such a symmetry requirement can be motivated by the requirement that in the absence of further information, all spatial directions should be equally treated (isotropy).

For a scale-space representation based on a multi-dimensional scale parameter, one may also consider a weaker requirement of rotational invariance at the level of the family of kernels, for example regarding a set of elongated kernels with different orientations in image space. Then, the family of kernels may capture image data of different orientation in a rotationally invariant manner, for example if all image orientations are explicitly represented or if the receptive fields corresponding to different orientations in image space can be related by linear combinations.

Affine covariance. The perspective mapping from the three-dimensional world to the two-dimensional image space gives rise to image deformations in the image domain. If we approximate the non-linear perspective mapping from a surface pattern in the world to the image plane by a local linear transformation (the derivative), then we can model this deformation by a local *affine transformation*

$$f' = \mathcal{A}\, f \quad \text{corresponding to} \quad f'(x') = f(x) \quad \text{with} \quad x' = A\, x + b. \tag{30}$$

A natural requirement on a vision system that observes objects whose projections on the image plane are being deformed in different ways depending on the viewing conditions, is that the vision system should be able to relate or match the different internal representations of external objects that are acquired under different viewing directions. Such a requirement is natural to enable a stable interpretation of objects in the world under variations of the orientation of the object relative to the observer.

To ensure that the internal representations behave nicely under image deformations, it is therefore natural to require a possibility of relating them under affine transformations

$$L'(x';\ s') = L(x;\ s) \quad \text{corresponding to} \quad \mathcal{T}_{A(s)}\, \mathcal{A}\, f = \mathcal{A}\, \mathcal{T}_s\, f \tag{31}$$

for some transformation $s' = A(s)$ of the scale parameter.

Within the class of linear operators $\mathcal{T}_s$ over a two-dimensional image domain, it is, however, not possible to realize such an affine covariance property over the full group of affine transformations within a scale-space concept based on a scalar scale parameter only. For two-dimensional image data, such affine covariance can, however, be accomplished within a three-parameter linear scale-space, which will be referred to as affine scale-space. The notions of scale covariance and rotational covariance can, however, be obtained based on a one-parameter spatial scale-space.

$$\begin{tikzcd}
\mathcal{T}_s\, f \arrow[r, "\mathcal{A}"] & L \\
f \arrow[u, "\mathcal{T}_s"'] \arrow[r, "\mathcal{A}"] & \mathcal{A}\, f \arrow[u, "\mathcal{T}_{A(s)}"']
\end{tikzcd}$$

Figure 8: Commutative diagram for scale-space representations computed under affine deformations of image space. Such an affine transformation may, for example, represent a local linear approximation of the projective mapping between two different perspective projections of a local surface patch.

5 Scale-space concepts for spatial image domains

5.1 General necessity result concerning Gaussian scale-spaces

Given the above mentioned requirements it can be shown that if we assume (i) linearity, (ii) shift-invariance over space, (iii) semi-group property over scale, (iv) sufficient regularity properties over space and scale in terms of Sobolev norms[6] and (v) non-enhancement of local extrema to hold for *any* smooth image function $f \in C^{\infty}(\mathbb{R}^N) \cap L^1(\mathbb{R}^N)$, then the scale-space representation over an N-dimensional spatial domain must satisfy (Lindeberg 2011, theorem 5, page 42)

$$\partial_s L = \frac{1}{2} \nabla_x^T \left(\Sigma_0 \nabla_x L \right) - \delta_0^T \nabla_x L \tag{32}$$

for some $N \times N$ covariance matrix Σ_0 and some N-dimensional vector δ_0 with $\nabla_x = (\partial_{x_1}, \dots, \partial_{x_N})^T$.

In terms of convolution kernels, this corresponds to convolutions with gradually growing elongated Gaussian kernels, which translate with drift velocity δ_0 with respect to the evolution parameter s. In terms of filtering operations, this scale-space can equivalently be constructed by convolution with *affine and translated Gaussian kernels*

$$g(x;\ \Sigma_s, \delta_s) = \frac{1}{(2\pi)^{N/2} \sqrt{\det \Sigma_s}} \, e^{-(x-\delta_s)^T \Sigma_s^{-1} (x-\delta_s)/2}, \tag{33}$$

which for a given $\Sigma_s = s \Sigma_0$ and a given $\delta_s = s \delta_0$ satisfy the diffusion equation (32). The Fourier transform of this shifted Gaussian kernel is

$$\hat{g}(\omega;\ \Sigma_s, \delta_s) = \int_{x \in \mathbb{R}^N} g(x;\ \Sigma_s, \delta_s) \, e^{-i \omega^T x} \, dx = e^{i \omega^T \delta_s - \omega^T \Sigma_s \omega / 2}. \tag{34}$$

From the diffusion equation formulation or the Fourier transform, it can be seen that these shifted and shape-adapted kernels satisfy the following generalized semi-group property

$$g(\cdot;\ \Sigma_1, v_1) * g(\cdot;\ \Sigma_2, v_2) = g(\cdot;\ \Sigma_1 + \Sigma_2, v_1 + v_2). \tag{35}$$

If we in addition require the convolution kernels to be *mirror symmetric* through the origin $T(-x;\ s) = T(x;\ s)$ then the offset vector δ_0 must be zero.

This formulation of the Gaussian scale-space representation L in terms of the diffusion equation (32) means that it is possible to interpret the intensity values of the input image f as a "temperature distribution" in the image plane and that the process that generates the scale-space representation as a function of the scale parameter s corresponds to heat diffusion in the image plane over virtual diffusion time s assuming that the thermal conductivity of the material equal to the arbitrarily chosen constant $1/2$ in the isotropic case when the covariance matrix Σ is equal to the unit matrix I. The covariance matrix Σ then describes how the thermal conductivity is modified in different directions in image space for heat diffusion

[6] To ensure sufficient differentiability properties such that an infinitesimal generator exists and the resulting multi-scale representation obtained by convolution with the semi-group of convolution kernels can be differentiated with respect to both space and scale such that the requirement of non-enhancement of local extrema can be applied, we do formally for an N-dimensional spatial domain require the semi-group $\mathcal{T}_s$ to be C_1-*continuous* such that $\lim_{h \downarrow 0} \left\| \frac{1}{h} \int_{s=0}^{h} \mathcal{T}(s) f \, ds - f \right\|_{H^k(\mathbb{R}^N)} = 0$ should hold for some $k > N/2$ and for all smooth functions $f \in L^1(\mathbb{R}^N) \cap C^{\infty}(\mathbb{R}^N)$ with $\| \cdot \|_{H^k(\mathbb{R}^N)}$ denoting the L^2-based Sobolev norm $\|u\|_{H^k(\mathbb{R}^N)} = \left(\int_{\omega \in \mathbb{R}^N} \left(1 + |\omega|^2\right)^k |\hat{u}(\omega)|^2 d\omega \right)^{1/2}$ and $\hat{u}$ denoting the Fourier transform of u over $\mathbb{R}^N$; see (Lindeberg 2011, Section 3.2 and Appendix A) regarding details.

in an anisotropic medium, and the term v represents a drift velocity with respect to virtual diffusion time s.

These relationships provide a general structure for linear scale-space concepts on shift-invariant continuous domains. Specifically, they comprise the following special cases:

5.2 Rotationally symmetric Gaussian scale-space

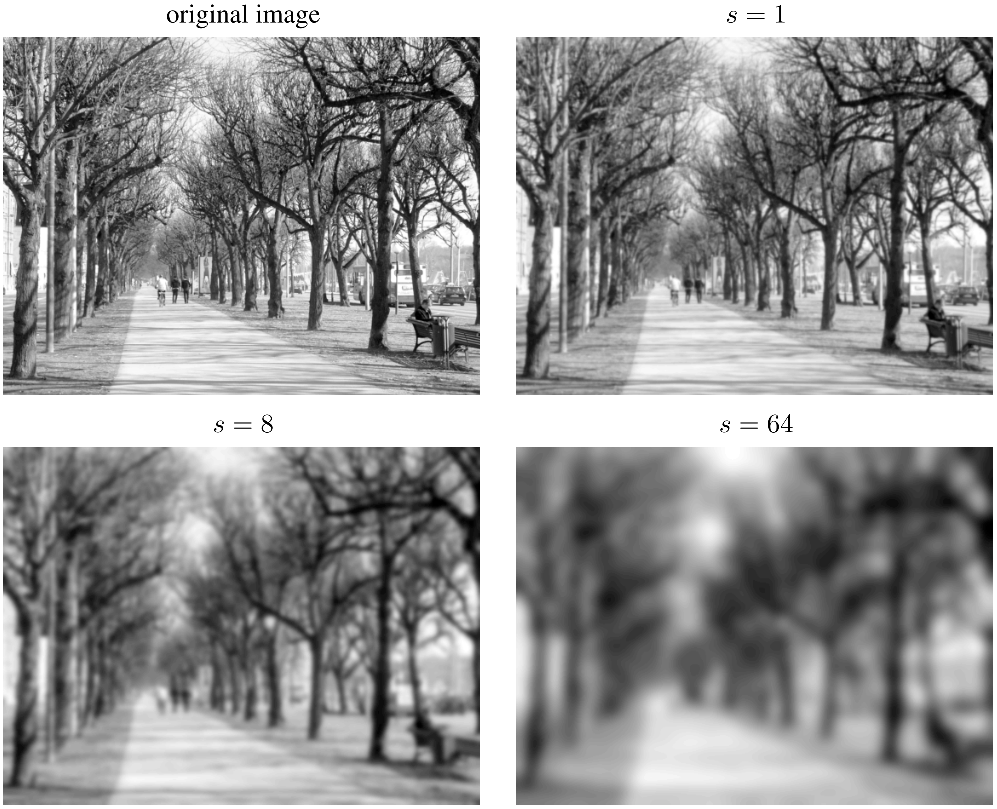


Figure 9: Illustration of the result of computing a rotationally symmetric scale-space representation of an image with perspective effects. Note how the Gaussian smoothing operation gives rise to a gradual suppression of finer-scale image structures, such that image structures having spatial extent smaller than $\sigma = \sqrt{s}$ have largely been suppressed in the scale-space representation at scale s.

If we require the scale-space representation generated by (32) or equivalently (33) to be rotationally symmetric, then it follows by necessity that the offsets δ_s and δ_0 must be zero and that the covariance matrices Σ_s and Σ_0 must be proportional to the unit matrix. Thus, the diffusion operator $\mathcal{A}$ will be proportional to the Laplacian operator, and the filter kernels will be rotationally symmetric Gaussians. In other words, this scale-space is obtained from

$$L(x;\ s) = \int_{\xi \in \mathbb{R}^N} f(x - \xi)\, g(\xi;\ s)\, d\xi \tag{36}$$

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

$$g(x;\ s) = \frac{1}{(2\pi s)^{N/2}}\, e^{-(x_1^2 + \dots + x_N^2)/2s} \tag{37}$$

Equivalently, this scale-space family can be obtained as the solution of the isotropic diffusion equation

$$\partial_s L = \frac{1}{2} \nabla^2 L \tag{38}$$

with initial condition $L(\cdot;\ 0) = f$. Earlier closely related necessity results regarding this representation in the rotationally symmetric case have been presented by (Koenderink 1984) based on the requirement that new level surfaces in scale-space must not be created with increasing scale (causality) in combination with isotropy and homogeneity, and in (Lindeberg 1996) based on a combination of a convolution semi-group structure with non-enhancement of local extrema. Another explicit necessity result in the one-dimensional case has also been given in (Lindeberg 1990, Theorem 5, page 241) (Lindeberg 1994*b*, Section 3.5.2, pages 89-91) based on a combination of a convolution semi-group structure with non-creation of local extrema with increasing scale, based on an earlier characterization of variation diminishing convolution transformations by (Schoenberg 1950); see also (Karlin 1968).

Figure 9 shows an illustration of computing different levels of a Gaussian scale-space representation for a grey-level image with significant perspective effects. Note how the Gaussian smoothing operation leads to gradual suppression of fine scale image structures, which can be used for separating image structures at different scales.

Gaussian derivative operators. From this scale-space representation, we can for any value of N (not necessarily coupled to the dimensionality of the signal) define the *multi-scale N-jet* by applying partial derivatives to the scale-space (Koenderink & van Doorn 1987, Koenderink & van Doorn 1992)

$$L_{x^\alpha} = \partial_{x^\alpha} L = \partial_{x_1^{\alpha_1} \dots x_N^{\alpha_N}} L \tag{39}$$

where we have introduced multi-index $\alpha = (\alpha_1, \dots, \alpha_N)$ to simplify the notation. Due to the linearity of the diffusion equation, all these *scale-space derivatives* L_{x^α} satisfy similar scale-space properties in terms of non-enhancement of local extrema as the original scale-space L. Due to the commutative property between convolution and differentiation, these scale-space derivatives can also be computed by applying Gaussian derivative operators (see figure 10) to the original signal

$$L_{x^\alpha}(\cdot;\ s) = \partial_{x^\alpha} L(\cdot;\ s) = (\partial_{x^\alpha} g(\cdot;\ s)) * f(\cdot). \tag{40}$$

For this reason, these derivative operators are also referred to as *Gaussian derivatives*. From linear combinations of partial derivatives, we can also compute directional derivatives in any direction $(\cos\varphi, \sin\varphi)$ (see figure 11) which also satisfy scale-space properties in terms of non-enhancement of local extrema. In two dimensions, we have

$$\partial_{\varphi^N} L = (\cos\varphi\, \partial_x + \sin\varphi\, \partial_y)^N L = \sum_{k=0}^{N} \left(\begin{array}{c} N \\ k \end{array} \right) \cos^k \varphi\, \sin^k \varphi\, L_{x^k y^{N-k}} \tag{41}$$

With regard to image deformations, the closedness properties of this original scale-space are restricted to translations, rotations and rescalings. On the other hand, this scale-space concept is *separable*

$$L(x;\ s) = g(x;\ s) * f(x) = g(x_1;\ s) * \dots * g(x_N;\ s) * f(x_1, \dots, x_N) \tag{42}$$

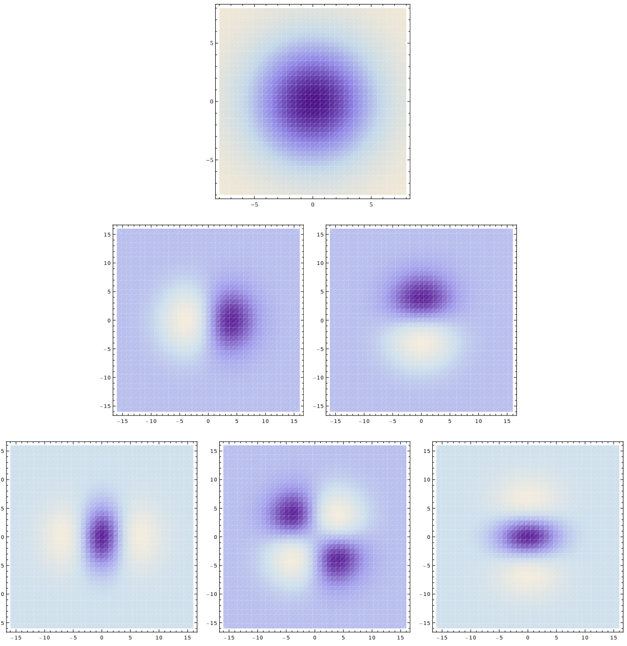


Figure 10: The Gaussian kernel in the 2-D case with its derivatives up to order two ($s = 16$).

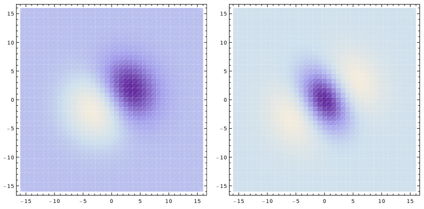


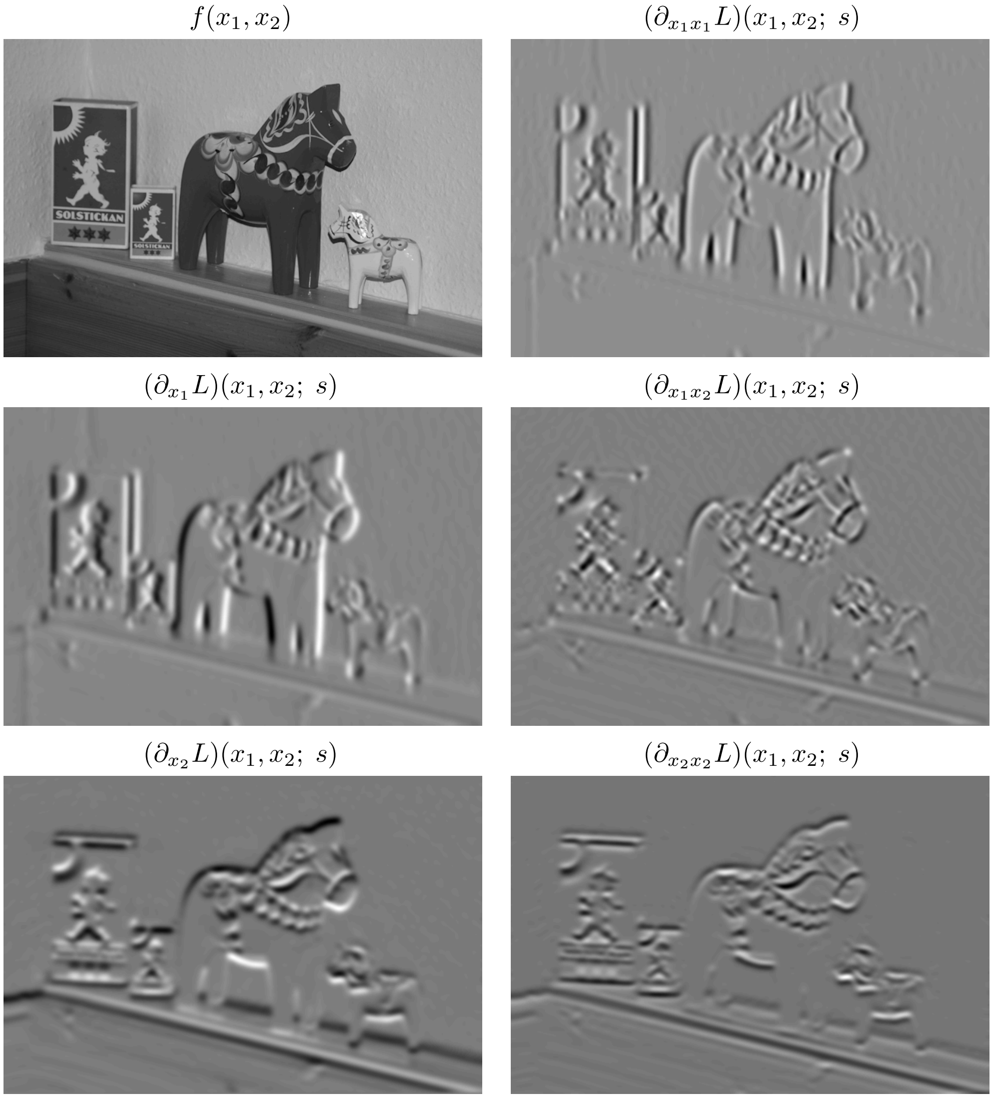
Figure 11: First- and second-order directional derivatives of the Gaussian kernel in the 2-D case computed from a linear combination of Cartesian partial derivatives according to equation (41) ($s = 16$, $\varphi = \pi/6$).

corresponding to the convolution[7] with one-dimensional Gaussian kernels along each dimension, which improves the computational efficiency in serial implementations. This separability carries over also to partial derivatives

$$L_{x^\alpha}(x;\ s) = g_{x^\alpha}(x;\ s) = g_{x_1^{\alpha_1}}(x_1;\ s) * \cdots * g_{x_N^{\alpha_N}}(x_N;\ s) * f(x_1, \ldots, x_N) \tag{43}$$

For derivatives up to order four, these expressions reduce to separable convolution with the

[7] In equations (42)–(43) we make an intentionally somewhat sloppy use of the convolution symbol in order to simplify the notation. These expression should be understood as one-dimensional convolutions carried out along each one of the dimensions. The presence of variable names as argument to the functions indicate over which dimension the convolution is performed. The correct notation for convolutions is of the form (40).



$f(x_1, x_2)$

$(\partial_{x_1 x_1} L)(x_1, x_2;\ s)$

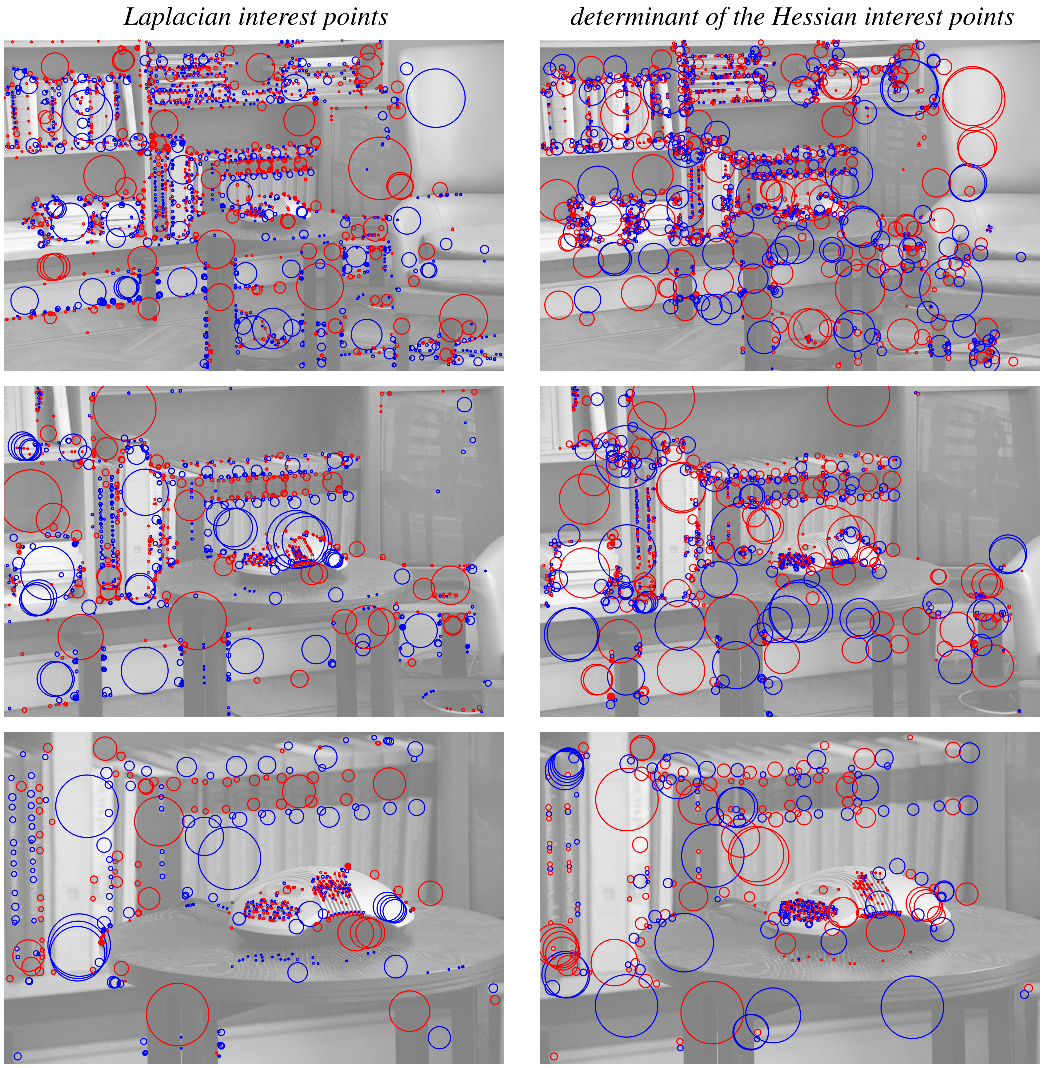
$(\partial_{x_1} L)(x_1, x_2;\ s)$

$(\partial_{x_1 x_2} L)(x_1, x_2;\ s)$

$(\partial_{x_2} L)(x_1, x_2;\ s)$

$(\partial_{x_2 x_2} L)(x_1, x_2;\ s)$

Figure 12: First- and second-order partial derivatives computed from a grey-level image at scale $s = 16$.



Laplacian interest points

determinant of the Hessian interest points

Figure 13: Examples of image features that can be computed based on the rotationally symmetric Gaussian scale-space concept. This figure shows scale-invariant interest points detected from three images obtained by gradually zooming in to structures in a library using two blob detectors with automatic scale selection proposed in (Lindeberg 1998) based on the detection of scale-space extrema of (left column) the scale-normalized Laplacian $\nabla_{norm} L = s\,(L_{x_1x_1} + L_{x_2x_2})$ and (right column) the scale-normalized determinant of the Hessian $\det \mathcal{H}_{norm} L = s^2\,(L_{x_1x_1} L_{x_2x_2} - L_{x_1x_2}^2)$. Each circle indicates an interest point with the radius of the circle equal to the detection scale of the feature in units of $\sigma = \sqrt{s}$, with red circles corresponding to positive values of the differential invariant and blue circles corresponding to negative values. Note how a large number of these features are preserved under rescalings in the image domain, which is a consequence of the scale invariance of the underlying feature detectors. Approximations of the these interest point detectors are used as primary feature detectors in the SIFT and SURF methods for image matching and object recognition proposed by (Lowe 2004) and (Bay et al. 2008), respectively. Because of the varying amount of image structures in these images caused by the size variations in the image domain due to zooming, the number of image features has been varied such that the images in the top row show the 2000 strongest interest points, the images in the middle row show the 1000 strongest interest points and the images in the bottom row show the 500 strongest interest points. (Scale range: $s \in [4, 4096]$, Image size: 1024×678 pixels.)

following one-dimensional Gaussian and Gaussian derivative kernels

$$g(x;\ s) = \frac{1}{\sqrt{2\pi s}} e^{-x^2/2s} \tag{44}$$

$$g_x(x;\ s) = -\frac{x}{s} g(x;\ s) = -\frac{x}{\sqrt{2\pi} s^{3/2}} e^{-x^2/2s} \tag{45}$$

$$g_{xx}(x;\ s) = \frac{(x^2 - s)}{s^2} g(x;\ s) = \frac{(x^2 - s)}{\sqrt{2\pi} s^{5/2}} e^{-x^2/2s} \tag{46}$$

$$g_{xxx}(x;\ s) = -\frac{(x^3 - 3sx)}{s^3} g(x;\ s) = -\frac{x(x^2 - 3s)}{\sqrt{2\pi} s^{7/2}} e^{-x^2/2s} \tag{47}$$

$$g_{xxxx}(x;\ s) = \frac{(x^4 - 6sx^2 + 3s^2)}{s^4} g(x;\ s) = \frac{(x^4 - 6sx^2 + 3s^2)}{\sqrt{2\pi} s^{9/2}} e^{-x^2/2s} \tag{48}$$

The Gaussian derivative operators do not obey a semi-group property. Instead, they (as well as any linear combination of them) satisfy the *cascade smoothing property*

$$L_{x^\alpha}(\cdot;\ s_1 + s_2) = g(\cdot;\ s_1) * L_{x^\alpha}(\cdot;\ s_2). \tag{49}$$

Such Gaussian derivative operators and differential invariants can be used as a *general basis* for expressing a large number of different visual operations including as feature detection, feature classification, surface shape, image matching and image-based recognition (Witkin 1983, Koenderink 1984, Koenderink & van Doorn 1992, Lindeberg 1994*b*, Lindeberg 1994*a*, Sporring et al. 1996, Florack 1997, ter Haar Romeny 2003, Lindeberg 2008); see specifically (Schiele & Crowley 2000, Linde & Lindeberg 2004, Lowe 2004, Bay et al. 2008, Tola et al. 2010, Linde & Lindeberg 2012, Larsen et al. 2012) for explicit approaches for object recognition based on Gaussian derivative operators or approximations thereof.

Figure 12 shown an illustration of computing Gaussian derivative operators up to order two from a grey-level image. Figure 13 shows an example of computing scale-invariant interest points from such Gaussian derivative responses of second order and applying this operation to different images of the same scene taken with different amount of physical zooming. The excellent repeatability properties of these interest point detectors under changing camera zoom, alternatively changing virtual distance between objects in the world and the observer, are a consequence of the scale invariant properties of the scale selection mechanism used in the interest point detectors, which in turn is based on the covariance properties of the underlying receptive fields under scaling transformations.

5.3 Affine Gaussian scale-space

If we relax the condition about rotational symmetry, while keeping a requirement that the corresponding Green's function should be mirror symmetric on every line through the origin (in the sense that the filters h should satisfy $h(-x_1, -x_2;\ s) = h(x_1, x_2;\ s)$ for every $(x_1, x_2) \in \mathbb{R}^2$), *i.e.*, to avoid spatial shifts the Fourier transform should be real, we obtain the *affine Gaussian scale-space representation*, generated by convolution with affine Gaussian kernels

$$g(x;\ \Sigma_s) = \frac{1}{(2\pi)^{N/2}\sqrt{\det \Sigma_s}}\, e^{-x^T \Sigma_s^{-1} x/2}, \tag{50}$$

where Σ_s is a symmetric positive definite (covariance) matrix. Besides the requirement of rotational symmetry, the affine scale-space basically satisfies similar scale-space properties as the linear scale-space. The main difference is that the affine scale-space is closed under the full group of non-singular affine transformations.

Transformation property under affine image transformations. If two image patterns f_L and f_R are related by an affine transformation

$$f_L(\xi) = f_R(\eta) \quad \text{where} \quad \eta = A\xi + b, \tag{51}$$

and if affine Gaussian scale-space representations of these images are defined by

$$L(\cdot;\ \Sigma_L, \delta_L) = g(\cdot;\ \Sigma_L, \delta_L) * f_L(\cdot), \quad R(\cdot;\ \Sigma_R, \delta_R) = g(\cdot;\ \Sigma_R, \delta_R) * f_R(\cdot), \tag{52}$$

then L and R are related by (Lindeberg & Gårding 1997)

$$L(x;\ \Sigma_L, \delta_L) = R(y;\ \Sigma_R, \delta_R), \tag{53}$$

where the covariance matrices Σ_L and Σ_R satisfy

$$\Sigma_R = A \Sigma_L A^T, \tag{54}$$

and the offset vectors δ_L and δ_R in the Gaussian kernels can be traded against coordinate shifts in x and y as long as the following relation is satisfied:

$$y - \delta_R = A(x - \delta_L) + b. \tag{55}$$

With regard to image processing and computer vision, this means that image data subjected to affine transformations can be perfectly captured with the extended class of affine scale-space operations. Specifically, for two-dimensional images arising as perspective projections of three-dimensional scenes, this notion of affine image deformations can be used as a first-order linear approximation of non-linear perspective effects.

This scale-space concept has been studied by (Lindeberg & Gårding 1994, Lindeberg 1994*b*, Griffin 1996) and is highly useful when computing surface shape under local affine distortion (Lindeberg & Gårding 1997) and performing affine invariant segmentation (Ballester & Gonzalez 1998) and matching (Baumberg 2000, Schaffalitzky & Zisserman 2001, Mikolajczyk & Schmid 2004, Mikolajczyk & Schmid 2004, Lazebnik et al. 2005, Rothganger et al. 2006) (see figure 17). Combined with derivative operations, this affine scale-space concept can also serve as a natural idealized model for filter banks (Freeman & Adelson 1991, Simoncelli et al. 1992) consisting of elongated directional filters (Perona 1992).

In practice, there are two principally different ways of computing scale-space representations under affine alignment — either by deforming the filter shapes or by deforming the

$$\begin{tikzcd}[column sep=huge]
L(x;\ \Sigma_L, \delta_L) \arrow[r, "\left\{\begin{array}{c} \eta = A\xi + b \\ \Sigma_R = A\Sigma_L A^T \\ y - \delta_R = A(x - \delta_L) + b \end{array}\right\}" description] & R(y;\ \Sigma_R, \delta_R) \\
f_L(\xi) \arrow[u, "*g(\cdot;\ \Sigma_L, \delta_L)" description] \arrow[r, "\eta = A\xi + b" description] & f_R(\eta) \arrow[u, "*g(\cdot;\ \Sigma_R, \delta_R)" description]
\end{tikzcd}$$

Figure 14: Explicit manifestation of the commutative diagram in figure 8 for the Gaussian scale-space concept under affine transformations of the spatial domain. The commutative property implies that the scale-space representations of two affine deformed image patches can be affine aligned, either by adapting the shapes of the Gaussian kernels or by deforming the image data prior to smoothing.

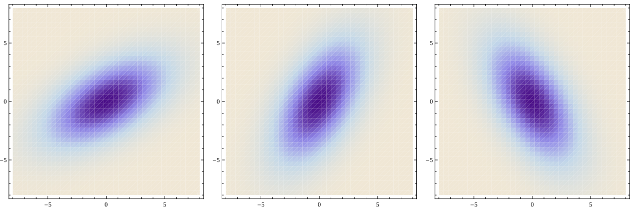


Figure 15: Examples of affine Gaussian kernels in the two-dimensional case (corresponding to $\lambda_1 = 16$, $\lambda_2 = 4$, $\beta = \pi/6, \pi/3, 2\pi/3$ in equation (56)).

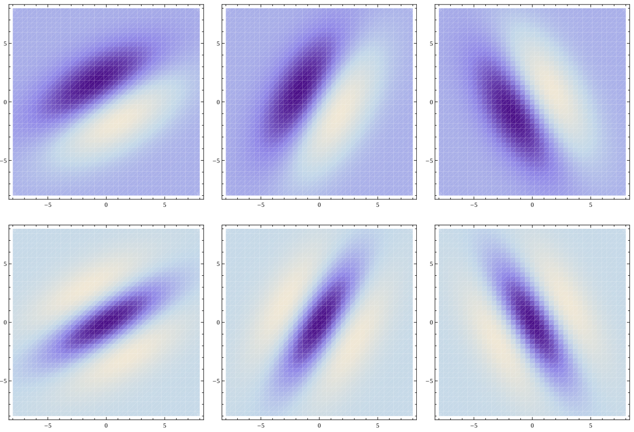


Figure 16: Elongated filters obtained by applying first- and second order directional derivatives to affine Gaussian kernels (corresponding to $\lambda_1 = 16$, $\lambda_2 = 4$, $\beta = \pi/6, \pi/3, 2\pi/3$, $\varphi = \beta + \pi/2$ in equation (56)).

image data before the smoothing operation. This equivalence is made explicit in the commutative diagram in figure 14, and the two approaches may have their respective advantages when expressing algorithms and computational mechanisms. In the ideal continuous case, the two approaches are mathematically equivalent. In a practical implementation, however, the first filter-based approach can be expected to be more accurate in the presence of noise whereas the second warping-based approach is usually faster with a serial implementation on a single-core computer, since the convolutions can then be performed by separable filters.

Parameterization of affine Gaussian kernels. To introduce more explicit notation for the affine Gaussian kernels, let us in the two-dimensional case parameterize such a covariance matrix by two eigenvalues λ_1, λ_2 and one orientation β. Then, the covariance matrix can be written

$$\begin{aligned}\Sigma' &= \left(\begin{array}{cc} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{array} \right) \left(\begin{array}{cc} \lambda_1 & 0 \\ 0 & \lambda_2 \end{array} \right) \left(\begin{array}{cc} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{array} \right) \\ &= \left(\begin{array}{cc} \lambda_1 \cos^2\beta + \lambda_2 \sin^2\beta & (\lambda_1 - \lambda_2) \cos\beta \, \sin\beta \\ (\lambda_1 - \lambda_2) \cos\beta \, \sin\beta & \lambda_1 \sin^2\beta + \lambda_2 \cos^2\beta \end{array} \right) \end{aligned} \tag{56}$$

with

$$\det \Sigma' = \lambda_1 \lambda_2 \tag{57}$$

Figure 15 shows a few examples of affine Gaussian filter kernels obtained in this way. Directional derivative operators of any order or orientation can then be obtained by combining equations (56) and (41) see figure 16 for a few illustrations.

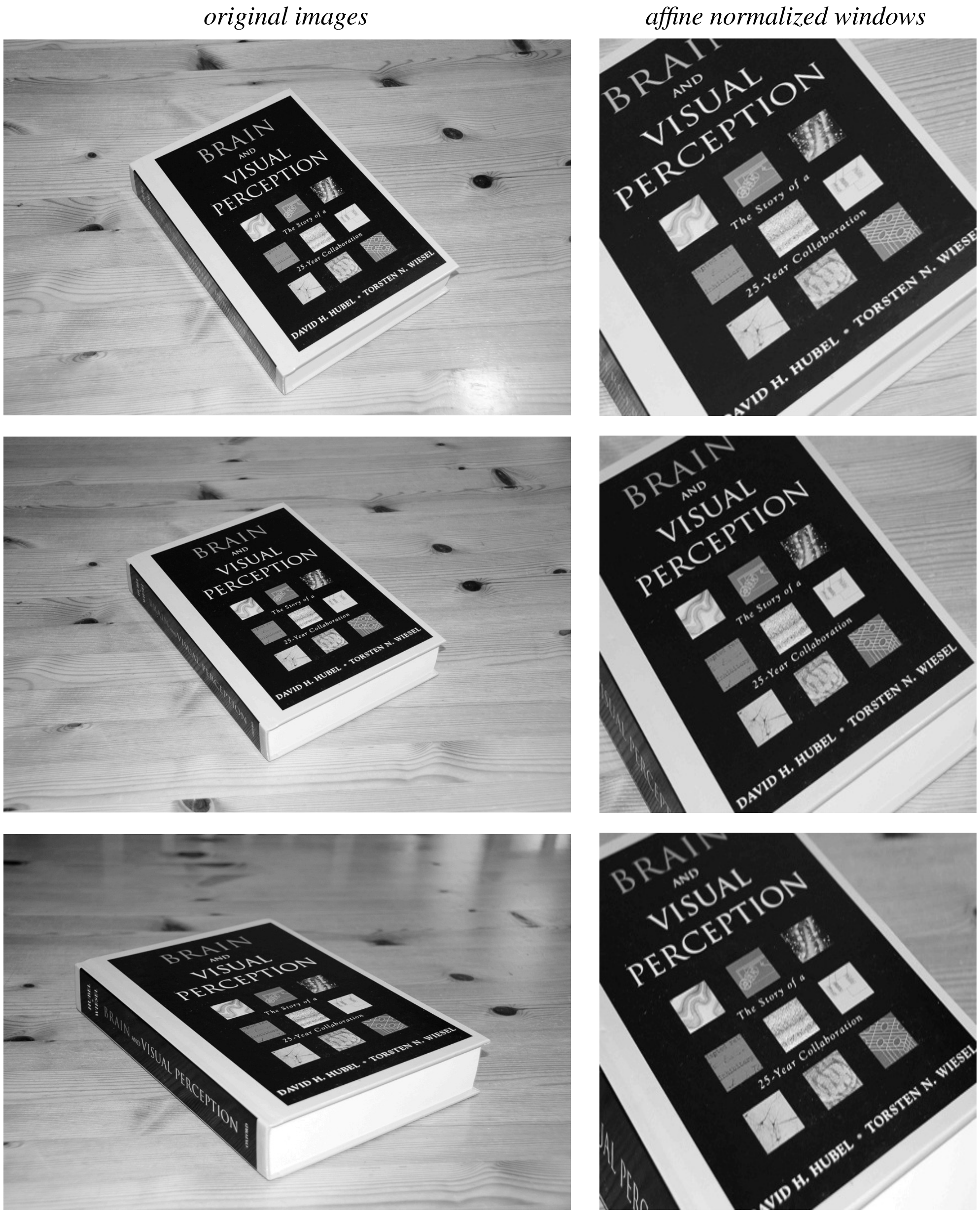


Figure 17: Illustration of how the affine Gaussian scale-space concept can be used for reducing the influence of perspective image deformations. The left column shows three views of a book with different amount of perspective foreshortening due to variations in the viewing direction relative to the surface normal of the front side of the book. The right column shows the result of performing an affine normalization of a central window in each image independently, by performing an affine warping to an affine invariant reference frame computed from an affine invariant fixed point in affine scale-space using the affine shape adaptation method proposed in (Lindeberg & Gårding 1997). Note how this leads to a major compensation for the perspective foreshortening, which can be used for significantly improving the performance of methods for image matching and object recognition under perspective projection. With regard to receptive fields, the use of an affine family of receptive field profiles makes it possible to define image operations in the image domain that are equivalent to the use of receptive fields based on rotationally symmetric smoothing operations in an affine invariant reference frame.

When computing directional derivatives from elongated affine Gaussian kernels, it should be noted that it is natural to align the orientations of the directional derivative operators (the angle φ in equation (41)) with the orientations of the eigen-directions of the covariance matrix in the affine Gaussian kernels (the angle β in equation (56)).

Under variations of the eigenvalues (λ_1, λ_2) and the eigen-direction β in equation (56), the covariance matrices Σ will span the variability of the affine transformations that arise from local linearizations of a smooth surfaces of objects seen from different viewing directions as illustrated in figure 18.

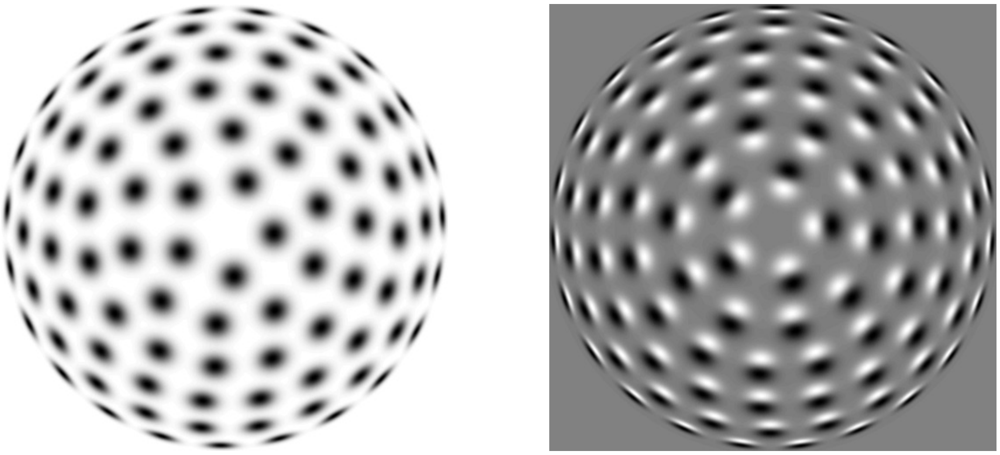


Figure 18: Affine Gaussian receptive fields generated for a set of covariance matrices Σ that correspond to an approximately uniform distribution on a hemisphere in the 3-D environment, which is then projected onto a 2-D image plane. (left) Zero-order receptive fields. (right) First-order receptive fields. If we consider surface patterns of three-dimensional objects in the world that are projected to a two-dimensional image plane, then these surface patterns will be deformed by the perspective projection. If we in the ideal theoretical case would l like to process these projected surface patterns by image operations that correspond to rotationally symmetric smoothing operations when backprojected to the tangent plane of surface, then the variability of this family of affine Gaussian receptive fields spans the full variability of the affine image deformations that arise from local linearizations of the non-linear perspective deformations. By allowing for a family of affine family of receptive fields, it will thereby be possible to substantially reduce the otherwise large mismatch between the receptive fields (illustrated in figure 5) that would occur when observing an object from different viewing directions.

5.4 Gaussian colour-opponent scale-space

To define a corresponding scale-space concept for colour images, the simplest approach would be by computing a Gaussian scale-space representation for each colour channel individually. Since the values of the colour channels will usually be highly correlated, it is, however, preferable to *decorrelate* the dependencies by computing a colour opponent representation. Such a representation is also in good agreement with human vision, where a separation into red/green and yellow/blue colour colour-opponent channels takes place at an early stage in the visual pathways.

Given three RGB channels obtained from a colour sensor, consider a colour-opponent

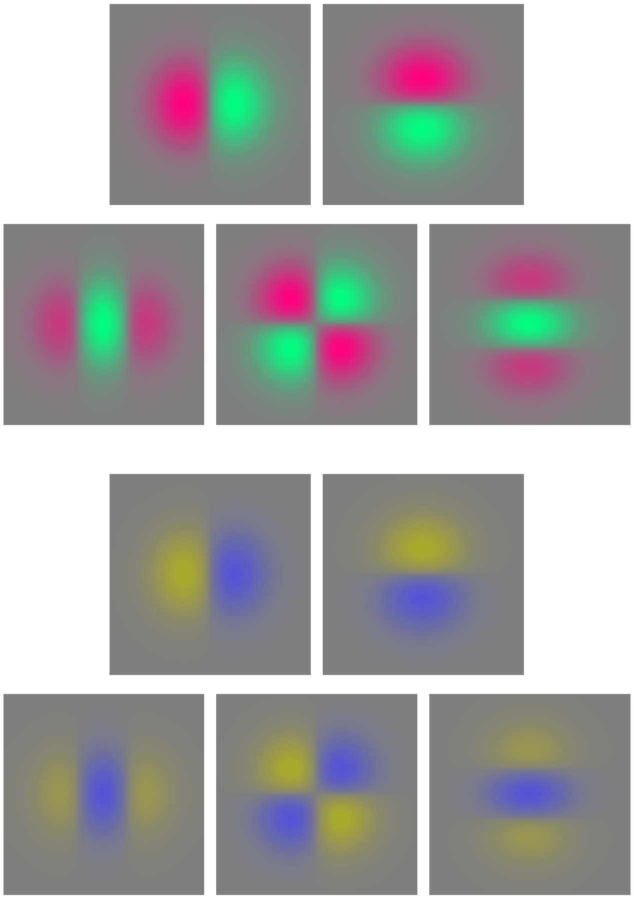


Figure 19: Spatio-chromatic receptive fields corresponding to the application of Gaussian derivative operators up to order two to red/green and yellow/blue colour opponent channels, respectively

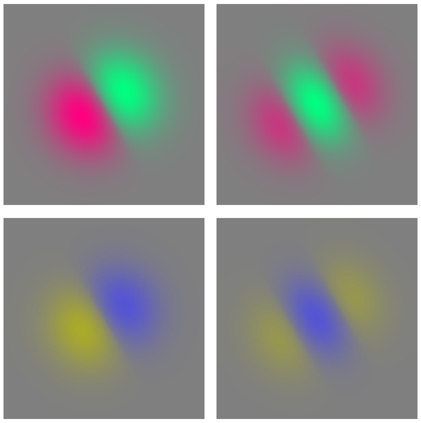
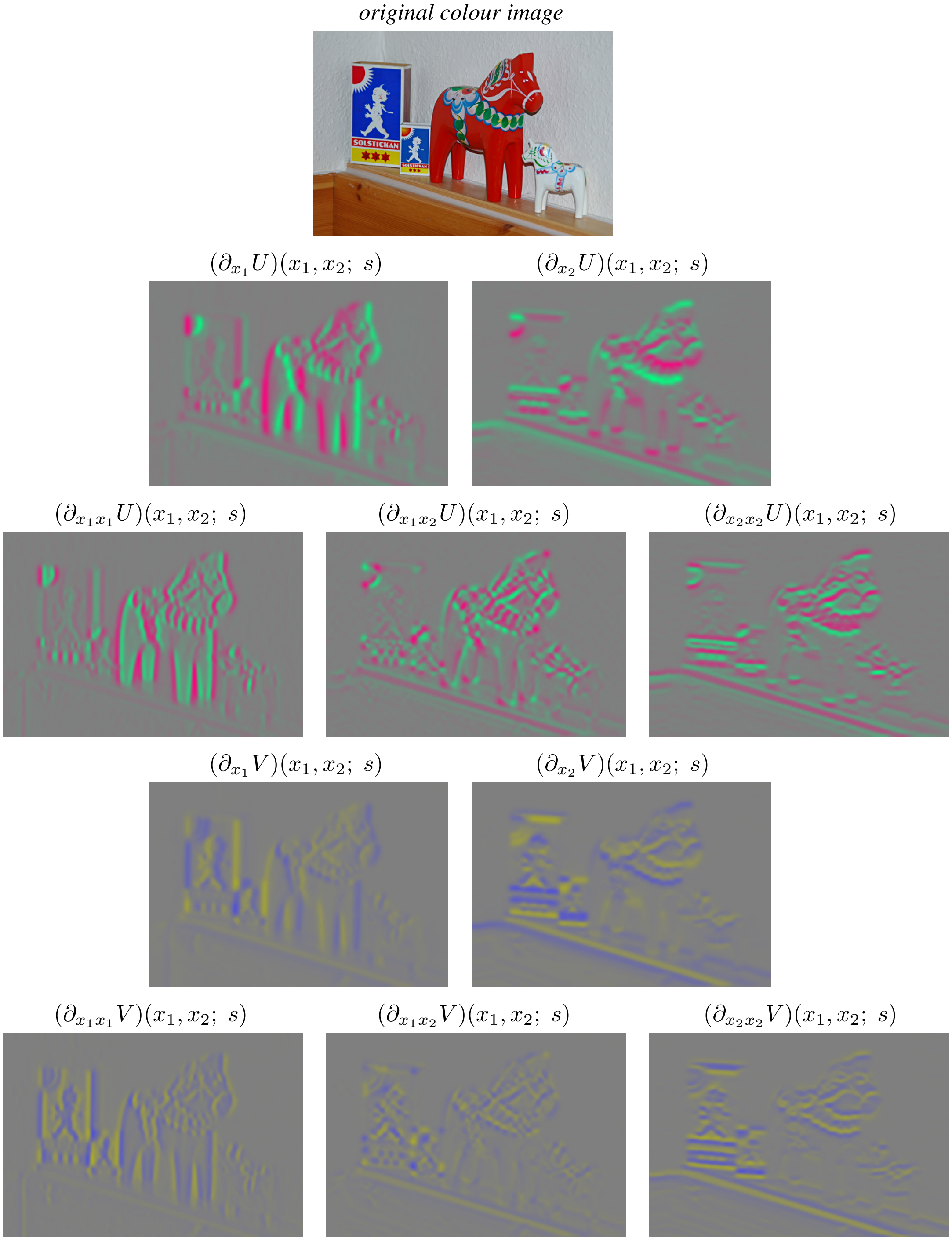


Figure 20: Spatio-chromatic receptive fields corresponding to the application of Gaussian directional derivatives up to order two along the direction $\varphi = \pi/6$ to red/green and yellow/blue colour opponent channels, respectively



original colour image

$(\partial_{x_1} U)(x_1, x_2;\ s)$ $(\partial_{x_2} U)(x_1, x_2;\ s)$

$(\partial_{x_1 x_1} U)(x_1, x_2;\ s)$ $(\partial_{x_1 x_2} U)(x_1, x_2;\ s)$ $(\partial_{x_2 x_2} U)(x_1, x_2;\ s)$

$(\partial_{x_1} V)(x_1, x_2;\ s)$ $(\partial_{x_2} V)(x_1, x_2;\ s)$

$(\partial_{x_1 x_1} V)(x_1, x_2;\ s)$ $(\partial_{x_1 x_2} V)(x_1, x_2;\ s)$ $(\partial_{x_2 x_2} V)(x_1, x_2;\ s)$

Figure 21: First- and second-order spatio-chromatic partial derivatives computed from a colour image at scale $s = 16$.

transformation of the form (Hall et al. 2000)

$$\begin{pmatrix} f \\ u \\ v \end{pmatrix} = \begin{pmatrix} f \\ c^{(1)} \\ c^{(2)} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & -1 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \tag{58}$$

where yellow is approximated by the average of the R and G channels and f can be defined as a channel of pure intensity information. Then, a *Gaussian colour-opponent scale-space representation* $(U, V)^T = (C^{(1)}, C^{(2)})^T$ can be defined by applying Gaussian convolution to the colour channels $(c^{(1)}, c^{(2)})$:

$$U = C^{(1)}(\cdot, \cdot;\ s) = g(\cdot, \cdot;\ s) * c^{(1)}(\cdot), \tag{59}$$

$$V = C^{(2)}(\cdot, \cdot;\ s) = g(\cdot, \cdot;\ s) * c^{(2)}(\cdot). \tag{60}$$

Figure 19 shows equivalent spatio-chromatic receptive fields corresponding to the application of Gaussian derivative operators according to (43) to such colour-opponent channels. Figure 20 shows examples of corresponding directional derivatives according to (41). Figure 21 shows an illustration of computing such spatio-chromatic derivatives in a red/green and yellow/blue colour-opponent space from a colour image.

In (Hall et al. 2000, Linde & Lindeberg 2004, Burghouts & Geusebroek 2009, van de Sande et al. 2010, Linde & Lindeberg 2012, Zhang et al. 2012) it is shown how such colour-opponent spatio-chromatic receptive fields in combination with regular spatial receptive fields can constitute a very effective basis for object recognition.

5.5 Uniqueness of the Gaussian kernel on a spatial domain

The Gaussian scale-space concept satisfies the requirements of (i) linearity, (ii) shift invariance, (iii) semi-group property, (iv) existence of an infinitesimal generator, (v) non-creation of local extrema, (vi) non-enhancement of local extrema, (vii) rotational symmetry, (viii) positivity, (ix) normalization and (x) scale invariance. In section 5.1 we described how the Gaussian scale-space concept is uniquely defined from four of these requirements in combination with sufficient regularity requirements. The Gaussian scale-space can also be uniquely derived from other combinations of basic scale-space axioms (Iijima 1962, Koenderink 1984, Babaud et al. 1986, Yuille & Poggio 1986, Lindeberg 1990, Lindeberg 1994*b*, Lindeberg 1994*a*, Pauwels et al. 1995, Lindeberg 1996, Sporring et al. 1996, Florack 1997, Weickert et al. 1999, ter Haar Romeny 2003). The Gaussian function is also special in the following respects:

- it minimizes the *uncertainty relation* (Folland & Sitaram 1997), which implies that in an N-dimensional space with $f \in L^2(\mathbb{R}^N)$ and with

$$\langle x \rangle^2 = \frac{\int_{x \in \mathbb{R}^N} |x - \overline{x}|^2 |f(x)|^2 dx}{\int_{x \in \mathbb{R}^N} |f(x)|^2 dx} \quad \text{where} \quad \overline{x} = \frac{\int_{x \in \mathbb{R}^N} x\, |f(x)|^2 dx}{\int_{x \in \mathbb{R}^N} |f(x)|^2 dx} \tag{61}$$

and

$$\langle \omega \rangle^2 = \frac{\int_{\omega \in \mathbb{R}^N} |\omega - \overline{\omega}|^2 |\hat{f}(\omega)|^2 d\omega}{\int_{\omega \in \mathbb{R}^N} |\hat{f}(\omega)|^2 d\omega} \quad \text{where} \quad \overline{\omega} = \frac{\int_{\omega \in \mathbb{R}^N} \omega\, |\hat{f}(\omega)|^2 d\omega}{\int_{\omega \in \mathbb{R}^N} |\hat{f}(\omega)|^2 d\omega} \tag{62}$$

then it holds for any f that

$$\langle x \rangle \, \langle \omega \rangle \geq \frac{N}{2} \tag{63}$$

and this relation is minimized by the Gaussian function

$$f(x) = g(x;\ s, m) = e^{-(x-m)^T(x-m)/2s^2}. \tag{64}$$

- if p is a probability density function on $\mathbb{R}^N$ with mean vector m

$$\int_{x \in \mathbb{R}^N} x\, p(x)\, dx = m \tag{65}$$

and covariance matrix Σ

$$\int_{x \in \mathbb{R}^N} (x-m)(x-m)^T\, p(x)\, dx = \Sigma \tag{66}$$

then the (possibly non-isotropic) Gaussian function

$$p(x) = g(x;\ \Sigma, m) = \frac{1}{(2\pi)^{N/2}\sqrt{\det \Sigma}}\, e^{-(x-m)^T\Sigma^{-1}(x-m)/2} \tag{67}$$

is the probability density function with *maximum entropy*

$$H = -\int_{x \in \mathbb{R}^N} p(x)\, \log p(x)\, dx \leq \frac{1}{2} \log\left((2\pi e)^N \det \Sigma\right). \tag{68}$$

The uncertainty relation means that the Gaussian function has maximally compact simultaneous localisation properties in the spatial and the frequency domains. The maximum entropy result can be interpreted as the Gaussian kernel requiring a minimum amount of information. These properties are also desirable when constructing a scale-space representation, since the uncertainty relation makes the smoothing operation well localized over space and scale, whereas the maximum entropy result means that the Gaussian kernel is maximally *uncommitted*.

The Gaussian kernel does also have the attractive property that after a non-infinitesimal amount of spatial smoothing, the Gaussian smoothed signal can be regarded as infinitely differentiable provided that the input signal is bounded. Thereby, the output from the Gaussian derivative operators can always be regarded as well-defined for any non-infinitesimal value of the scale parameter.

6 Scale-space axioms for spatio-temporal image domains

6.1 Scale-space axioms transferred from spatial to spatio-temporal domain

For spatio-temporal image data $f(x, t)$ defined on an $N+1$-dimensional spatio-temporal domain indexed by $p = (x, t)^T = (x_1, \dots, x_N, t)^T$, it is natural to inherit the symmetry requirements over the spatial domain. Given that we are interested in defining a spatio-temporal scale-space representation that comprises both a spatial scale parameter $s \in \mathbb{R}^M$ and a temporal scale parameter $\tau \in \mathbb{R}_+$, we would therefore like to determine a family of operators $\mathcal{T}_{s,\tau}$ that are to act on spatio-temporal image data $f \colon \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ to produce a family of intermediate representations $L \colon \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^M \times \mathbb{R}_+ \to \mathbb{R}$ according to

$$L(\cdot, \cdot;\ s, \tau) = \mathcal{T}_{s,\tau}\, f(\cdot, \cdot) \tag{69}$$

where $(\cdot, \cdot)$ denote the arguments for the spatial and temporal coordinates, respectively.

Linearity. If we want the initial visual processing stages to make as few irreversible decisions as possible, it is natural to initially require $\mathcal{T}_s$ to be a *linear operator* such that

$$\mathcal{T}_{s,\tau}(a_1 f_1 + a_2 f_2) = a_1 \mathcal{T}_{s,\tau} f_1 + a_2 \mathcal{T}_{s,\tau} f_2 \tag{70}$$

holds for all functions $f_1, f_2 \colon \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ and all scalar constants $a_1, a_2 \in \mathbb{R}$.

Again, linearity implies that a number of special properties of receptive fields (to be described below) will transfer to spatio-temporal derivatives of these and do therefore imply that different types of spatio-temporal image structures will be treated in a similar manner irrespective of what types of linear filters they are captured by.

Specifically, such spatio-temporal derivative operators will respond to relative variations in image intensities and will therefore be less sensitive to local illumination variations than zero-order image intensities.

Translational invariance. Let us also require $\mathcal{T}_{s,\tau}$ to be a *shift-invariant operator* in the sense that it commutes with the spatio-temporal shift operator $\mathcal{S}_{(\Delta x, \Delta t)}$ defined by $(\mathcal{S}_{(\Delta x, \Delta t)} f)(x) = f(x - \Delta x, t - \Delta t)$, such that

$$\mathcal{T}_{s,\tau} \left(\mathcal{S}_{(\Delta x, \Delta t)} f \right) = \mathcal{S}_{(\Delta x, \Delta t)} \left(\mathcal{T}_{s,\tau} f \right) \tag{71}$$

holds for all $\Delta x \in \mathbb{R}^N$ and $\Delta t \in \mathbb{R}$. The motivation behind this assumption is the basic requirement that the representation of a visual object should be similar irrespective of its position in space-time. Alternatively stated, the operator $\mathcal{T}_{s,\tau}$ can be said to be *homogeneous across space-time*.

Convolution structure. Together, the assumptions of linearity and shift-invariance imply that the internal representations $L(\cdot, \cdot;\; s, \tau)$ are given by *convolution transformations*

$$L(x, t;\; s, \tau) = (T(\cdot,\, \cdot;\; s, \tau) * f)(x, t) = \int_{\xi \in \mathbb{R}^N} \int_{\eta \in \mathbb{R}} T(\xi, \eta;\; s, \tau)\, f(x - \xi, t - \eta)\, d\xi\, d\eta \tag{72}$$

where $T(\cdot, \cdot;\; s, \tau)$ denotes some family of convolution kernels. These convolution kernels and their spatio-temporal derivatives can also be referred to as spatio-temporal receptive fields.

Regularity. To be able to use tools from functional analysis, we will initially assume that both the original signal f and the family of convolution kernels $T(\cdot, \cdot;\; s, \tau)$ are in the Banach space $L^2(\mathbb{R}^N \times \mathbb{R})$, *i.e.* that $f \in L^2(\mathbb{R}^N \times \mathbb{R})$ and $T(\cdot, \cdot;\; s,\; \tau) \in L^2(\mathbb{R}^N \times \mathbb{R})$ with the norm

$$\|f\|_2^2 = \int_{x \in \mathbb{R}^N} \int_{t \in \mathbb{R}} |f(x, t)|^2\, dx\, dt. \tag{73}$$

Then also the intermediate representations $L(\cdot, \cdot;\; s, \tau)$ will be in the same Banach space, and the operators $\mathcal{T}_{s,\tau}$ can be regarded as well-defined.

Positivity (non-negativity). Concerning the convolution kernels, one may require these to be non-negative in order to constitute smoothing transformations.

$$T(x, t;\; s,\; \tau) \geq 0. \tag{74}$$

Normalization. Furthermore, it may be natural to require the convolution kernels to be normalized to unit mass

$$\int_{x \in \mathbb{R}^N} \int_{t \in \mathbb{R}} T(x, t;\ s, \tau) \, dx \, dt = 1. \tag{75}$$

to leave a constant signal unaffected by the smoothing transformation.

Quantitative measurement of the spatio-temporal extent and the spatio-temporal offset of non-negative scale-space kernels. For a non-negative convolution kernel, we can measure its spatial offset $\bar{p} = (\bar{x}, \bar{t})^T$ by the mean operator

$$m = \bar{p} = M(T(\cdot;\ s, \tau)) = \frac{\int_{p=(x,t)^T \in \mathbb{R}^N \times \mathbb{R}} p \, T(x, t;\ s, \tau) \, dx \, dt}{\int_{p=(x,t)^T \in \mathbb{R}^N \times \mathbb{R}} T(x, t;\ s, \tau) \, dx \, dt} \tag{76}$$

and its spatial extent by the spatial covariance matrix

$$\Sigma = C(T(\cdot;\ s, \tau)) = \frac{\int_{p=(x,t)^T \in \mathbb{R}^N \times \mathbb{R}} ((p - \bar{p}) \, (p - \bar{p})^T \, T(x, t;\ s,\ \tau) \, dx \, dt}{\int_{p=(x,t)^T \in \mathbb{R}^N \times \mathbb{R}} T(x, t;\ s, \tau) \, dx \, dt}. \tag{77}$$

Using the additive properties of mean values and covariance matrices under convolution, which hold for non-negative distributions, it follows that

$$m = M(T(\cdot, \cdot;\ s_1, \tau_1) * T(\cdot, \cdot;\ s_2,\ \tau_2)) = M(T(\cdot, \cdot;\ s_1, \tau_1)) + M(T(\cdot, \cdot;\ s_2, \tau_2)) = m_1 + m_2, \tag{78}$$

$$\Sigma = C(T(\cdot, \cdot;\ s_1, \tau_1) * T(\cdot, \cdot;\ s_2, \tau_2)) = C(T(\cdot, \cdot;\ s_1, \tau_1)) + C(T(\cdot, \cdot;\ s_2, \tau_2)) = \Sigma_1 + \Sigma_2 \tag{79}$$

Identity operation with continuity. To guarantee that the limit case of the internal scale-space representations when the scale parameters s and τ tend to zero should correspond to the original image data f, we will assume that

$$\lim_{s \downarrow 0, \tau \downarrow 0} L(\cdot, \cdot;\ s, \tau) = \lim_{s \downarrow 0, \tau \downarrow 0} \mathcal{T}_{s,\tau} f = f. \tag{80}$$

Hence, the intermediate image representations $L(\cdot, \cdot; s, \tau)$ can be regarded as a family of derived representations parameterized by a spatial scale parameter s and a temporal scale parameter τ. Since $s \in \mathbb{R}^M$ and $\tau \in \mathbb{R}_+$ together span a multi-dimensional scale parameter $r = (s, \tau) \in \mathbb{R}^M \times \mathbb{R}_+$, equation (80) should be interpreted as $\lim_{|r| \downarrow 0} L(\cdot, \cdot;\ s, \tau) = \lim_{|r| \downarrow 0} \mathcal{T}_{s,\tau} f = f$ with $|r| = \sqrt{s_1^2 + \cdots + s_M^2 + \tau^2}$.

Semi-group structure. For such image measurements to be properly related *between* different spatio-temporal scales, it is natural to require the operators $\mathcal{T}_{s,\tau}$ with their associated convolution kernels $T(\cdot, \cdot;\ s, \tau)$ to form a *semi-group* over both s and τ

$$\mathcal{T}_{s_1,\tau_1} \mathcal{T}_{s_2,\tau_2} = \mathcal{T}_{s_1+s_2,\tau_1+\tau_2} \tag{81}$$

with a corresponding semi-group structure for the convolution kernels

$$T(\cdot, \cdot;\ s_1, \tau_1) * T(\cdot, \cdot;\ s_2, \tau_2) = T(\cdot, \cdot;\ s_1 + s_2, \tau_1 + \tau_2). \tag{82}$$

Then, the transformation between any different and ordered scale levels (s_1, τ_1) and (s_2, τ_2) with $s_2 \geq s_1$ and $\tau_2 \geq \tau_1$ will obey the *cascade property*

$$L(\cdot, \cdot;\; s_2, \tau_2) = T(\cdot, \cdot;\; s_2 - s_1, \tau_2 - \tau_1) * T(\cdot, \cdot;\; s_1, \tau_1) * f = T(\cdot, \cdot;\; s_2 - s_1, \tau_2 - \tau_1) * L(\cdot, \cdot;\; s_1, \tau_1) \tag{83}$$

i.e. a similar type of transformation as from the original data f. An image representation having these properties is referred to as a *spatio-temporal multi-scale representation*.

6.2 Additional scale-space axioms for time-dependent image data

For spatio-temporal image data, the following covariance requirements are natural to impose motivated by the special nature of time and space-time.

Temporal covariance. If the same scene is observed by two different cameras that sample the spatio-temporal image data with different temporal sampling rates, or if a camera observes similar types of motion patterns that occur with different speed, it seems natural that the visual system should be able to relate the spatio-temporal scale-space representations that are computed from the time-dependent image data. Therefore, one may require that if the temporal dimension is rescaled by a uniform scaling factor

$$f' = \mathcal{B}\, f \quad \text{corresponding to} \quad f'(t') = f(t) \quad \text{with} \quad t' = B\, t, \tag{84}$$

then there should exist some transformation of the temporal scale parameter $\tau' = B(\tau)$ such that the corresponding spatio-temporal scale-space representations are equal (here with the spatial dimension and the spatial scale parameter(s) suppressed; see also figure 22):

$$L'(t';\; \tau') = L(t;\; \tau) \quad \text{corresponding to} \quad \mathcal{T}_{B(\tau)}\, \mathcal{B}\, f = \mathcal{B}\, \mathcal{T}_\tau\, f. \tag{85}$$

$$\begin{tikzcd}
\mathcal{T}_\tau\, f \arrow[r, "\mathcal{B}"] & L \\
f \arrow[r, "\mathcal{B}"] \arrow[u, "\mathcal{T}_\tau"'] & \mathcal{B}\, f \arrow[u, "\mathcal{T}_{B(\tau)}"']
\end{tikzcd}$$

Figure 22: Commutative diagram for scale-space representations computed under uniform scalings of the temporal dimension. Such a temporal scaling transformation may, for example, represent spatio-temporal image data that have been acquired using visual sensors that sample the image data with different temporal sampling rates or motion patterns alternatively spatio-temporal events in the world that occur with different speed alternatively of different temporal extent.

Galilean covariance. For time-dependent spatio-temporal image data, we may have *relative motions* between objects in the world and the observer, where a constant velocity translational motion can be modelled by a *Galilean transformation*

$$f' = \mathcal{G}_v\, f \quad \text{corresponding to} \quad f'(x', t') = f(x, t) \quad \text{with} \quad x' = x + v\, t \tag{86}$$

where v denotes the image velocity.

To enable a consistent visual interpretation under different relative motions, it is natural to require that it should be possible to relate internal representations $L(\cdot, \cdot;\; s, \tau)$ that are computed from spatio-temporal image data under different relative motions

$$L'(x', t';\; s', \tau') = L(x, t;\; s, \tau) \quad \text{corresponding to} \quad \mathcal{T}_{G_v(s,\tau)}\, \mathcal{G}_v\, f = \mathcal{G}_v\, \mathcal{T}_{s,\tau}\, f. \tag{87}$$

Such a property is referred to as *Galilean covariance* (see figure 23).

$$\begin{tikzcd}
\mathcal{T}_{s,\tau}\, f \arrow[r, "\mathcal{G}_v"] & L \\
f \arrow[u, "\mathcal{T}_{s,\tau}"'] \arrow[r, "\mathcal{G}_v"] & \mathcal{G}_v\, f \arrow[u, "\mathcal{T}_{G_v(s,\tau)}"']
\end{tikzcd}$$

Figure 23: Commutative diagram for a spatio-temporal scale-space representation computed under a Galilean transformation of space-time. Such a constant velocity motion may, for example, represent a local linear approximation of the projected motion field for corresponding image points under relative motions between objects in the world and the visual observer.

6.2.1 Specific scale-space axioms for a non-causal spatio-temporal domain

Depending on the conditions under which the spatio-temporal image data are accessed, we can consider two different types of cases. For pre-recorded spatio-temporal image data such as video, we may in principle assume access to image information all temporal moments simultaneously and thereby apply similar types of operations as are used for processing purely spatial image data. For real-time vision or when modelling biological vision, there is, however, no way of having access to the future, which imposes fundamental additional structural requirements on a spatio-temporal visual front-end.

In this section, we shall develop a set of spatio-temporal scale-space axioms that can be used when processing pre-recorded image data in an offline situation, where temporal causality can be disregarded.

Infinitesimal generator for non-causal spatio-temporal domain. For theoretical analysis it is preferably if the spatial scale parameter s and the temporal scale parameter τ can be treated as continuous parameters and if image representations at adjacent scales can be related by partial differential equations. With $r = (s, \tau) = (s_1, \dots, s_M, \tau)$ denoting a multi-dimensional spatio-temporal scale parameter, we define the *directional derivative of the semi-group* along any *positive direction* $u = (u_1, \dots, u_M, u_{M+1})$ in the parameter space

$$(\mathcal{D}_u L)(x, t;\ s, \tau) = (\mathcal{B}(u)\, L)(x, t;\ s, \tau) = (u_1 \mathcal{B}_1 + \dots + u_M \mathcal{B}_M + u_{M+1} \mathcal{B}_{M+1})\ L(x, t;\ s, \tau) \tag{88}$$

where each $\mathcal{B}_k$ $(k = 1 \dots M, M+1)$ constitutes the infinitesimal generator for the parameter r_k along the unit direction e_k in the $M+1$-dimensional parameter space

$$\mathcal{B}_k L = \lim_{h \downarrow 0} \frac{T(\cdot, \cdot;\ h\, e_k) * f - f}{h} \tag{89}$$

and with the notion of a "positive direction" in parameter space defined in a corresponding way as in footnote 4.

In (Lindeberg 2011, Section 3.2 and Appendix A) it is shown how such differential relationships can be ensured given a proper selection of functional spaces and sufficient regularity requirements over space-time (x, t) and spatio-temporal scales (s, τ) in terms of Sobolev norms. We shall therefore henceforth regard the internal representations $L(\cdot, \cdot;\ s, \tau)$ as differentiable with respect to the spatio-temporal scale parameters s and τ.

Non-enhancement of local extrema for non-causal spatio-temporal domain A natural way to express the requirement of non-enhancement of local extrema for spatio-temporal image data is by requiring the value to not be allowed to increase in any positive direction in the parameter space of spatio-temporal scales. In other worlds, if a point $(x_0, t_0;\ s_0, \tau_0)$ is a local (spatial) maximum of the mapping $(x, t) \mapsto L(x, t;\ s_0, \tau_0)$ then the value must not increase with scale in any positive direction in parameter space. Similarly, if a point $(x_0, t_0;\ s_0, \tau_0)$ is a local (spatial) minimum of the mapping $(x, t) \mapsto L(x, t;\ s_0, t_0)$, then the value must not decrease with scale in any positive direction in parameter space. Given the above mentioned differentiability property with respect to scale, we say that the multi-scale representation constitutes a *scale-space representation* if it satisfies the following conditions

$$(\mathcal{D}_u L)(x_0, t_0;\ s_0, \tau_0) \leq 0 \qquad \text{at any non-degenerate local maximum,} \tag{90}$$

$$(\mathcal{D}_u L)(x_0, t_0;\ s_0,\ \tau_0) \geq 0 \qquad \text{at any non-degenerate local minimum,} \tag{91}$$

for any positive direction $u = (u_1, \ldots, u_M, u_{M+1})$ in the parameter space.

6.2.2 Special scale-space axioms for a time-causal spatio-temporal domain

When processing spatio-temporal image data in a real-time scenario, the following additional temporal and spatio-temporal requirements are instead needed:

Temporal causality. For a vision system that interacts in with the environment in a real-time setting, a fundamental constraint on the convolution kernels (the spatio-temporal receptive fields) is that there is no way of having access to future information, which implies that the temporal smoothing kernels must be *time-causal* in the sense that the convolution kernel must be zero for any relative time moment that would imply access to the future:

$$T(x, t;\ s, \tau) = 0 \quad \text{if} \quad t < 0. \tag{92}$$

Note that the possibly pragmatic solution of using a truncated symmetric filter of finite support in combination with a temporal delay is not appropriate for a time-critical real-time system, since it would need to unnecessarily long time delays in particular at coarser temporal scales. Therefore, a dedicated theory for truly time-causal spatio-temporal scale-space concepts is needed.

Time-recursivity. Another fundamental constraint on a real-time system is that it cannot be expected to keep a full record of everything that has happened in the past. Too keep down memory requirements it is therefore desirable that the computations can be based on a limited internal *temporal buffer* $M(x, t)$, which should provide:

- a sufficient record of past information and
- sufficient information to update its internal state in a recursive manner over time as new information arrives.

A particularly useful solution in this context is to use the internal spatio-temporal representations L at different temporal scales τ as a sufficient memory buffer of the past. Depending on whether the temporal scale parameter is regarded as as a continuous parameter or a set of discrete temporal scale levels, two different types of special cases can be distinguished:

Time-recursivity in the context of a continuum of temporal scale levels. For a spatio-temporal scale-space representation enabling a continuum of temporal scale levels $\tau \in \mathbb{R}_+$, such a requirement of a time-recursive structure over over time can in combination with a semi-group structure over image space $x \in \mathbb{R}^N$ with an associated spatial scale parameter $s \in \mathbb{R}^M$ be formalized in terms of a spatio-temporal time-recursive updating rule of the form (Lindeberg 2011, section 5.1.3, page 57)

$$
\begin{aligned}
L(x, t_2; s_2, \tau) = & \int_{\xi \in \mathbb{R}^N} \int_{\zeta \geq 0} U(x - \xi, t_2 - t_1;\ s_2 - s_1, \tau, \zeta)\, L(\xi, t_1;\ s_1, \zeta)\, d\zeta\, d\xi \\
& + \int_{\xi \in \mathbb{R}^N} \int_{u=t_1}^{t_2} B(x - \xi, t_2 - u;\ s_2, \tau)\, f(\xi, u)\, d\xi\, du
\end{aligned}
$$

which is required to hold for any pair of scale levels $s_2 \geq s_1$ and any two time moments $t_2 \geq t_1$, where

- the kernel U updates the internal state,
- the kernel B incorporates new image data into the representation and
- $\zeta \in \mathbb{R}_+$ is an integration variable referring to internal temporal buffers at different temporal scales.

Note that this algebraic structure comprises increments over both time $t_2 \geq t_1$ and spatial scales $s_2 \geq s_1$.

Time-recursivity in the context of discrete temporal scale levels. For a spatio-temporal scale-space representation $L(x, t;\ \sigma, k)$ restricted to a discrete set of scale levels τ_k for $k = 1 \dots K$, temporal recursivity can in combination with the requirement of a semi-group property over space be expressed in terms of a spatio-temporal time-recursive structure of the form

$$
\begin{aligned}
L(x, t_2; s_2, k) = & \int_{\xi \in \mathbb{R}^N} \sum_{\zeta=0}^{K} U(x - \xi, t_2 - t_1;\ s_2 - s_1, k, \zeta)\, L(\xi, t_1;\ s_1, \zeta)\, d\zeta\, d\xi \\
& + \int_{\xi \in \mathbb{R}^N} \int_{u=t_1}^{t_2} B(x - \xi, t_2 - u;\ s_2, k)\, f(\xi, u)\, d\xi\, du
\end{aligned}
$$

where the kernel U updates the internal state, the kernel B incorporates new information and ζ constitutes an index over the internal temporal scale levels. Again, this time-recursive structure does also comprise increments over both time $t_2 \geq t_1$ and scale $s_2 \geq s_1$.

Non-enhancement of local extrema in a time-recursive setting. For a time-recursive spatio-temporal visual front-end having a continuous scale parameter, it is natural to express the notion of non-enhancement of local extrema such that it is required to hold both with respect to increasing spatial scales s and evolution over time t (instead of increasing temporal scales τ as for a non-causal spatio-temporal scale-space). Thus, if at some spatial scale s_0 and time moment t_0 a point (x_0, τ_0) is a local maximum (minimum) for the mapping

$$
(x, \tau) \rightarrow L(x, t_0;\ s_0, \tau) \tag{93}
$$

then for *every positive direction* $u = (u_1, \dots, u_N, u_{N+1})$ in the $N+1$-dimensional space spanned by (s, t), the directional derivative $(\mathcal{D}_u L)(x, t;\ s, \tau)$ must satisfy

$$(\mathcal{D}_u L)(x_0, t_0;\ s_0, \tau_0) \leq 0 \qquad \text{at any local maximum,} \tag{94}$$

$$(\mathcal{D}_u L)(x_0, t_0;\ s_0, \tau_0) \geq 0 \qquad \text{at any local minimum.} \tag{95}$$

This formulation reflects the basic fact that for any given temporal moment t_0, the only information that is available for the visual front-end is the continuum of spatio-temporal scale-space representations over space x, spatial scales s and temporal scales τ. Thereby, no additional explicit memory of past information or access to the future is needed to make the requirement of non-enhancement of local extrema operational.

Temporal scale-space kernel in the context of discrete temporal scale levels. For a spatio-temporal scale-space representation involving a discrete set of scale levels only, the non-enhancement of local extrema condition can obviously not be applied. In this case, one can instead build on the requirement of non-creation of local extrema as expressed for a one-dimensional temporal signal depending on time t only. In analogy with the one-dimensional spatial case, let us therefore regard a one-dimensional temporal smoothing kernel T_{time} as a temporal scale-space kernel if and only if the kernel is time-causal and in addition for any purely temporal signal f, the number of local extrema in $T_{time} * f$ is guaranteed to not exceed the number of local extrema in f.

As will be shown later, axiomatic derivations show that both the non-causal and the time-causal spatio-temporal scale-space concepts give rise to spatio-temporal smoothing kernels of the form

$$T_{space-time}(x, t;\ s, v, \tau) = T_{space}(x - vt;\ s)\, T_{time}(t;\ \tau) \tag{96}$$

when combined with the requirement of Galilean covariance. If we therefore take this algebraic structure for granted, it therefore seems natural that the temporal smoothing kernel T_{time} should not be allowed to create new image structures in terms of new local extrema or new zero-crossings when applied to purely temporal image data.

7 Scale-space concepts for spatio-temporal image domains

7.1 Non-causal Gaussian spatio-temporal scale-space

If we for the purpose of analyzing pre-recorded video data allow for unlimited freedom of accessing image data at all temporal moments simultaneously, we can apply a similar way of reasoning as in section 5 with space x replaced by space-time $p = (x, t)$, thus disregarding both temporal causality and temporal recursivity.

Necessity result. Given image data $f \colon \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ defined over an N+1-dimensional spatio-temporal domain, let us therefore again assume that the first stage of visual processing as represented by the operator $\mathcal{T}_s$ should be (i) *linear*, (ii) *shift invariant* and (iii) obey a *semi-group structure over both spatial and temporal scales* s, where we also have to assume (iv) certain *regularity properties* of the semi-group $\mathcal{T}_s$ *over scale* s to guarantee sufficient differentiability properties with respect to space x, time t and spatio-temporal scales s.[8] Let us furthermore require (v) *non-enhancement of local extrema* to hold for *any* smooth image function $f \in C^\infty(\mathbb{R}^N \times \mathbb{R}) \cap L^1(\mathbb{R}^2 \times \mathbb{R})$ and for any positive scale direction s.

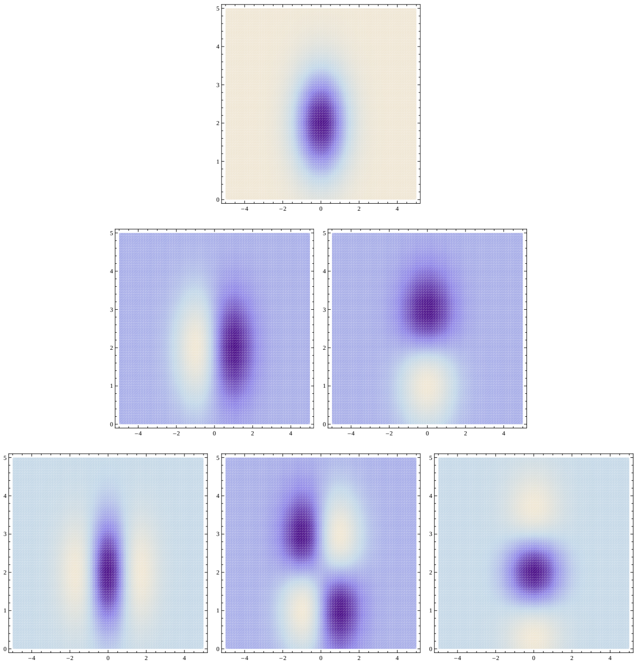


Figure 24: *Space-time separable kernels* $g_{x^\alpha t^\gamma}(x, t;\ s, \tau, \delta)$ up to order two obtained from the *Gaussian spatio-temporal scale-space* in the case of a 1+1-D space-time ($s = 1$, $\tau = 1$, $\delta = 2$). (Horizontal axis: space x. Vertical axis: time t.)

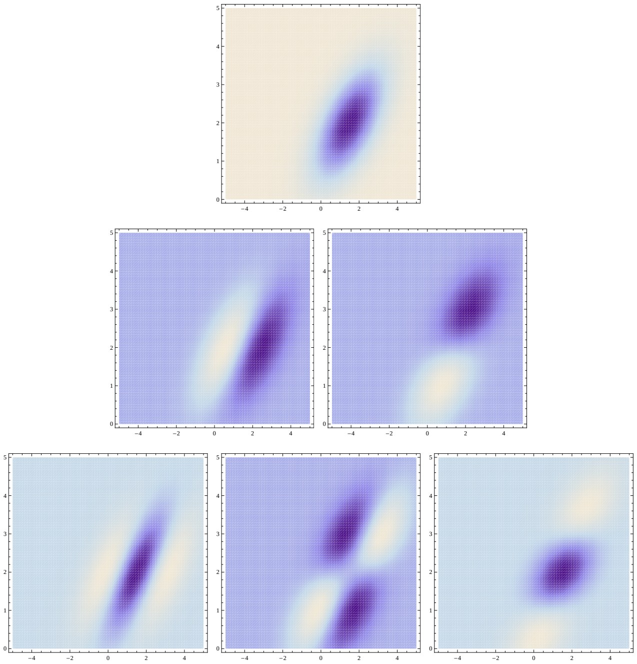
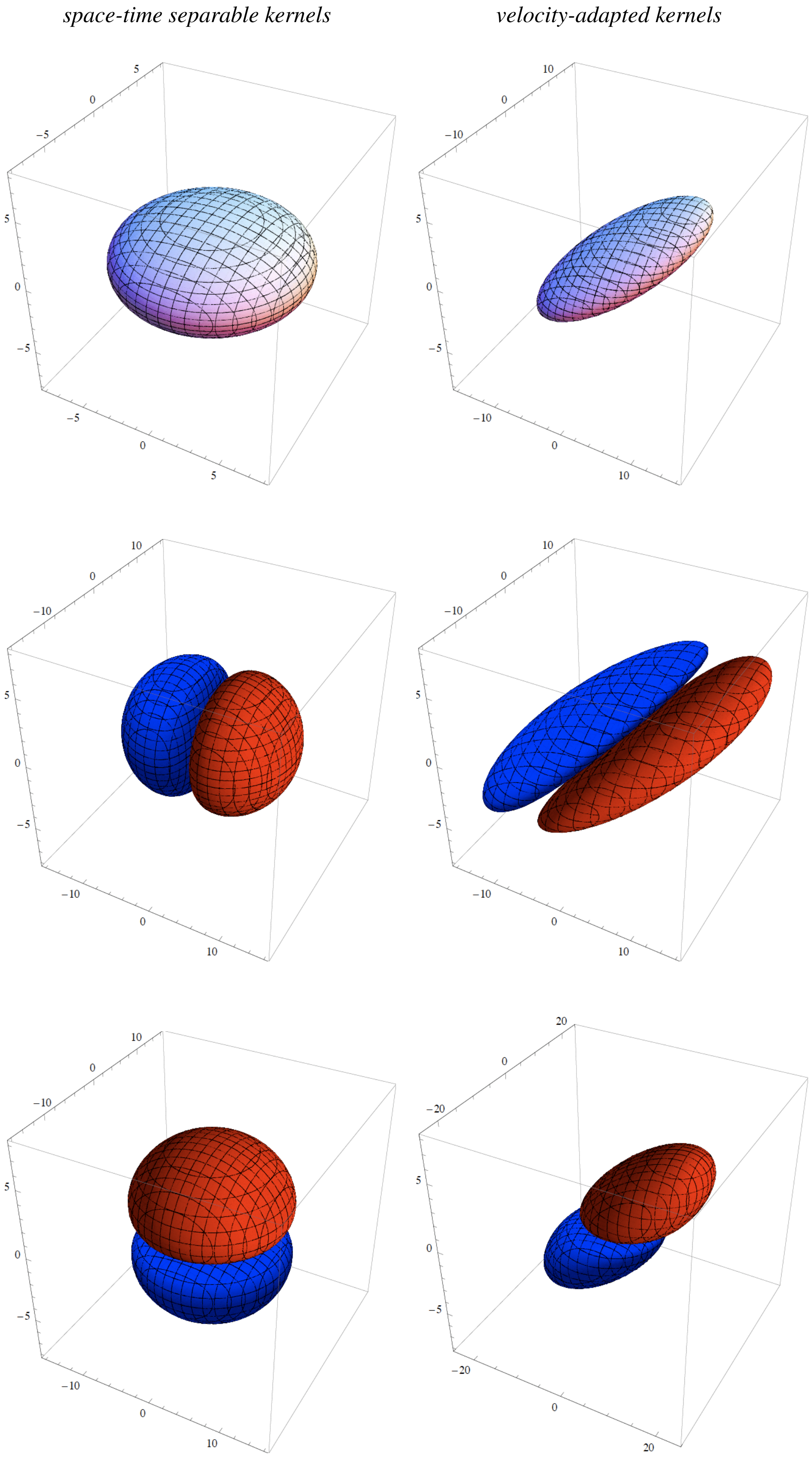


Figure 25: *Velocity-adapted spatio-temporal kernels* $g_{\bar{x}^\alpha \bar{t}^\gamma}(x, t;\ s, \tau, v, \delta)$ up to order two obtained from the *Gaussian spatio-temporal scale-space* in the case of a 1+1-D space-time ($s = 1$, $\tau = 1$, $v = 0.75$, $\delta = 2$). (Horizontal axis: space x. Vertical axis: time t.)


space-time separable kernels

velocity-adapted kernels

Figure 26: Gaussian spatio-temporal scale space kernels over a 2+1-D space-time. The left column shows *space-time separable kernels* with $v_{x_1} = 0$ and the right column corresponding *velocity-adapted kernels* with $v_{x_1} = 2$: (top row) Zero-order smoothing kernels $g(x_1, x_2, t;\ \Sigma, v)$, (middle row) First-order spatial derivative $g_{x_1}(x_1, x_2, t;\ \Sigma, v)$, (bottom row) First-order temporal derivative $g_{\bar{t}}(x_1, x_2, t;\ \Sigma, v)$. (Horizontal dimensions: space $x = (x_1, x_2)$. Vertical dimension: time t. Filter parameters: $\lambda_1 = \lambda_2 = 16$, $\lambda_t = 4$, $v_{x_2} = 0$ according to (100)).

Then, it follows from (Lindeberg 2011, theorem 5, page 42) that these conditions together imply that the scale-space family L must satisfy a diffusion equation of the form

$$\partial_s L = \frac{1}{2} \nabla_{(x,t)}^T \left(\Sigma_0 \nabla_{(x,t)} L \right) - \delta_0^T \nabla_{(x,t)} L \tag{97}$$

with the notation $\nabla_{(x,t)} = (\partial_{x_1}, \dots, \partial_{x_N}, \partial_t)^T$ for the spatio-temporal gradient operator, and with initial condition $L(x, t;\ 0, 0) = f(x, t)$ for some positive semi-definite $(N+1) \times (N+1)$ covariance matrix Σ_0 and for some $N + 1$-dimensional vector δ_0. Equivalently, this spatio-temporal scale-space representation at scale s can be obtained by convolution with *spatio-temporal Gaussian kernels* of the form

$$g(p;\ \Sigma_s, \delta_s) = \frac{1}{(2\pi)^{(N+1)/2} \sqrt{\det \Sigma_s}} \, e^{-(p-\delta_s)^T \Sigma_s^{-1} (p-\delta_s)/2s}. \tag{98}$$

with $p = (x, t)^T = (x_1, \dots, x_N, t)^T$, where the covariance matrix $\Sigma_s = s \, \Sigma_0$ constitutes a simultaneous covariance matrix over space and time and $\delta_s = s \, \delta_0$ denotes a corresponding translation vector over space and time.

Interpretation. By a suitable interpretation of the covariance matrix Σ_s and the offset δ_s, this non-causal scale-space concept can be used for modelling smoothing operations for spatio-temporal image data.

On a temporal domain, the non-zero offset in the Gaussian kernel over the temporal domain can be used as a simplified model of the fact that all computations require a non-zero computation time and time averages can only be computed from data that have occurred in the past. This requirement of temporal causality implies that any temporal receptive field has to be associated with a non-zero time delay, and introducing a temporal delay in the spatio-temporal smoothing operations constitutes a simple model of these effects within the paradigm based on Gaussian smoothing operations.

On a spatio-temporal domain, we may furthermore want the receptive fields to follow the direction of motion, in such a way that the centres and the shapes of the receptive fields are adapted to the direction of motion; see figure 27 for an illustration. Such *velocity adaptation* (Lindeberg 1997) is useful for reducing the temporal blur induced by observing objects that move relative to the camera and is a natural mechanism to include in modules for multi-scale motion estimation (Nagel & Gehrke 1998, Florack et al. 1998) and for recognizing spatio-temporal activities or events (Laptev & Lindeberg 2004*c*, Laptev & Lindeberg 2004*b*, Laptev et al. 2007).

With respect to temporal implementation, however, the filters in this Gaussian filter class do not respect temporal causality in a strict sense. Although the total mass of the filter coefficients that imply access to the future can be made arbitrarily small, by a suitable choice of time delay associated with the scale parameter in the scale direction, all filters in this

[8] To ensure sufficient differentiability properties such that an infinitesimal generator exists and the resulting multi-scale representation obtained by convolution with the semi-group of convolution kernels can be differentiated with respect to both space-stime and spatio-temporal scales such that the requirement of non-enhancement of local extrema can be applied, we do formally for an $N + 1$-dimensional space-time require the semi-group $\mathcal{T}_s$ to be C_1-*continuous* in the sense that $\lim_{h \downarrow 0} \left\| \frac{1}{h} \int_{s=0}^{h} \mathcal{T}(s) f \, ds - f \right\|_{H^k(\mathbb{R}^N \times \mathbb{R})} = 0$ should hold for some $k > (N+1)/2$ and for all smooth functions $f \in L^1(\mathbb{R}^N \times \mathbb{R}) \cap C^\infty(\mathbb{R}^N \times \mathbb{R})$ with $\| \cdot \|_{H^k(\mathbb{R}^2 \times \mathbb{R})}$ denoting the L^2-based Sobolev norm $\|u\|_{H^k(\mathbb{R}^N \times \mathbb{R})} = \left(\int_{\omega \in \mathbb{R}^N \times \mathbb{R}} \left(1 + |\omega|^2\right)^k \, |\hat{u}(\omega)|^2 d\omega \right)^{1/2}$ and $\hat{u}$ denoting the Fourier transform of u over $\mathbb{R}^N \times \mathbb{R}$; see (Lindeberg 2011, Section 3.2 and Appendix A) regarding details.

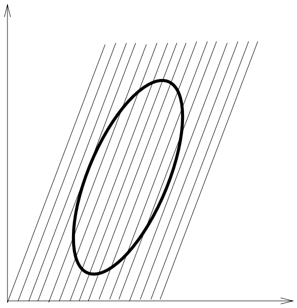


Figure 27: By adapting the shape and the position of a spatio-temporal smoothing kernel to the direction of motion, we can compute image descriptors that are invariant to constant velocity motion. This property can for example be used for reducing the effect of motion blur when computing image descriptors of moving objects at coarse temporal scales.

filter class have support regions that cover the entire time axis and are not suitable for real-time processing of temporal image data. Nevertheless, they are highly useful as the simplest possible model for studying properties of temporal and spatio-temporal scale-spaces. They are also highly useful for off-line processing. We shall later describe spatio-temporal scale-space concepts that respect temporal causality in a strict sense in section 7.2.

Parameterization of Gaussian spatio-temporal kernels for a 2+1-D spatio-temporal domain. By combining the parameterization of a general spatial 2×2 covariance matrix in equation (56) with a general Galilean transformation

$$\left\{ \begin{array}{l} x_1' = x_1 + v_1 t \\ x_2' = x_2 + v_2 t \\ t' = t \end{array} \right. \tag{99}$$

it can be shown (Lindeberg 2011, equation (61), page 46) that such a *spatio-temporal covariance matrix* can be parameterized as

$$\Sigma_s = \left(\begin{array}{ccc} \lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta + v_1^2 \lambda_t & (\lambda_2 - \lambda_1) \cos\theta \, \sin\theta + v_1 v_2 \lambda_t & v_1 \lambda_t \\ (\lambda_2 - \lambda_1) \cos\theta \, \sin\theta + v_1 v_2 \lambda_t & \lambda_1 \sin^2\theta + \lambda_2 \cos^2\theta + v_2^2 \lambda_t & v_2 \lambda_t \\ v_1 \lambda_t & v_2 \lambda_t & \lambda_t \end{array} \right) \tag{100}$$

where λ_1 and λ_2 are the eigenvalues of the spatial component of the covariance matrix with orientation θ and hence determine the spatial extent of the kernel, whereas $\lambda_t = \tau$ determines the temporal extent. The vector δ_s can in turn be parameterized as

$$\delta_s = \left(\begin{array}{c} v_1 t \\ v_2 t \\ \delta \end{array} \right) \tag{101}$$

where δ denotes a *temporal delay* associated with the spatio-temporal scale-space kernel. This parameter can be used for modelling the temporal delay that will be associated with any time-causal temporal kernel. Specifically, if we replace all the values of a temporal Gaussian scale-space kernel that would extend into the future by zeros, an increasing value of this temporal delay parameter will thereby reduce the influence of such truncation effects.

For the specific case with one spatial dimension and one temporal dimension, we obtain

$$\det \Sigma' = \lambda_x \lambda_t = s\tau \tag{102}$$

$$(X-\delta)^T \Sigma'^{-1} (X-\delta) = \frac{(x-vt)^2}{s} + \frac{(t-\delta_t)^2}{\tau} \tag{103}$$

which after insertion into equation (33) implies that these Gaussian spatio-temporal kernels assume the form

$$\begin{aligned} g(x,t;\ s,\tau,v,\delta) &= \frac{1}{\sqrt{2\pi s}} e^{-\frac{(x-vt)^2}{2s}} \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{(t-\delta)^2}{2\tau}} \\ &= g(x-vt;\ s)\, g(t;\ \tau,\delta). \end{aligned} \tag{104}$$

Velocity-adapted spatio-temporal derivative kernels. Corresponding *velocity-adapted spatio-temporal derivatives* are given by

$$\partial_{\bar{x}_1} = \partial_{x_1}, \quad \dots, \quad \partial_{\bar{x}_N} = \partial_{x_N}, \qquad \partial_{\bar{t}} = v^T \nabla_x + \partial_t = v_1\, \partial_{x_1} + \dots + v_N\, \partial_{x_N} + \partial_t. \tag{105}$$

Figures 24–25 show a examples of spatio-temporal scale-space kernels generated in this way in the case of a 1+1-dimensional space-time for (i) the space-time separable case with $v = 0$ and (ii) the non-separable case with a non-zero image velocity $v \neq 0$, in the special case when the spatial smoothing operation is rotationally symmetric ($\lambda_1 = \lambda_2$). Figure 26 shows corresponding kernels for a $2+1$-dimensional space-time.

Such *Gaussian spatio-temporal scale-space kernels* have been successfully used for computing spatio-temporal image features from video data (Laptev & Lindeberg 2003, Willems et al. 2008) and for performing spatio-temporal recognition (Laptev & Lindeberg 2004*a*, Kläser et al. 2008, Laptev et al. 2008, Wang et al. 2009, Shao & Mattivi 2010). Specifically, it was shown in (Laptev & Lindeberg 2004*c*, Lindeberg et al. 2004, Laptev et al. 2007) that the computation of *Galilean invariant image descriptors* improves the ability to perform image-based recognition under unknown relative motions between the objects/events and the observer. These Galilean invariant properties are made possible by the Galilean covariant property of the underlying spatio-temporal scale-space; see (Lindeberg 2011, Section 4.1.4 and Appendix C) for a formal proof regarding Galilean covariant fixed-points in a velocity-adapted spatio-temporal scale-space.

Combined Galilean and affine covariance. Consider two spatio-temporal patterns $f_L \colon \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ and $f_R \colon \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ that are related by a space-time transformation of the form

$$f_L(\xi) = f_R(\eta) \quad \text{where} \quad \eta = G_v\ (A\,\xi + b) = A\xi + vt + b \tag{106}$$

where

$$G_v = \begin{pmatrix} 1 & & & v_1 \\ & \ddots & & \vdots \\ & & 1 & v_N \\ & & & 1 \end{pmatrix} \tag{107}$$

corresponds to a Galilean transformation of space-time according to (99) and

$$A = \begin{pmatrix} a_{11} & \dots & a_{1N} & \\ \vdots & \ddots & \vdots & \\ a_{N1} & \dots & a_{NN} & \\ & & & 1 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_N \\ c \end{pmatrix} \tag{108}$$

represent an affine deformation over the spatial domain x determined by A and b complemented by a temporal shift $t_R = t_L + c$ given by c.

Define spatio-temporal scale-space representations L and R of these spatio-temporal patterns according to

$$L(\cdot;\ \Sigma_L, \delta_L) = g(\cdot;\ \Sigma_L, \delta_L) * f_L(\cdot), \quad R(\cdot;\ \Sigma_R, \delta_R) = g(\cdot;\ \Sigma_R, \delta_R) * f_R(\cdot), \tag{109}$$

where $g(\cdot;\ \Sigma, \delta)$ denote spatio-temporal Gaussian kernels of the form (98). Then, for corresponding points $p_L = (x_L, t_L)^T$ and $p_R = (x_R, t_R)^T$ in space-time, the spatio-temporal scale-space representations L and R will be related by

$$L(p_L;\ \Sigma_L, \delta_L) = R(p_R;\ \Sigma_R, \delta_R) \tag{110}$$

if the covariance matrices Σ_L and Σ_R satisfy

$$\Sigma_R = G_v\, A\, \Sigma_L\, A^T\, G_v^T \tag{111}$$

provided that the velocity terms δ_L and δ_R in the Gaussian kernels can be traded against coordinate shifts in p_L and p_R as long as the following relation is satisfied:

$$p_R - \delta_R = G_v\, A\, (p_L - \delta_L) + b. \tag{112}$$

Hence, this Gaussian spatio-temporal scale-space concept allows for simultaneous affine covariance over space and Galilean covariance over space-time. The general parameterization of the corresponding spatio-temporal kernels in equation (100) reflects this property, by simultaneously allowing for elongated and directionally adapted operations over the spatial domain and velocity-adapted operations along the direction of motion.

Velocity adaptation vs. image stabilization or filter banks When implementing a velocity-adapted spatio-temporal scale-space representation in practice, there are different alternatives to consider. The simplest approach is to use the same velocity vector at all image positions, and is equivalent to global stabilization. More generally, one may also consider different image velocities at different image positions.[9] In this way, the corresponding velocity-adapted spatio-temporal scale-space representations will for appropriate values of the velocity parameters correspond to filtering along the particle trajectories. Thereby, the system will be able to handle multiple moving objects and will also have the ability to derive a Galilean invariant representation for each object. Alternatively, we may at each image position even consider an *ensemble* of spatio-temporal filters that are tuned to different image velocities — a design with close relations to velocity-tuned receptive fields biological vision (see (Lindeberg 2011)). Such a parallel treatment of velocity adaption for different image velocities does also have the potential to handle transparent motion.

7.2 Time-causal spatio-temporal scale-space

When constructing a vision system for real-time processing of visual information, a fundamental constraint on the spatio-temporal smoothing kernels is that they have to be time-causal. As previously mentioned, the ad hoc solution of using a truncated symmetric filter

[9] A spatial counterpart of this idea has been developed in (Almansa & Lindeberg 2000), where the spatial covariance matrix in an affine scale-space representation is allowed to vary in space, to allow for enhancements of local directional image structures in fingerprint images.

of finite temporal extent in combination with a temporal delay is not appropriate in a time-critical context. Because of computational and memory efficiency, the computations should also be based on a compact temporal buffer that contains sufficient information for representing image information at multiple temporal scales and computing image feature therefrom. Corresponding requirements are also necessary in computational modelling of biological vision.

In this section, we shall describe two types of time-causal spatio-temporal scale-space concepts in the cases of (i) an *a priori* discretization of the temporal scale parameter and (ii) a continuous scale parameter.

7.2.1 Time-causal spatio-temporal scale-space based on discrete temporal scale levels (without true temporal covariance)

Time-causal scale-space kernels for a purely temporal domain. Given the requirement on a temporal scale-space kernel in terms of non-creation of local extrema over a purely temporal domain, truncated exponential kernels

$$h_{exp}(t;\ \mu_i) = \begin{cases} \frac{1}{\mu_i} e^{-t/\mu_i} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{113}$$

arise as a basic model for modelling temporal smoothing operations over a continuous time-causal temporal domain (Lindeberg 1990, Lindeberg & Fagerström 1996). The Laplace transform of such a kernel is given by

$$H_{exp}(q;\ \mu_i) = \int_{t=-\infty}^{\infty} h_{exp}(t;\ \mu_i)\, e^{-qt}\, dt = \frac{1}{1+\mu_i q} \tag{114}$$

By coupling k such kernels in cascade, we obtain a composed filter

$$h_{composed}(t;\ \mu) = *_{i=1}^{k} h_{exp}(t;\ \mu_i) \tag{115}$$

having a Laplace transform of the form

$$H_{composed}(q;\ \mu) = \int_{t=-\infty}^{\infty} (*_{i=1}^{k} h_{exp}(t;\ \mu_i))\, e^{-qt}\, dt = \prod_{i=1}^{k} \frac{1}{1+\mu_i q} \tag{116}$$

The composed filter has mean value

$$M(h_{composed}(\cdot;\ \mu)) = \sum_{t=1}^{k} \mu_i \tag{117}$$

and variance

$$\tau_k = V(h_{composed}(\cdot;\ \mu)) = \sum_{t=1}^{k} \mu_i^2 \tag{118}$$

In terms of physical models, repeated convolution with this set of truncated exponential kernels corresponds to coupling a series of first-order integrators with time constants μ_k in cascade:

$$\partial_t L(t;\ \tau_k) = \frac{1}{\mu_k} \left(L(t;\ \tau_{k-1}) - L(t;\ \tau_k)\right) \tag{119}$$

with $L(t;\ 0) = f(t)$. These temporal smoothing kernels satisfy scale-space properties in the sense that the number of local extrema or the number of zero-crossings in the temporal signal are guaranteed to not increase with the temporal scale. In this respect, these kernels have a desirable and well-founded smoothing property that can be used for defining multi-scale observations over time. A limitation of this type of temporal scale-space representation, however, is that the *scale levels are required to be discrete* and that the scale-space representation does hence not admit a continuous scale parameter. In this respect, this discrete temporal scale-space representation differs from the other temporal scale-space concepts considered here. Specifically, the lack of scale continuity implies that the transformation properties under rescalings of time will be more complex than a regular covariance property under scaling transformations. Moreover, this temporal scale-space having only discrete scale levels cannot satisfy a differential equation over temporal scales. Computationally, however, the scale-space representation based on truncated exponential kernels can be highly efficient and admits for direct implementation in terms of hardware (or wetware) that emulates first-order integration over time (see figure 28 for illustration of a corresponding electric wiring diagram corresponding to a set of first-order integrators coupled in cascade).

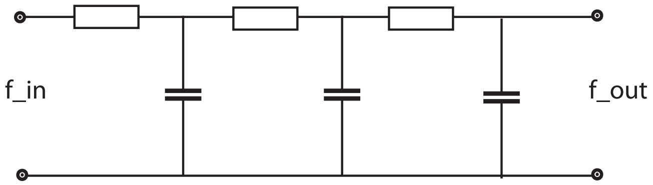


Figure 28: Electric wiring diagram consisting of a set of resistors and capacitors that emulate a series of first-order integrators coupled in cascade, if we regard the time-varying voltage f_{in} as representing the time varying input signal and the resulting output voltage f_{out} as representing the time varying output signal at a coarser temporal scale. According to the theory of temporal scale-space kernels for one-dimensional signals (Lindeberg 1990, Lindeberg & Fagerström 1996), the corresponding equivalent truncated exponential kernels are the only primitive temporal smoothing kernels that guarantee both temporal causality and non-creation of local extrema (alternatively zero-crossings) with increasing temporal scale. Such first-order temporal integration can be used as a straightforward computational model for temporal processing in biological neurons (see also (Koch 1999, Chapters 11–12) regarding physical modelling of the information transfer in dendrites of neurons).

Time-recursive computation of temporal derivatives. Temporal scale-space derivatives of order r can be defined from this scale-space model according to

$$L_{t^r}(\cdot;\ \tau_k) = \partial_{t^k} L(\cdot;\ \tau_k) = (\partial_{t^k} (*_{i=1}^{k} h_{exp}(t;\ \mu_i)) * f \qquad (120)$$

and the Laplace transform of the composed (equivalent) derivative kernel is

$$H_{composed}^{(r)}(q;\ \tau_k) = q^r \prod_{i=1}^{k} \frac{1}{1 + \mu_i q} \qquad (121)$$

For this kernel to have a net integration effect, and to enable well-posed derivative operators, an obvious requirement is that the total order of differentiation should not exceed the total order of integration. Thereby, $r < k$ is a necessary requirement. As a consequence, the composed transfer function must have a finite L_2-norm.

A very useful observation that can be made concerning derivative computations is that temporal derivatives can equivalently be computed from differences between different temporal channels. Let us first assume that all time constants μ_i are different in (121). Then, a

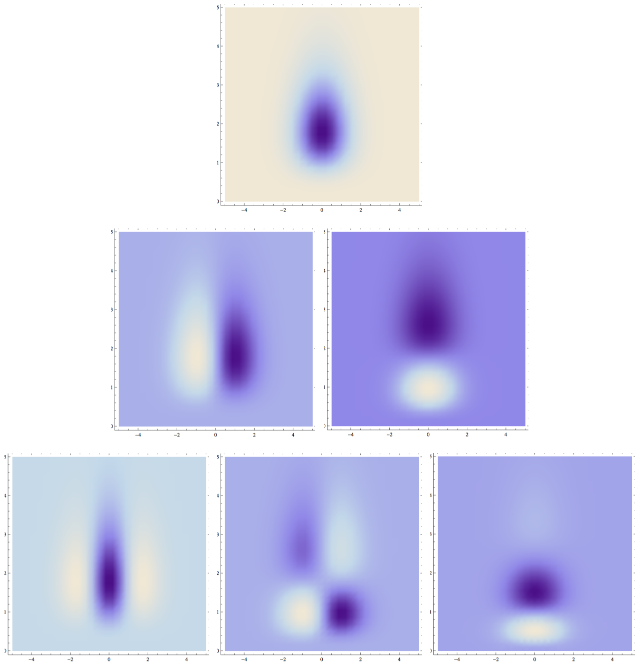


Figure 29: *Space-time separable kernels* $g_{x^\alpha t^\gamma}(x,t;\ s,\tau)$ up to order two corresponding to the combination of a cascade of $k = 7$ time-causal and time-recursive first-order integrators over the temporal domain with a Gaussian scale-space over the spatial domain in the case of a 1+1-D space-time ($s = 1$, $\tau = 1$) and using a self-similar distribution of the scale levels according to equations (129) and (131). (Horizontal axis: space x. Vertical axis: time t.)

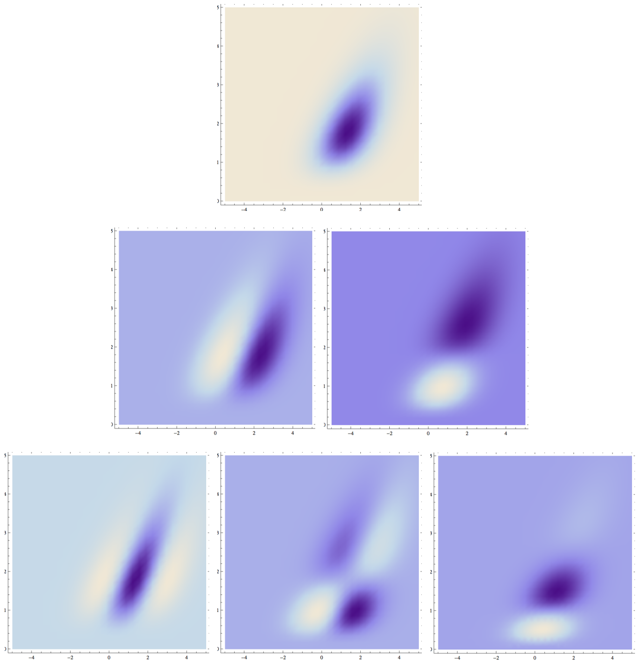


Figure 30: *Velocity-adapted spatio-temporal kernels* $g_{\bar{x}^\alpha \bar{t}^\gamma}(x,t;\ s,\tau,v)$ up to order two obtained by combining a cascade of $k = 7$ time-causal and time-recursive first-order integrators over the temporal domain with a Gaussian scale-space over the spatial domain in the case of a 1+1-D space-time ($s = 1$, $\tau = 1$, $v = 0.75$) and using a self-similar distribution of the scale levels according to equations (129) and (131). (Horizontal axis: space x. Vertical axis: time t.)

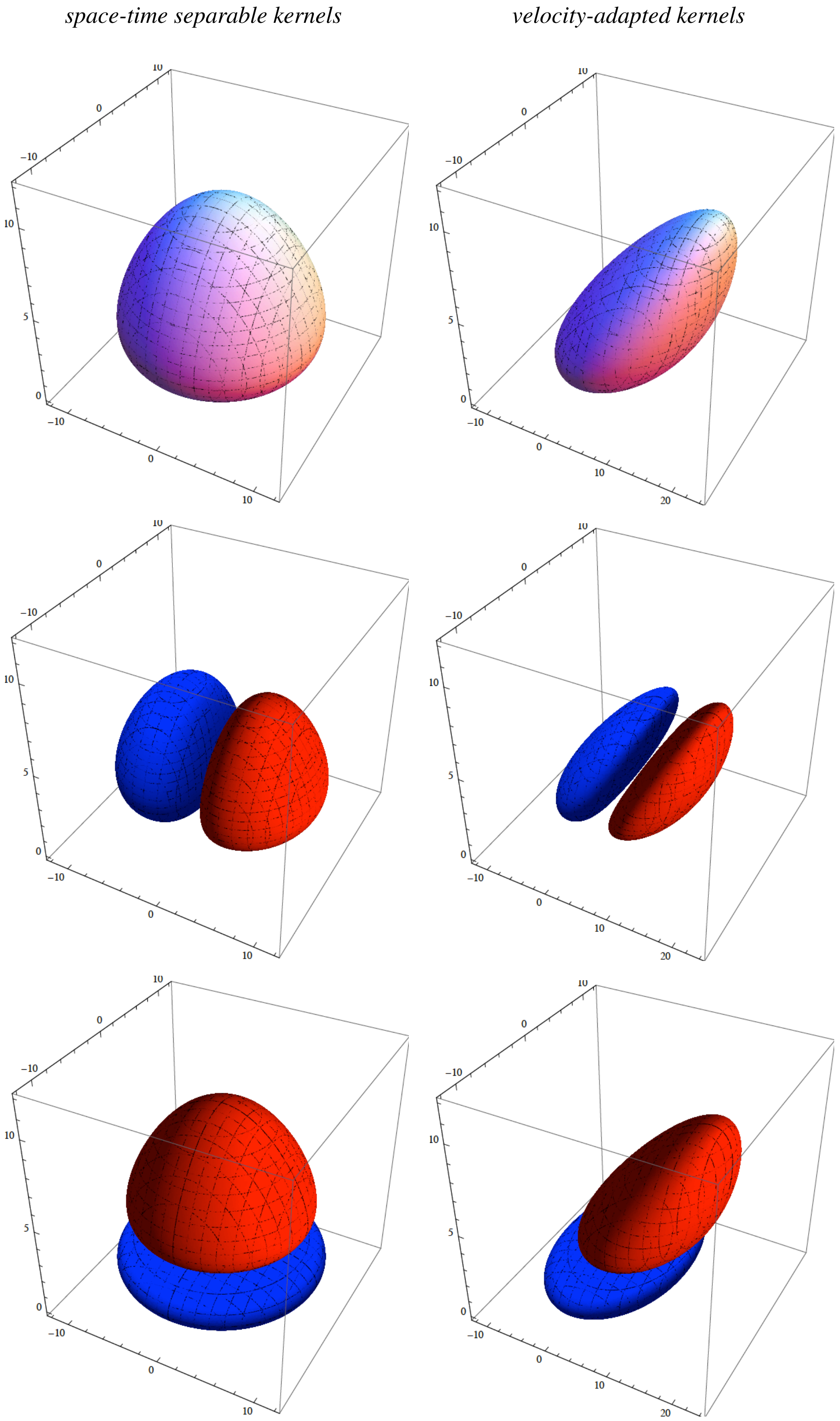


Figure 31: Time-causal and time-recursive spatio-temporal scale-space kernels over a 2+1-D space-time obtained by combining $k = 7$ first-order integrators over the temporal domain with Gaussian smoothing over the spatial domain for (left column) the space-time separable case with $v = 0$ and (right column) the velocity-adapted case with $v \neq 0$: (top row) Original smoothing kernel $h(x, y, t;\ \Sigma, v, \tau)$, (middle row) First-order spatial derivative $h_x(x, y, t;\ \Sigma, v, \tau)$, (bottom row) First-order temporal derivative $h_{\bar{t}}(x, y, t;\ \Sigma, v, \tau)$. (Parameters: $\lambda_1 = \lambda_2 = 16$, $\tau = 4$, $v_x = 0$ *vs.* $3/2$ $v_y = 0$).

partial fraction division gives

$$H^{(r)}_{composed}(q;\ \tau_k) = \sum_{i=1}^{k} A_i \, H_{prim}(q;\ \mu_i) \tag{122}$$

where

$$A_i = \frac{(-1)^r}{\mu_i^r} \prod_{j=1, j \neq i}^{k} \frac{1}{(1 - \mu_j/\mu_i)} \qquad (1 \leq i \leq k) \tag{123}$$

showing that *each temporal derivative can be computed as a linear combination of the representations at the different time-scales*.

More realistically, however, the channels that we can regard as available at a certain temporal scale with index k will not be the results of direct filtering with different time constants μ_i. Rather, we would like to use the intermediate outputs from the cascade coupled recursive filters $H_{composed}(q;\ \tau_i)$ for $k - r \leq i \leq k$. Decomposition of $H^{(r)}_{composed}$ into a sum of r such transfer functions

$$H^{(r)}_{composed}(q;\ \tau_k) = \sum_{i=k-r}^{k} B_i \, H_{composed}(q;\ \tau_i) \tag{124}$$

shows that the weights B_i are given as the solution of a triangular system of equations provided that the necessary condition $r < k$ is satisfied

$$\frac{(-1)^r}{\mu_i^r} \prod_{j=i+1}^{k} \frac{1}{(1 - \mu_j/\mu_i)} = B_i + \sum_{\nu=i+1}^{k} B_\nu \prod_{j=i+1}^{\nu} \frac{1}{(1 - \mu_j/\mu_i)} \qquad (k - r \leq i \leq k). \tag{125}$$

After a few more calculations it can be shown that the Laplace transforms of the equivalent derivative computation kernels satisfy the recurrence relation (Lindeberg & Fagerström 1996)

$$H^{(r)}_{composed}(q;\ \tau_k) = \frac{1}{\mu_k} \left(H^{(r-1)}_{composed}(q;\ \tau_{k-1}) - H^{(r-1)}_{composed}(q;\ \tau_k) \right), \tag{126}$$

implying that higher-order temporal derivatives can be computed from small-support finite differences of lower-order derivatives (analogous to pure finite differences in the spatial domain) where the temporal scale-space representations at different temporal scales serve as a sufficient temporal buffer of what has occurred in the past (see figure 32 for a schematic illustration). Derivative computations will therefore be highly efficient. Specifically, it follows that both the temporal smoothing operation and the computation of temporal derivatives are time-recursive.

Spatio-temporal extension. The axiomatic restriction of temporal scale-space kernels to truncated exponential kernels coupled in cascade has been made for the case of a one-dimensional temporal domain (Lindeberg & Fagerström 1996) based on a more general classification of continuous scale-space kernels (Lindeberg 1990) building upon earlier results by (Schoenberg 1950). When applying this result to spatio-temporal image data, we can make analogies with the axiomatic derivations of the non-causal Gaussian spatio-temporal scale-space in section 7.1 and the time-causal spatio-temporal scale-space concept to be considered in section 7.2.2 for which the requirement of Galilean covariance do both imply that the spatio-temporal scale-space kernels should be of the form

$$T_{space-time}(x, t;\ s, v, \tau) = T_{space}(x - vt;\ s) \, T_{time}(t;\ \tau) \tag{127}$$

where $T_{space}(x;\ s)$ denotes a spatial Gaussian scale-space kernel with scale parameter s, v denotes an image velocity and $T_{time}(t;\ \tau)$ denotes a temporal smoothing kernel with temporal scale parameter τ. Let us therefore define spatio-temporal extensions of the truncated exponential filters coupled in cascade over an $N+1$-dimensional spatio-temporal domain according to

$$T(x,t;\ s,v,\tau) = g(x - vt;\ s)\, h_{composed}(t;\ k) \tag{128}$$

where $g(x;\ s)$ denotes an N-dimensional spatial Gaussian kernel and $h_{composed}(t;\ \tau_k)$ the convolution of k truncated exponential filters with time constants μ_i according to (115).

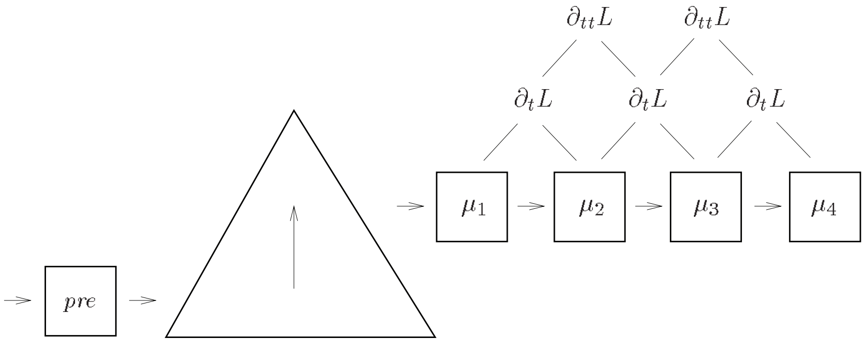


Figure 32: Composed architecture for a spatio-temporal visual front-end based on a time-causal scale-space concept defined from a set of truncated exponential filters corresponding to first-order integrators coupled in cascade: (i) Optional temporal preprocessing stage. (ii) Spatial scale-space representation. (iii) One set of time-recursive temporal smoothing stages associated with each spatial scale. (iv) Temporal derivatives obtained from linear combinations of temporal channels.

In the absence of further information, it is natural to distribute the temporal scale levels according to a geometric series, corresponding to a uniform distribution in units of *effective temporal scale* $\tau_{eff} = \log \tau$:

$$\tau_k = \gamma^{k-1}\, \tau_{min} \qquad \text{where} \qquad \gamma = \left(\frac{\tau_{max}}{\tau_{min}} \right)^{\frac{1}{K-1}} \qquad (1 \leq k \leq K) \tag{129}$$

which by the additive property of variances between adjacent scales

$$\tau_{k+1} = \tau_k + \mu_k^2 \tag{130}$$

and implies that the time constants of the individual temporal smoothing stages should be chosen according to

$$\mu_k = \sqrt{\tau_{min}\,(\gamma - 1)}\, \gamma^{(k-1)/2}. \tag{131}$$

Figures 29 and 30 show examples of spatio-temporal kernels and spatio-temporal derivatives of these computed in this way in the case of a $1+1$-dimensional space-time. Figure 31 shows examples of corresponding kernels over a $2+1$-dimensional space-time.

7.2.2 Time-causal spatio-temporal scale-space based on continuous temporal scale parameter (with true temporal covariance)

To process *real-time image data* with a spatio-temporal scale-space concept having a continuous temporal scale parameter, we do in addition to (i) *linearity* and (ii) spatial and temporal *shift-invariance* require the scale-space kernels to be (iii) *time-causal* and require the visual front-end to be (iv) *time-recursive* in the sense that the internal image representations $L(x,t;\ s,\tau)$ at different spatial scales s and temporal scales τ to also constitute a sufficient internal temporal memory $M(x,t)$ of the past, without any further need for temporal buffering. To adapt the convolution semi-group structure to a time-recursive setting, we require the spatio-temporal scale-space concept

$$L(\cdot,t;\ s,\cdot) = \mathcal{T}_{s,t}\, L(\cdot,0;\ 0,\cdot) \tag{132}$$

to be generated by a (v) *two-parameter semi-group* over spatial scales s and time s

$$\mathcal{T}_{s_1,t_1}\, \mathcal{T}_{s_2,t_2} = \mathcal{T}_{s_1+s_2,t_1+t_2}. \tag{133}$$

Then, it can be shown (Lindeberg 2011, theorem 17, page 78) that provided we express (vi) certain *regularity properties* on the semi-group in terms of Sobolev norms, the requirement of (vii) the *time-recursive formulation of non-enhancement of local extrema* in equations (94)–(95) implies that the semi-group must satisfy the following system of diffusion equations

$$\partial_s L = \frac{1}{2}\nabla_x^T(\Sigma\nabla_x L), \tag{134}$$

$$\partial_t L = -v^T\nabla_x L + \frac{1}{2}\partial_{\tau\tau} L \tag{135}$$

and combined boundary and initial condition $L(x,t;0,0;\ \Sigma,v) = f(x,t)$.

A physical interpretation of the second equation in this scale-space concept, given fixed values of the spatial scale parameter s and the covariance matrix Σ, is that the time-varying spatio-temporal image data $f(x,t)$ are considered as a time-varying thermal distribution that constitutes the boundary condition at $\tau = 0$ for heat diffusion over time t in an $N+1$-dimensional solid with extent over x and τ, where the thermal diffusion with respect to real time t only takes place in the τ-direction and the parameter v corresponds to an overall spatial drift velocity with respect to real time t. The internal representations at distance τ from the boundary of the solid do thereby correspond to the image representations at coarser temporal scales τ. The combination of the second equation with the first equation then corresponds to an additional diffusion process that operates over the spatial domain only and which is evolving with respect to a second diffusion evolution parameter ("additional virtual time") s, with the spatial covariance matrix Σ describing the spatial anisotropy of this medium.

Time-causal spatio-temporal scale-space kernels. In terms of integral expressions, it can be shown (Lindeberg 2011, equations (90)–(93), page 53) that the solution of these equations with initial condition $L(x,0;0,\tau;\ \Sigma,v) = 0$ and combined boundary and initial condition $L(x,t;0,0;\ \Sigma,v) = f(x,t)$ can be written

$$L(x,t;\ s,\tau;\ \Sigma,v) = \int_{u=0}^{t}\int_{\xi\in\mathbb{R}^N} f(\xi,u)\, h(x-\xi,t-u;s,\tau;\ \Sigma,v)\, d\xi\, du \tag{136}$$

where the notation with double semi-colons in the list of variables indicates that s and τ are parameters while Σ and v are meta-parameters. The convolution kernel h is in turn given by

$$\begin{aligned} h(x,t;\ s,\tau;\ \Sigma,v) &= g(x-vt;\ s;\ \Sigma)\,\phi(t;\ \tau) \\ &= \frac{1}{(2\pi s)^{N/2}\sqrt{\det\Sigma}}\, e^{-(x-vt)^T\Sigma^{-1}(x-vt)/2s}\, \frac{1}{\sqrt{2\pi}\,t^{3/2}}\,\tau\, e^{-\tau^2/2t} \end{aligned} \tag{137}$$

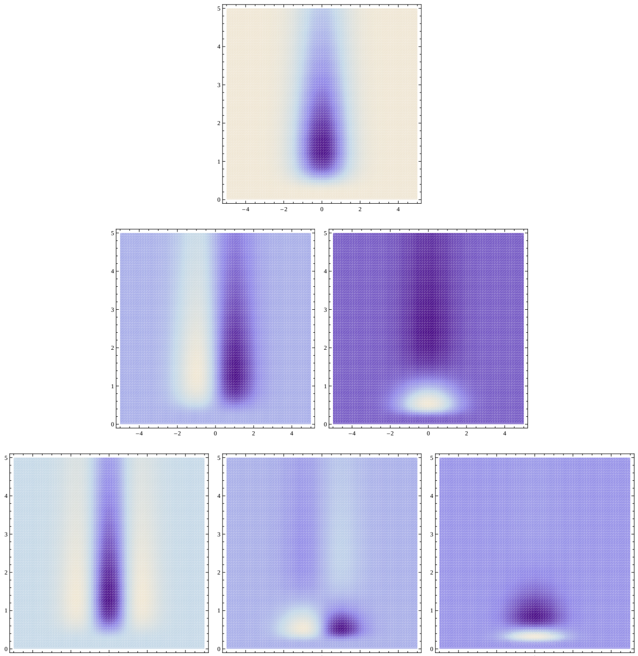


Figure 33: *Space-time separable kernels* $h_{x^\alpha t^\gamma}(x, t;\ s, \tau, v)$ up to order two obtained from the *time-causal spatio-temporal scale-space* in the case of a 1+1-D space-time ($s = 1$, $\tau = 2$). (Horizontal axis: space x. Vertical axis: time t.)

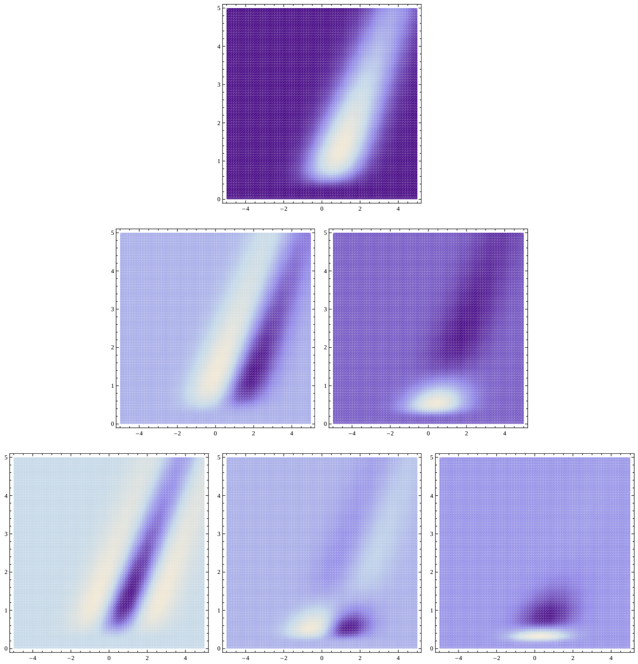


Figure 34: *Velocity-adapted spatio-temporal kernels* $h_{\bar{x}^\alpha \bar{t}'^\gamma}(x, t;\ s, \tau, v)$ up to order two obtained from the *time-causal spatio-temporal scale-space* in the case of a 1+1-D space-time ($s = 1$, $\tau = 2$, $v = 0.75$). (Horizontal axis: space x. Vertical axis: time t.)

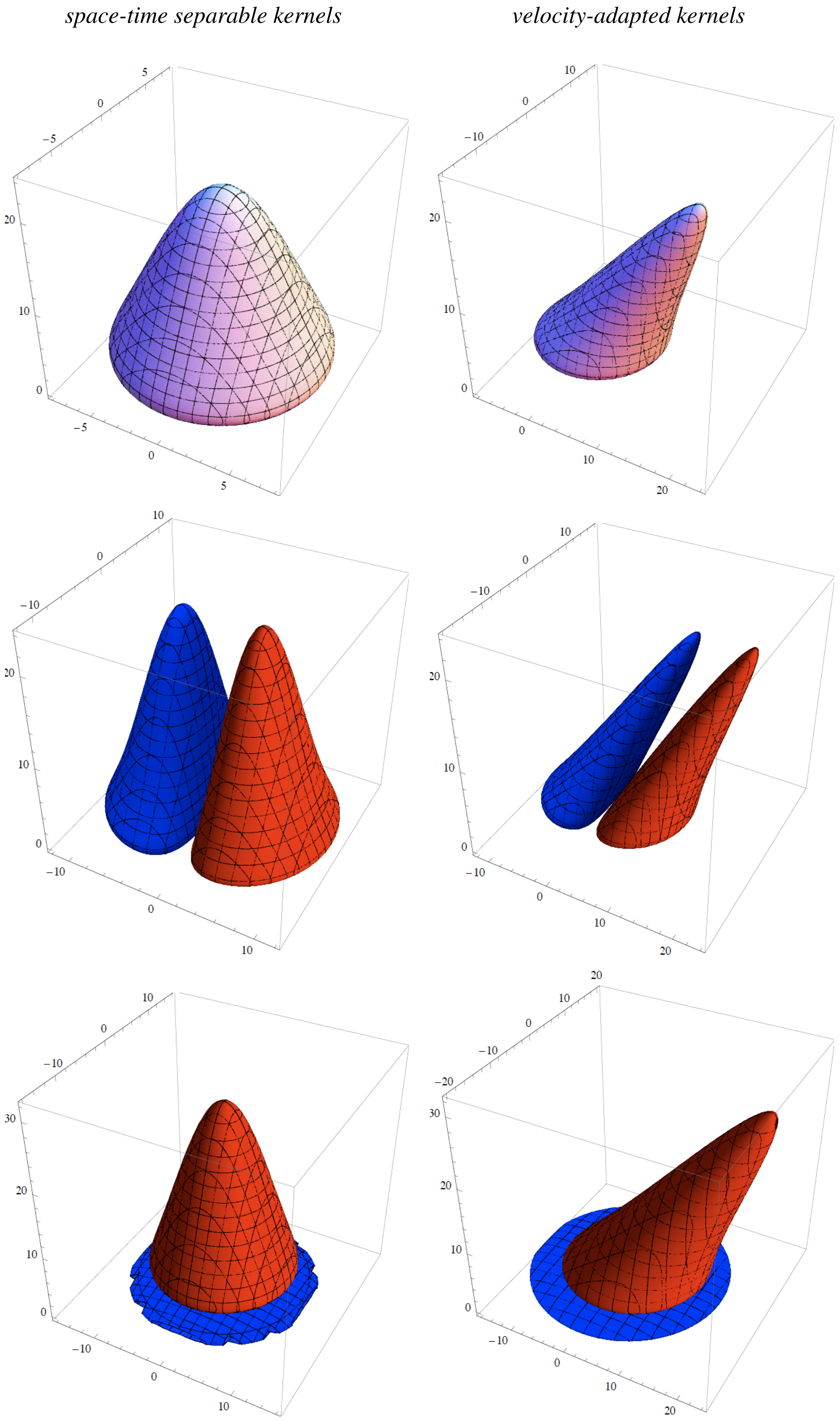


Figure 35: Time-causal spatio-temporal scale-space kernels over a 2+1-D space-time for (left column) the space-time separable case with $v = 0$ and (right column) the velocity-adapted case with $v \neq 0$: (top row) Original smoothing kernel $h(x, y, t;\ \Sigma, v, \tau)$, (middle row) First-order spatial derivative $h_x(x, y, t;\ \Sigma, v, \tau)$, (bottom row) First-order temporal derivative $h_{\bar{t}}(x, y, t;\ \Sigma, v, \tau)$. (Parameters: $\lambda_1 = \lambda_2 = 16$, $\tau = 2$, $v_x = 0$ *vs.* $3/4$ $v_y = 0$).

where $g(x - vt;\ s;\ \Sigma)$ corresponds to an N-dimensional affine Gaussian kernel with covariance matrix Σ that moves an image velocity v in space-time, and $\phi(t;\ \tau)$ denotes *a time-causal smoothing kernel over time t* with temporal scale parameter τ. Hence, given original image data of dimensionality $N + 1$, the time-causal scale-space representation will (at least) comprise $N + 3$ dimensions.[10]

This form of time-causal spatio-temporal scale-space has also been considered by (Fagerström 2007) in the special case when $\Sigma = I$ as a member in a family of self-similar kernels derived from the different argument of scale invariance in combination with a semi-group structure. The additional degree of freedom in the spatial covariance matrix Σ obtained here has the additional advantage that it allows for non-isotropic smoothing kernels over the spatial domain, which may be useful when dealing with local image deformations over time or when considering motion boundaries.

The interpretation of this temporal scale parameter is, however, somewhat different than for the previously considered Gaussian spatio-temporal concept. Specifically, the *temporal delay* associated with the temporal smoothing kernel is (Lindeberg 2011, equation (123), page 57)

$$\delta = \frac{\int_{t=0}^{\infty} t\, \phi^2(t;\ \tau)\, dt}{\int_{t=0}^{\infty} \phi^2(t;\ \tau)\, dt} = \tau^2. \tag{138}$$

Time-causal spatio-temporal derivative operators. Figures 33–34 show examples of such time-causal spatio-temporal kernels with their partial spatio-temporal derivatives for a $1+1$-dimensional spatio-temporal domain in the space-time separable case with $v = 0$

$$(\partial_{x^\alpha t^\beta} h)(x, t;\ s, \tau;\ \Sigma, 0) = (\partial_{x^\alpha} g)(x;\ s;\ \Sigma)\, (\partial_{t^\beta} \phi)(t;\ \tau) \tag{139}$$

and for the non-separable velocity-adapted case with $v \neq 0$

$$(\partial_{x^\alpha \bar{t}^\beta} h)(x, t;\ s, \tau;\ \Sigma, v) = (\partial_{x^\alpha} g)(x - vt;\ s;\ \Sigma)\, (\partial_{t^\beta} \phi)(t;\ \tau). \tag{140}$$

Figure 35 shows corresponding kernels over a $2 + 1$-dimensional space-time in the special case when the spatial covariance matrix is proportional to the unit matrix.

Relations to regular Gaussian smoothing We can note that there is also a very close link to regular Gaussian smoothing. By inspection, it can be seen that the time-causal spatio-temporal smoothing can be interpreted as as a first-order derivative with respect to temporal scale τ of a one-dimensional Gaussian over temporal scales, i.e.

$$\phi(t;\ \tau) = -\partial_\tau g(\tau;\ t), \tag{141}$$

and an N-dimensional Galilean-transformed affine Gaussian kernel

$$g_N(x - vt;\ \Sigma) = \frac{1}{(\sqrt{2\pi})^N \sqrt{\det \Sigma}}\, e^{-(x-vt)^T \Sigma^{-1} (x-vt)/2} \tag{142}$$

[10] If the full group of spatial covariance matrices Σ and velocity vectors v is considered as well, the dimensionality of the affine- and velocity-adapted scale-space will be $\dim(x) + \dim(t) + \dim(\Sigma) + \dim(v) + \dim(\tau) = N + 1 + N(N+1)/2 + N + 1 = (N^2 + 5N + 4)/2$. To handle such high-dimensional scale-spaces in practice, some sorts of intelligent search strategies are obviously required, such as combinations of lower-dimensional subgroups. The affine shape adaptation and Galilean velocity adaptation algorithms constitute examples of such simplifying search strategies. When using a massively parallel architecture, such as in biological vision, however, one could afford to represent a richer family of affine-adapted and/or velocity-adapted filters than would be possible to handle with a serial single-core computer.

over space x. For sake of convenience, we will henceforth change to the following notation:

$$L(x,t;\ \Sigma,v,\tau) = \int_{u=0}^{t} \int_{\xi\in\mathbb{R}^N} f(\xi,u)\, h(x-\xi,t-u;\Sigma,v,\tau)\, d\xi\, du \tag{143}$$

where

$$h(x,t;\ \Sigma,v,\tau) = g_N(x-vt;\ \Sigma)\, \phi(t;\ \tau) \tag{144}$$

and

$$g_N(x;\ \Sigma) = \frac{1}{(\sqrt{2\pi})^N \sqrt{\det\Sigma}}\, e^{-x^T\Sigma^{-1}x/2} \tag{145}$$

$$\phi(t;\ \tau) = -\partial_\tau g_1(\tau;\ t) = \frac{1}{\sqrt{2\pi}\, t^{3/2}}\, e^{-\tau^2/2t} \tag{146}$$

Please, note the shift of the order of the arguments between ϕ and g_1 in equation (146).

Temporal cascade-recursive formulation. When computing a spatio-temporal scale-space representation at time $t_2 > t_1$, a very attractive property is if this can be done in a *time-recursive* manner, such that it sufficient to use the following sources of information:

- the internal buffer of the spatio-temporal scale-space representation L at time t_1, and
- information about the spatio-temporal input data f during the time interval $[t_1, t_2]$.

This property means that it is sufficient to use the internal states of the spatio-temporal scale-space representation as internal memory, *and we do not need to have any complementary buffer of what else has occurred in the past.*

Such a property can indeed be established for the time-causal scale-space representation, based the fact that the time-causal scale-space kernel $\phi(t;\ \tau)$ satisfies the following time-recursive cascade smoothing property over a pure temporal domain (derived in (Lindeberg 2011, Appendix D.3)):

$$\phi(t_2;\ \tau) = \int_{\zeta=0}^{\infty} \phi(t_1;\ \zeta)\, (g(\tau-\zeta;t_2-t_1) - g(\tau+\zeta;\ t_2-t_1))\ d\zeta \tag{147}$$

From this relation it follows that the time-causal spatio-temporal scale-space representation satisfies the following cascade-recursive structure over time t and spatial scales s

$$\begin{aligned} L(x,t_2;s_2,\tau) = & \int_{\xi\in\mathbb{R}^N} \int_{\zeta\geq 0} T(x-\xi,t_2-t_1;\ s_2-s_1,\tau,\zeta)\, L(\xi,t_1;\ s_1,\zeta)\, d\zeta\, d\xi \\ & + \int_{\xi\in\mathbb{R}^N} \int_{u=t_1}^{t_2} B(x-\xi,t_2-u;\ s_2,\tau)\, f(\xi,u)\, d\xi\, du \end{aligned} \tag{148}$$

where the kernel T for updating the internal memory representation L is given by

$$T(x,t;\ s,\tau,\zeta) = g_N(x-vt;\ s)\ (g(\tau-\zeta;t) - g(\tau+\zeta;\ t)) \tag{149}$$

and the kernel B for incorporating new knowledge from the input signal f at the boundary is

$$B(x,t;\ s,\tau) = g_N(x-vt;\ s)\, \phi(t;\ \tau). \tag{150}$$

Please, note that we have here dropped the arguments for the meta-parameters Σ and v in order to simplify the notation.

Properties of the time-causal smoothing functions. To describe the evolution properties over temporal scales τ is however somewhat different than for the Gaussian spatio-temporal scale-space. Whereas the integral of h over space-time is finite

$$\int_{t=0}^{\infty} \int_{x \in \mathbb{R}^N} h(x, t;\ \Sigma, v, \tau)\, dx\, dt = 1, \tag{151}$$

we cannot compute regular first- or second-order moments of h over time t, since the corresponding integrals do not converge

$$\int_{t=0}^{\infty} \int_{x \in \mathbb{R}^N} t\, h(x, t;\ \Sigma, v, \tau)\, dx\, dt \to \infty, \tag{152}$$

$$\int_{t=0}^{\infty} \int_{x \in \mathbb{R}^N} t^2\, h(x, t;\ \Sigma, v, \tau)\, dx\, dt \to \infty. \tag{153}$$

Hence, we cannot parameterize the time-causal kernels h in terms of mean vectors and covariance matrices over space-time, as is a natural approach for the other spatio-temporal scale-spaces considered in this article, based on non-causal spatio-temporal Gaussian kernels or truncated exponential kernels coupled in cascade in combination with a spatial scale-space and velocity adaptation. Nevertheless, we can compute the position in space-time of the local maximum of $h(x, t;\ \Sigma, v, \tau)$

$$\left(\begin{array}{c} \hat{x} \\ \hat{t} \end{array} \right) = \frac{1}{3} \left(\begin{array}{c} v \\ 1 \end{array} \right) \tau^2 \tag{154}$$

and we can also compute the spatial mean $\bar{x}$ and the spatial covariance matrix $C(x, x)$ as

$$\bar{x} = \frac{\int_{\tau=0}^{\infty} \int_{x \in \mathbb{R}^N} x\, h(x, t;\ \Sigma, v, \tau)\, dx\, dt}{\int_{\tau=0}^{\infty} \int_{x \in \mathbb{R}^N} h(x, t;\ \Sigma, v, \tau)\, dx\, dt} = vt, \tag{155}$$

$$C(x, x) = \frac{\int_{\tau=0}^{\infty} \int_{x \in \mathbb{R}^N} x x^T\, h(x, t;\ \Sigma, v, \tau)\, dx\, dt}{\int_{\tau=0}^{\infty} \int_{x \in \mathbb{R}^N} h(x, t;\ \Sigma, v, \tau)\, dx\, dt} - \bar{x}\bar{x}^T = s\, \Sigma. \tag{156}$$

For the temporal derivatives of $h(x, t;\ \Sigma, v, \tau)$, we can also obtain finite moments over time, by squaring the temporal derivatives. Hence, we can measure the spatio-temporal mean of the squared velocity-adapted derivatives $h_{\bar{t}}^2(x, t;\ \Sigma, v, \tau)$ and $h_{\bar{t}\bar{t}}^2(x, t;\ \Sigma, v, \tau)$ according to

$$M(h_{\bar{t}}^2) = \frac{\int_{\tau=0}^{\infty} \int_{x \in \mathbb{R}^N} \left(\begin{array}{c} x \\ t \end{array} \right) h_{\bar{t}}^2(x, t;\ \Sigma, v, \tau)\, dx\, dt}{\int_{\tau=0}^{\infty} \int_{x \in \mathbb{R}^N} h_{\bar{t}}^2(x, t;\ \Sigma, v, \tau)\, dx\, dt} = \frac{1}{5} \left(\begin{array}{c} v \\ 1 \end{array} \right) \tau^2 \tag{157}$$

$$M(h_{\bar{t}\bar{t}}^2) = \frac{\int_{\tau=0}^{\infty} \int_{x \in \mathbb{R}^N} \left(\begin{array}{c} x \\ t \end{array} \right) h_{\bar{t}\bar{t}}^2(x, t;\ \Sigma, v, \tau)\, dx\, dt}{\int_{\tau=0}^{\infty} \int_{x \in \mathbb{R}^N} h_{\bar{t}\bar{t}}^2(x, t;\ \Sigma, v, \tau)\, dx\, dt} = \frac{1}{9} \left(\begin{array}{c} v \\ 1 \end{array} \right) \tau^2 \tag{158}$$

To summarize, a main message from the estimates in this subsection is that (i) the spatial shape of the spatio-temporal kernel $h(x, t;\ \Sigma, v, \tau)$ is described by the spatial covariance matrix Σ, (ii) the temporal extent is proportional to τ^2 and (iii) the velocity vector v specifies the orientation of the kernel in space-time.

Laplace transforms and semi-group/cascade smoothing structure over temporal scale. For a one-sided purely temporal signal $f(t)$ with $f(t) = 0$ for $t \leq 0$, the Laplace transform is defined by

$$(\mathcal{L}f)(q) = \bar{f}(q) = \int_{t=0}^{\infty} f(t)\, e^{-qt} dt. \tag{159}$$

With the one-sided and finite support convolution operation defined by

$$(f * g)(t) = \int_{u=0}^{t} f(u)\, g(t-u)\, du = \int_{u=0}^{t} f(t-u)\, g(u)\, du, \tag{160}$$

a corresponding convolution theorem holds for the Laplace transforms of f and g

$$\mathcal{L}(f * g) = (\mathcal{L}f)(\mathcal{L}g) \tag{161}$$

With regard to the time-causal kernel $\phi(t;\; \tau)$, its Laplace transform is given by

$$(\mathcal{L}\phi)(q;\; \tau) = \bar{\phi}(q;\; \tau) = \int_{t=0}^{\infty} \phi(t;\; \tau)\, e^{-qt} dt = \int_{t=0}^{\infty} \frac{1}{\sqrt{2\pi}\, t^{3/2}}\, \tau\, e^{-\tau^2/2t}\, e^{-qt} dt = e^{-\sqrt{2q}\,\tau}, \tag{162}$$

and the result of multiplying two such Laplace transforms is of the form

$$\bar{\phi}(q;\; \tau_1)\, \bar{\phi}(q;\; \tau_2) = e^{-\sqrt{2q}\,\tau_1} e^{-\sqrt{2q}\,\tau_2} = e^{-\sqrt{2q}\,(\tau_1+\tau_2)} = \bar{\phi}(q;\; \tau_1 + \tau_2) \tag{163}$$

corresponding to the linear semi-group structure of $\phi(t;\; \tau)$ under additions of the temporal scale parameter τ. In terms of Laplace transforms, we have

$$\mathcal{L}\left(\phi(\cdot;\; \tau_1) * \phi(\cdot;\; \tau_2)\right) = (\mathcal{L}\phi(\cdot;\; \tau_1))(\mathcal{L}\phi(\cdot;\; \tau_2)) = \mathcal{L}\phi(\cdot;\; \tau_1 + \tau_2) \tag{164}$$

or more explicitly in terms of one-sided and finite support convolution operations

$$\phi(\cdot;\; \tau_1) * \phi(\cdot;\; \tau_2) = \phi(\cdot;\; \tau_1 + \tau_2). \tag{165}$$

Due to this semi-group structure, the time-causal scale-space also satisfies the cascade smoothing property

$$L(\cdot;\; \tau_2) = \phi(\cdot;\; \tau_2 - \tau_1) * L(\cdot;\; \tau_1). \tag{166}$$

and so do all temporal scale-space derivatives

$$L_{t^\alpha}(\cdot;\; \tau_2) = \phi(\cdot;\; \tau_2 - \tau_1) * L_{t^\alpha}(\cdot;\; \tau_1). \tag{167}$$

Alternatively, we can also obtain the temporal scale-space derivatives by convolution with temporal derivatives of the time-causal kernel

$$L_{t^\alpha}(\cdot;\; \tau) = \phi_{t^\alpha}(\cdot;\; \tau) * f(\cdot). \tag{168}$$

Geometric covariance properties. This spatio-temporal scale-space concept is *closed* under (i) *rescalings* of the spatial and temporal dimensions, (ii) *affine transformations* in the spatial domain and (iii) *Galilean transformations* of space-time (Lindeberg 2011, section 5.1.2). Therefore, it satisfies the natural transformation properties that allow it to handle:

- image data acquired with different spatial and/or temporal *sampling rates*,
- image structures of different spatial and/or temporal *extent*,
- objects at different *distances* from the camera,
- the linear component of *perspective deformations* and
- the linear component of *relative motions* between objects in the world and the observer.

Similar covariance properties hold also for the Gaussian spatio-temporal scale-space.

8 Temporal smoothing kernels

This section gives explicit expressions for a number of kernels that can be used for modelling the temporal smoothing step in the spatio-temporal scale-space concepts described in this article.

8.1 The truncated and time-delayed Gaussian kernel

The regular (non-centered) one-dimensional time-delayed Gaussian kernel is of the form

$$g(t;\ \tau, \delta) = \frac{1}{\sqrt{2\pi\tau}} e^{-(t-\delta)^2/2\tau} \tag{169}$$

with its regular first- and second-order derivatives

$$g_t(t;\ \tau, \delta) = -\frac{(t-\delta)}{\tau}\, g(t;\ \tau, \delta) = -\frac{(t-\delta)}{\sqrt{2\pi}\, t\, \tau^{3/2}} e^{-(t-\delta)^2/2\tau} \tag{170}$$

$$g_{tt}(t;\ \tau, \delta) = \frac{((t-\delta)^2 - \tau)}{\tau^2}\, g(t;\ \tau, \delta) = \frac{((t-\delta)^2 - \tau)}{\sqrt{2\pi}\, t^2\, \tau^{5/2}} e^{-(t-\delta)^2/2\tau} \tag{171}$$

Graphs of these kernels are shown in figure 36. Notably, these kernels are not strictly time causal. To arbitrary degree of accuracy, however, they can by truncation be approximated by truncated time-causal kernels, provided that the time delay δ is chosen sufficiently long in relation to the temporal scale τ. Hence, the choice of δ leads to a trade-off between the computational accuracy of the implementation and the temporal response properties as delimited by a non-zero time delay. This problem, however, arises only for real-time analysis. For off-line computations, the time delay can in many cases be set to zero. In this respect, the truncated and time-shifted Gaussian kernels may serve as a simplest possible model for a temporal scale-space representation, provided that the requirements of temporal causality and temporal recursivity are relaxed.

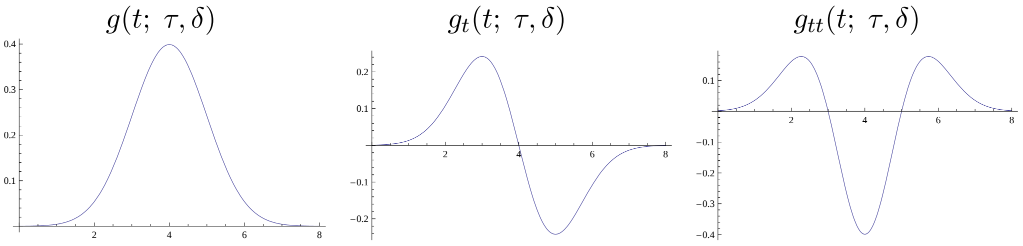


Figure 36: The time-shifted Gaussian kernel $g(t;\ \tau, \delta) = 1/\sqrt{2\pi\tau}\exp(-(t-\delta)^2/2\tau)$ for $\tau = 1$ and $\delta = 4$ with its first- and second-order temporal derivatives.

8.2 Truncated exponential kernels

When coupling a set of truncated exponential filter in cascade, the primitive time constants μ_i should as previously describe preferably be chosen such that the composed time constants τ_k are distributed according to a geometric series (129). The explicit expression for $h_{composed}^{(r)}$ will, however, therefore in general be rather complex. If we, however, for the ease of theoretical analysis consider the specific modelling case with all μ_i being equal, *i.e.*, $\mu_i = \mu$, then

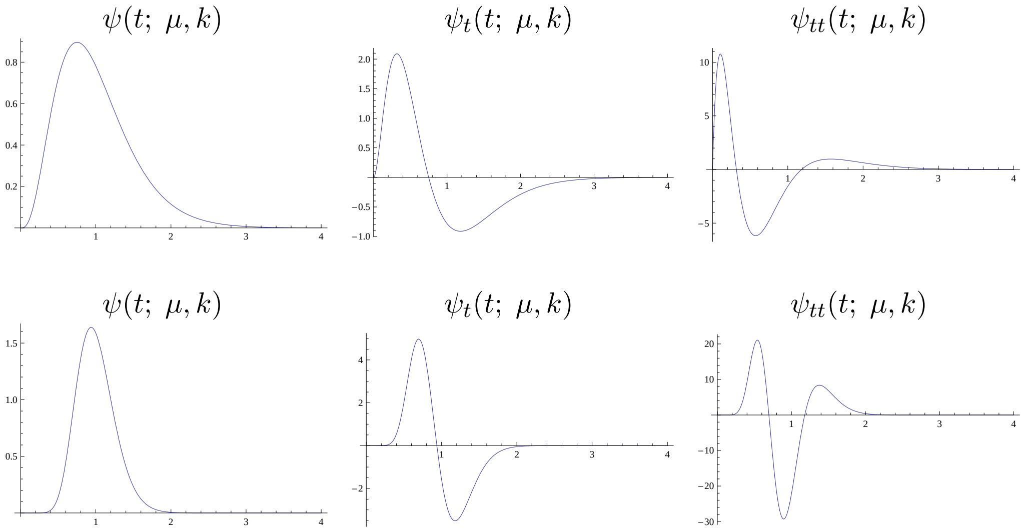


Figure 37: Equivalent kernels $h_{composed}(t;\ \mu) = *_{i=1}^{k} h_{exp}(t;\ \mu)$ corresponding to the composition of k truncated exponential kernels $h_{exp}(t;\ \mu) = \frac{1}{\mu} \exp -t/\mu$ having the same time constant μ, with their first- and second-order derivatives. (top row) $k = 4$ and $\mu = 1/4$. (bottom row) $k = 16$ and $\mu = 1/16$.

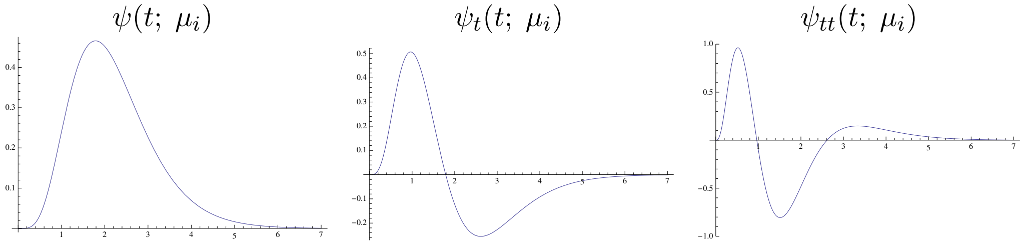


Figure 38: Equivalent kernels $h_{composed}(t;\ \mu) = *_{i=1}^{k} h_{exp}(t;\ \mu_i)$ corresponding to the composition of $k = 7$ truncated exponential kernels with different time constants defined from a self-similar distribution of the temporal scale levels according to equations (129) and (131) and corresponding to a uniform distribution in terms of effective temporal scale $\tau_{eff} = \log \tau$ with $\tau_{min} = 0.1$ and $\tau_{max} = 1$.

a closed form analysis becomes much simpler. Straightforward computation of the inverse Laplace transform of (116) shows that the equivalent convolution kernel is of the form

$$h_{composed}(t;\ \mu, k) = \mathcal{L}^{-1}\left(\frac{1}{(1+\mu q)^k}\right) = \frac{t^{k-1}\, e^{-t/\mu}}{\mu^k\, \Gamma(k)} \qquad (t > 0) \tag{172}$$

where the composed kernel has mean value $M = k\mu$ and variance $V = k\mu^2$. Note that in contrast to the primitive truncated exponentials, which are discontinuous at the origin, these kernels are continuous of order $k-1$, thus allowing for differentiation up to order $k-1$. The corresponding expressions for the first- and second-order derivatives are

$$
\begin{aligned}
h_{composed,t}(t;\ \mu, k) &= \mu^{-k-1} t^{k-2} \frac{((k-1)\mu - t)}{\Gamma(k)} e - t/\mu \\
&= -\frac{(t-(k-1)\mu)}{\mu t}\, h_{composed,t}(t;\ \mu, k) \qquad (173) \\
h_{composed,tt}(t;\ \mu, k) &= \mu^{-k-2} t^{k-3} \frac{\left((k^2-3k+2)\mu^2 - 2(k-1)t\mu + t^2\right)}{\Gamma(k)} e^{-t/\mu} \\
&= \frac{\left(\left(k^2-3k+2\right)\mu^2 - 2(k-1)t\mu + t^2\right)}{\mu^2 t^2}\, h_{composed,t}(t;\ \mu, k) \qquad (174)
\end{aligned}
$$

Figure 37 shows graphs of these kernels for two combinations of μ and k corresponding to a similar value of the mean $m = k\,\mu$. As can be seen from the graphs, the kernels are highly asymmetric for small values of k, whereas they become gradually more symmetric as the value of k increases. Figure 38 shows examples of kernels corresponding to a set of truncated exponential kernels having different time constants as defined from equations (129) and (131) and corresponding to a self-similar distribution in terms of effective scale.

8.3 The time-causal semi-group and non-enhancement kernel $\phi(t;\ \tau)$

The time-causal kernel also previously studied in the context of heat conduction in solids (Carslaw & Jaeger 1959) has the explicit expression

$$
\phi(t;\ \tau) = \frac{\tau}{\sqrt{2\pi}\, t^{3/2}} e^{-\tau^2/2t} \qquad (175)
$$

with its first- and second-order derivatives given by

$$
\begin{aligned}
\phi_t(t;\ \tau) &= -\frac{\tau(3t-\tau^2)}{2\sqrt{2\pi}\, t^{7/2}} e^{-\tau^2/2t} = -\frac{(3t-\tau^2)}{2t^2}\,\phi(t;\ \tau) \qquad (176) \\
\phi_{tt}(t;\ \tau) &= \frac{(15t^2 - 10t\tau^2 + \tau^4)}{4\sqrt{2\pi}\, \tau^{11/2}} e^{-\tau^2/2t} = \frac{(15t^2 - 10t\tau^2 + \tau^4)}{4t^4}\,\phi(t;\ \tau) \qquad (177)
\end{aligned}
$$

see figure 39 for graphs.

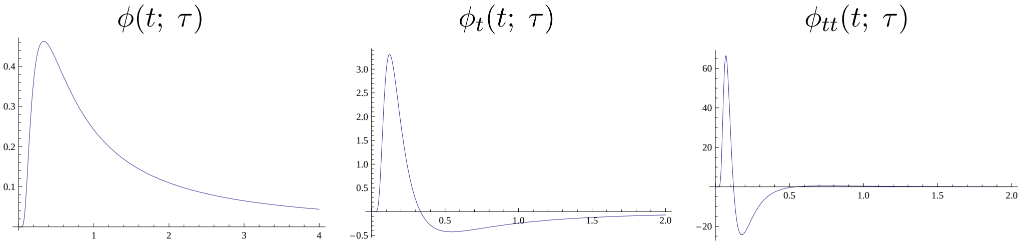


Figure 39: The time-causal kernel $\phi(t;\ \tau) = 1/\sqrt{2\pi t^3}\,\tau \exp(-\tau^2/2t)$ for $\tau = 1$ with its first- and second-order temporal derivatives.

To visualize the temporal response properties of the one-dimensional time-causal kernel $\phi(t;\ \tau)$, we can also compute the response to a step function $f_{step}(t) = H(t) = 1$ for $t > 0$ and $f_{step}(t) = H(t) = 0$ for $t < 0$

$$
L_{step}(t;\ \tau) = \operatorname{erfc}\left(\frac{\tau}{\sqrt{2t}}\right) \qquad (178)
$$

and to a linear ramp $f_{ramp}(t) = t$ (see figure 40)

$$L_{ramp}(t;\ \tau) = (t+\tau^2)\,\mathrm{erfc}\left(\frac{\tau}{\sqrt{2t}}\right) - e^{-\frac{\tau^2}{2t}}\sqrt{\frac{2}{\pi}}\tau\sqrt{t}. \tag{179}$$

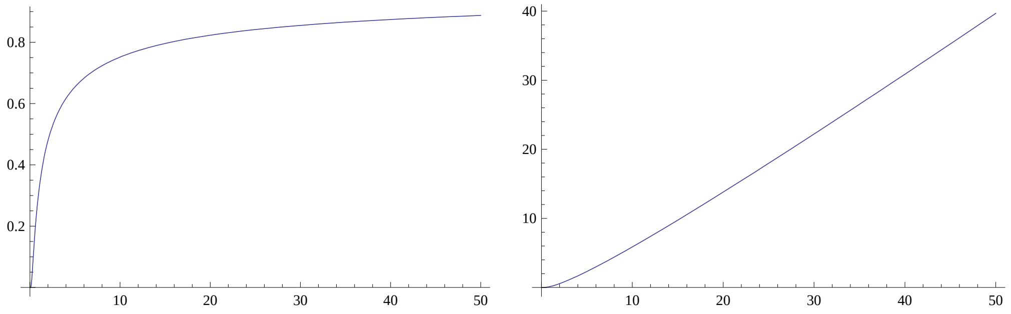


Figure 40: The response dynamics of the one-dimensional time-causal scale-space kernel $\phi(t;\ \tau)$ to (left) a unit step function and (right) a linear ramp at temporal scale $\tau = 1$.

9 History of axiomatic scale-space formulations

When (Witkin 1983) coined the term “scale-space”, he was concerned with one-dimensional signals and observed that new local extrema cannot be created under Gaussian convolutions. Specifically, he applied this property to zero-crossings of the second-order derivative to construct so-called “fingerprints”. This observation shows that Gaussian convolution satisfies certain sufficiency results for being a smoothing operation. The first proof in the Western literature of the necessity of Gaussian smoothing for generating a scale-space was given by (Koenderink 1984), who also gave a formal extension of the scale-space theory to higher dimensions. He introduced the concept of *causality*, which means that new level surfaces must not be created in the scale-space representation when the scale parameter is increased. By combining causality with the notions of *isotropy* and *homogeneity*, which essentially mean that all spatial positions and all scale levels must be treated in a similar manner, he showed that the scale-space representation must satisfy the diffusion equation

$$\partial_t L = \frac{1}{2}\nabla^2 L. \tag{180}$$

Related necessity results were given by (Babaud et al. 1986) and by (Yuille & Poggio 1986).

(Lindeberg 1990) considered the problem of characterizing those kernels in one dimension that share the property of not introducing new local extrema or new zero-crossings in a signal under convolution. Such scale-space kernels can be completely classified using classical results by (Schoenberg 1950, Schoenberg 1953). For continuous signals, it can be shown that all such non-trivial scale-space kernels can be decomposed into Gaussian kernels and truncated exponential functions. By imposing a *semi-group* structure on scale-space kernels, the Gaussian kernels will then be singled out as a unique choice. For discrete signals, the corresponding result is that all discrete scale-space kernels can be decomposed into generalized binomial smoothing, moving average or first-order recursive filtering and infinitesimal smoothing with the discrete analogue of the Gaussian kernel. To express a corresponding theory for higher-dimensional signals, (Lindeberg 1990) reformulated Koenderink's causality requirement into *non-enhancement of local extrema* and combined this requirement with a semi-group structure as well as an *infinitesimal generator* and showed that all such discrete scale-spaces must satisfy semi-discrete diffusion equations. A corresponding scale-space formulation for continuous signals based on non-enhancement of local extrema for rotationally symmetric smoothing kernels was presented in (Lindeberg 1996).

A formulation by (Florack et al. 1992) with continued work by (Pauwels et al. 1995) shows that the class of allowable scale-space kernels can also be restricted by combining a semi-group structure of convolution operations with *scale invariance* and rotational symmetry. When (Florack et al. 1992) studied this approach, they used separability in Cartesian coordinates as an additional constraint and showed that this lead to the Gaussian kernel. Separability should, however, not be counted as a scale-space axiom, since it is a coordinate dependent property related to issues of implementation. In addition, the requirement of separability in combination with rotational symmetry would *per se* fixate the smoothing kernel to be a Gaussian.[11] If the requirement about separability on the other hand is relaxed, (Pauwels et al. 1995) showed that this leads to a one-parameter family of scale-spaces, with Fourier transforms of the form

$$\hat{h}(\omega;\ s) = e^{-\alpha|\sigma\omega|^p}. \tag{181}$$

where $\sigma = \sqrt{s}$. Within this class, it can furthermore be shown that only the exponents p that are even integers lead to differential equations that have local infinitesimal generators of a classical form. Specifically, out of this countable set in turn, only the choice $p = 2$ gives rise to a non-negative convolution kernel, which leads to the Gaussian kernel.

There are, however, also possibilities of defining scale-space representations for other values of p. The specific case with $p = 1$ has been studied by (Felsberg & Sommer 2004), who showed that the corresponding scale-space representation is in the two-dimensional case given by convolution with Poisson kernels of the form

$$P(x;\ s) = \frac{s}{2\pi(\left(\frac{s}{2}\right)^2 + |x|^2)^{3/2}} \tag{182}$$

(Duits et al. 2003, Duits et al. 2004) have investigated the cases with other non-integer values of p in the range $]0, 2[$ and showed that such families of self-similar α-scale-spaces (with $\alpha = p/2$) can be modelled so-called pseudo-partial differential equations of the form

$$\partial_s L = -\frac{1}{2}(-\Delta)^{p/2} L \tag{183}$$

These scale-spaces can be related to the theory of Lévy processes and infinitely divisible distributions. For example, according to this theory a non-trivial probability measure on $\mathbb{R}^N$ is α-stable with $0 < \alpha \leq 2$ if and only if its Fourier transform is of the form (181) with $p = \alpha$ (Sato 1999, page 86). These scale-space do, however, not obey non-enhancement of local extrema.

[11] This result can be easily verified as follows: Consider for simplicity the two-dimensional case. Rotational symmetry and separability imply that h must satisfy $h(r\cos\phi, r\sin\phi) = h_1(r) = h_2(r\cos\phi)\,h_2(r\sin\phi)$ for some functions h_1 and h_2 (where (r, ϕ) are polar coordinates). Inserting $\phi = 0$ shows that $h_1(r) = h_2(r)\,h_2(0)$. With $\psi(\xi) = \log(h_2(\xi)/h_2(0))$ this relation reduces to $\psi(r\cos\phi) + \psi(r\sin\phi) = \psi(r)$. Differentiating this relation with respect to r and ϕ and combining these derivatives shows that $\psi'(r\sin\phi) = \psi'(r)\sin\phi$. Differentiation gives $1/r = \psi''(r)/\psi'(r)$ and integration $\log r = \log\psi'(r) - \log b$ for some b. Hence, $\psi'(\xi) = b\xi$ and $h_2(\xi) = a\exp(b\xi^2/2)$ for some a and b. Hence, if we would include both separability and rotational symmetry as scale-space axioms, we would not be able to derive any other kernels than the Gaussian.

For the specific family of Gaussian scale-space representations (Koenderink & van Doorn 1992) carried out a closely related study, where they showed that Gaussian derivative operators are natural operators to derive from a scale-space representation, given the assumption of scale invariance. Axiomatic derivations of image processing operators based on scale invariance have also been given in earlier Japanese literature (Weickert et al. 1999).

With regard to temporal data, the first proposal about a scale-space for temporal data was given by (Koenderink 1988) by applying Gaussian smoothing to a logarithmically transformed time axes. Such temporal smoothing filters have been considered in follow-up works by (Florack 1997) and (ter Haar Romeny et al. 2001). These approaches, however, appear to require infinite memory of the past and have so far not been developed for computational applications. To handle time-causality in a manner more suitable for real-time implementation, (Lindeberg & Fagerström 1996) expressed a strictly time-recursive space-time separable spatio-temporal scale-space model based on cascades of temporal scale-space kernels in terms of either truncated exponential functions or first-order recursive filters. These temporal scale-space models also had the attractive and memory saving property that temporal derivatives could be computed from differences between temporal channels at different scales, thus eliminating the need for complimentary time buffering. A similar computation of temporal derivatives has been used by (Fleet & Langley 1995). Early work on non-separable spatio-temporal scale-spaces with velocity adaptation was presented in (Lindeberg 1997, Lindeberg 2002) which was then developed into applications regarding recognition of activities and Galilean invariant image descriptors in (Laptev & Lindeberg 2004*c*, Laptev & Lindeberg 2004*a*, Lindeberg et al. 2004, Laptev & Lindeberg 2004*b*, Laptev et al. 2007) based on a Gaussian spatio-temporal scale-space. (Fagerström 2005, Fagerström 2007) then studied scale-invariant continuous scale-space models that allows for the construction of continuous semi-groups over the internal memory representation and in a special case lead to a diffusion formulation. An extension and combination of several of these linear spatial, affine and spatio-temporal concepts into a unified framework was recently presented in (Lindeberg 2011).

Outside the class of linear operations, there is also a large literature on non-linear scale-spaces (ter Haar Romeny 1994). In particular, the works by (Alvarez et al. 1993) and (Guichard 1998) have many structural similarities to the linear/affine/spatio-temporal scale-space formulations in terms of semi-group structure, infinitesimal generator and invariance to rescalings and affine or Galilean transformations. Non-linear scale-space that obey similar properties as non-enhancement of local extrema have been studied in particular by (Weickert 1998). With close relationship to non-enhancement of local extrema, the maximum principle has been used as a *sufficient* condition for defining linear or non-linear scale-space representations (Hummel & Moniot 1989, Alvarez et al. 1993).

10 Summary and conclusions

We have presented a generalized theory for Gaussian scale-space representation of spatial and spatio-temporal data. Starting from a general condition about non-creation of spurious image structures with increasing scale formalized in terms of non-enhancement of local extrema, we have described the semi-groups of convolution transformations that obey this requirement on different types of spatial and spatio-temporal image domains based on general theoretical necessity results in (Lindeberg 2011). The resulting theory comprises the existing continuous scale-space theory on symmetric spatial domains, with extensions to non-symmetric anisotropic spatial domains as well as both non-causal and time-causal spatio-temporal domains. Specifically, we have shown that this combination of scale-space axioms makes it possible to axiomatically derive the notions of:

- rotationally symmetric Gaussian scale-space on isotropic spatial domains,
- affine Gaussian scale-space on anisotropic spatial domains,
- Gaussian spatio-temporal scale-space on non-causal spatio-temporal domains, and
- two time-causal and time-recursive spatio-temporal scale-spaces on time-causal spatio-temporal domains.

A main message is that a much richer structure of affine as well as spatio-temporal filters can be obtained if we start from a reformulation of Koenderink's causality requirement into non-enhancement of local extrema, and then relax the requirement of spatial symmetry that was prevalent in the earliest scale-space formulations as well as most follow-up works. We have also described how a different time-causal and time-recursive temporal and spatio-temporal scale-space concept with weaker theoretical properties can be constructed by coupling a set truncated exponential filters corresponding to first-order integrators in cascade and be extended from a purely temporal to a spatio-temporal domain in a structurally similar way as two other spatio-temporal scale-space concepts are obtained from fully axiomatic derivations from a set of natural spatio-temporal scale-space axioms.

In companion works, such spatial, affine and spatio-temporal scale-spaces have been shown to be highly useful for different tasks in computer vision, by allowing the vision system to take into explicit account as well as to compensate for the following type of image transformations that arise when a vision system observes a real world:

- objects composed of different types of image structures at different scales,
- objects observed at different distances between the observer (camera) and the object,
- affine transformations arising from the first-order linearized component of the perspective mapping, and
- Galilean transformations arising because of relative motions between the observer and objects in the world.

Indeed, by considering more general covariance matrices for anisotropic handling of different dimensions and as well as spatial and/or spatio-temporal derivative operators applied to corresponding filters, a much richer family of filter shapes can be generated than from rotationally symmetric Gaussian kernels. For these Gaussian or Gaussian-related scale-spaces, all the generalized derivative filters resulting from the theory do also obey non-enhancement of local extrema as well as a transfer of the semi-group property into a cascade smoothing property. For the time-recursive scale-space based on truncated exponential filters coupled in cascade, the temporal smoothing stage guarantees non-creation of new local extrema or equivalently new zero-crossings when the temporal smoothing operation is applied to a purely temporal signal. In (Lindeberg 2011, Section 6) and (Lindeberg 2013*a*, Lindeberg 2013*b*) it is shown that the spatial and spatio-temporal derivative operations resulting from this theory give rise to receptive field profiles with high similarities to receptive fields recorded from biological vision. Indeed, from spatial and spatio-temporal derivatives of spatial or spatio-temporal scale-space kernels derived from this theory, it is possible to generate idealized receptive field models similar to *all* the basic types of receptive fields reported in the surveys of classical receptive fields in the lateral geniculate nucleus (LGN) and primary visual cortex (V1) by (DeAngelis et al. 1995, DeAngelis & Anzai 2004). In (Lindeberg 2013a, Lindeberg 2013b) it is furthermore proposed that we can *explain* the basic types of receptive fields found in the first stages of biological vision, which are tuned to different scales and orientations in space as well as different motion directions in space-time, from the requirement that the visual system should have the ability to compute invariant image representations from the image data with regard to the basic image transformations (symmetry properties) that occur relative to the environment corresponding to variations in viewing distance, viewing direction and relative motion between objects and the observer (see figure 4). If the underlying families of receptive fields would not allow for the computation of covariant image representations under basic image transformations or approximations thereof, there would be systematic errors arising from the resulting image representations corresponding to the amount of mismatch between the backprojections of the receptive fields onto the physical world (as illustrated in figure 5). This treatment does hence show that a very rich and both general and biologically plausible set of visual front-end operations can be obtained from a unified and generalized Gaussian scale-space theory that has been derived in an axiomatic way from first principles that reflect structural symmetry properties in relation to the environment.

For modelling and describing the properties of the resulting scale-space operations, we have here throughout used the corresponding spatial, spatio-chromatic or spatio-temporal receptive fields as primary objects in the theory. In a practical implementation, however, it should be noted that it may not at all be necessary to implement the corresponding receptive field operators in terms of explicit linear filters. Instead, the spatial and spatio-temporal smoothing operations can be implemented using diffusion equations, possibly in combination with corresponding temporal recurrence relations for the time-recursive scale-spaces. By varying the conductivities between neighbouring picture elements, local image features corresponding the application of equivalent receptive fields with different shapes (elongation, orientation and/or orientation) in space or space-time can thereby be computed directly by applying local derivative approximations to the scale-space smoothed image data. This also opens up interesting possibilities for adaptive smoothing schemes, where the local conductivities in the diffusion equations and/or the temporal recurrence relations are adapted to the local spatial or spatio-temporal image structure, which in addition to achieving covariance with respect to local affine or local Galilean image deformations could also be used for achieving a larger amount of local smoothing along *e.g.* edge or ridge structures than across them. Such locally adapted image operations could be of particular interest for expressing locally adapted imaging or image restoration schemes.

There are also other types of non-Gaussian scale-space theories, such as the self-similar scale-space families arising from equation (181) or its affine generalization $\hat{h}(\omega;\; s) = e^{-\alpha|B\omega|^p}$, alternatively $\hat{h}(\omega;\; s) = e^{-\alpha(\omega^T B^T B\omega)^{p/2}}$, where B is a non-singular $N \times N$ matrix. The resulting kernels will then be affine warpings of the Poisson kernels in equation (182) or the solutions of the pseudo-partial differential equation (183). In this context it should, however, be stressed that the generalized Gaussian scale-space theory presented in this article constitutes a particularly convenient class of scale-spaces with most attractive properties. For example, compared to the Poisson kernel in equation (182), the Gaussian smoothing filter decreases much faster towards infinity and faster than any polynomial, which implies a very strong regularizing property for any scale-space derivative. Compared to the α-scale-spaces, the Gaussian scale-spaces have classical infinitesimal generators, straightforward closed-form expressions in the spatial domain and obey non-enhancement of local extrema. The Gaussian scale-spaces are also maximally uncommitted in the sense that their smoothing kernels have maximum entropy and minimize the uncertainty relation.

We propose that this generalized axiomatic scale-space framework constitutes both a natural, theoretically well-founded and general basis to consider (i) when designing visual front-end operations for computer vision or image analysis systems and (ii) when modelling some of the earliest processing stages in biological vision.

References

Almansa, A. & Lindeberg, T. (2000), 'Fingerprint enhancement by shape adaptation of scale-space operators with automatic scale-selection', *IEEE Transactions on Image Processing* **9**(12), 2027–2042.

Alvarez, L., Guichard, F., Lions, P.-L. & Morel, J.-M. (1993), 'Axioms and fundamental equations of image processing', *Arch. for Rational Mechanics* **123**(3), 199–257.

Babaud, J., Witkin, A. P., Baudin, M. & Duda, R. O. (1986), 'Uniqueness of the Gaussian kernel for scale-space filtering', *IEEE Trans. Pattern Analysis and Machine Intell.* **8**(1), 26–33.

Ballester, C. & Gonzalez, M. (1998), 'Affine invariant texture segmentation and shape from texture by variational methods', *J. of Mathematical Imaging and Vision* **9**, 141–171.

Baumberg, A. (2000), Reliable feature matching across widely separated views, *in* 'Proc. CVPR', Hilton Head, SC, pp. I:1774–1781.

Bay, H., Ess, A., Tuytelaars, T. & van Gool, L. (2008), 'Speeded up robust features (SURF)', *Computer Vision and Image Understanding* **110**(3), 346–359.

Burghouts, G. J. & Geusebroek, J.-M. (2009), 'Performance evaluation of local colour invariants', *Computer Vision and Image Understanding* **113**(1), 48–62.

Carslaw, H. S. & Jaeger, J. C. (1959), *Conduction of Heat in Solids*, Clarendon Press, Oxford.

DeAngelis, G. C. & Anzai, A. (2004), A modern view of the classical receptive field: Linear and non-linear spatio-temporal processing by V1 neurons, *in* L. M. Chalupa & J. S. Werner, eds, 'The Visual Neurosciences', Vol. 1, MIT Press, pp. 704–719.

DeAngelis, G. C., Ohzawa, I. & Freeman, R. D. (1995), 'Receptive field dynamics in the central visual pathways', *Trends in Neuroscience* **18**(10), 451–457.

Duits, R., Felsberg, M., Florack, L. & Platel, B. (2003), α-scale-spaces on a bounded domain, *in* L. Griffin & M. Lillholm, eds, 'Proc. Scale-Space Methods in Computer Vision: Scale-Space'03', Vol. 2695 of *Lecture Notes in Computer Science*, Springer-Verlag, Isle of Skye, Scotland, pp. 494–510.

Duits, R., Florack, L., de Graaf, J. & ter Haar Romeny, B. (2004), 'On the axioms of scale space theory', *J. of Mathematical Imaging and Vision* **22**, 267—298.

Fagerström, D. (2005), 'Temporal scale-spaces', *Int. J. of Computer Vision* **2–3**, 97–106.

Fagerström, D. (2007), Spatio-temporal scale-spaces, *in* F. Gallari, A. Murli & N. Paragios, eds, 'Proc. 1st International Conference on Scale-Space Theories and Variational Methods in Computer Vision', Vol. 4485 of *Lecture Notes in Computer Science*, Springer, pp. 326–337.

Felsberg, M. & Sommer, G. (2004), 'The monogenic scale-space: A unifying approach to phase-based image processing in scale-space', *J. of Mathematical Imaging and Vision* **21**, 5–26.

Fleet, D. J. & Langley, K. (1995), 'Recursive filters for optical flow', *IEEE Trans. Pattern Analysis and Machine Intell.* **17**(1), 61–67.

Florack, L. M. J. (1997), *Image Structure*, Series in Mathematical Imaging and Vision, Springer.

Florack, L. M. J., ter Haar Romeny, B. M., Koenderink, J. J. & Viergever, M. A. (1992), 'Scale and the differential structure of images', *Image and Vision Computing* **10**(6), 376–388.

Florack, L., Niessen, W. & Nielsen, M. (1998), 'The intrinsic structure of optic flow incorporating measurement duality', *Int. J. of Computer Vision* **27**(3), 263–286.

Folland, G. B. & Sitaram, A. (1997), 'The uncertainty principle: A mathematical survey', *Journal of Fourier Analysis and Applications* **3**(3), 207–238.

Freeman, W. T. & Adelson, E. H. (1991), 'The design and use of steerable filters', *IEEE Trans. Pattern Analysis and Machine Intell.* **13**(9), 891–906.

Griffin, L. (1996), Critical point events in affine scale space, *in* J. Sporring, M. Nielsen, L. Florack & P. Johansen, eds, 'Gaussian Scale-Space Theory: Proc. PhD School on Scale-Space Theory', Springer, Copenhagen, Denmark, pp. 165–180.

Guichard, F. (1998), 'A morphological, affine, and Galilean invariant scale-space for movies', *IEEE Trans. Image Processing* **7**(3), 444–456.

Hall, D., de Verdiere, V. & Crowley, J. (2000), Object recognition using coloured receptive fields, *in* 'Proc. ECCV'00', Vol. 1842 of *Lecture Notes in Computer Science*, Springer-Verlag, Dublin, Ireland, pp. I:164–177.

Hille, E. & Phillips, R. S. (1957), *Functional Analysis and Semi-Groups*, Vol. XXXI, American Mathematical Society Colloquium Publications.

Hirschmann, I. I. & Widder, D. V. (1955), *The Convolution Transform*, Princeton University Press, Princeton, New Jersey.

Hubel, D. H. & Wiesel, T. N. (2005), *Brain and Visual Perception: The Story of a 25-Year Collaboration*, Oxford University Press.

Hummel, R. A. & Moniot, R. (1989), 'Reconstructions from zero-crossings in scale-space', *IEEE Trans. Acoustics, Speech and Signal Processing* **37**(12), 2111–2130.

Iijima, T. (1962), Observation theory of two-dimensional visual patterns, Technical report, Papers of Technical Group on Automata and Automatic Control, IECE, Japan.

Karlin, S. (1968), *Total Positivity*, Stanford Univ. Press.

Kläser, A., Marszalek, M. & Schmid, C. (2008), A spatio-temporal descriptor based on 3D-gradients, *in* 'Proc. British Machine Vision Conference', Leeds, U.K.

Koch, C. (1999), *Biophysics of Computation: Information Processing in Single Neurons*, Oxford University Press.

Koenderink, J. J. (1984), 'The structure of images', *Biological Cybernetics* **50**, 363–370.

Koenderink, J. J. (1988), 'Scale-time', *Biological Cybernetics* **58**, 159–162.

Koenderink, J. J., Kaeppers, A. & van Doorn, A. J. (1992), Local operations: The embodiment of geometry, *in* G. Orban & H.-H. Nagel, eds, 'Artificial and Biological Vision Systems', pp. 1–23.

Koenderink, J. J. & van Doorn, A. J. (1987), 'Representation of local geometry in the visual system', *Biological Cybernetics* **55**, 367–375.

Koenderink, J. J. & van Doorn, A. J. (1992), 'Generic neighborhood operators', *IEEE Trans. Pattern Analysis and Machine Intell.* **14**(6), 597–605.

Laptev, I., Caputo, B., Schuldt, C. & Lindeberg, T. (2007), 'Local velocity-adapted motion events for spatio-temporal recognition', *Computer Vision and Image Understanding* **108**, 207–229.

Laptev, I. & Lindeberg, T. (2003), Space-time interest points, *in* 'Proc. 9th Int. Conf. on Computer Vision', Nice, France, pp. 432–439.

Laptev, I. & Lindeberg, T. (2004*a*), Local descriptors for spatio-temporal recognition, *in* 'Proc. ECCV'04 Workshop on Spatial Coherence for Visual Motion Analysis', Vol. 3667 of *Lecture Notes in Computer Science*, Springer, Prague, Czech Republic, pp. 91–103.

Laptev, I. & Lindeberg, T. (2004*b*), Velocity adaptation of space-time interest points, *in* 'International Conference on Pattern Recognition', Vol. 2, Cambridge, pp. 1–6.

Laptev, I. & Lindeberg, T. (2004*c*), 'Velocity-adapted spatio-temporal receptive fields for direct recognition of activities', *Image and Vision Computing* **22**(2), 105–116.

Laptev, I., Marszalek, M., Schmid, C. & Rozenfeld, B. (2008), Learning realistic human actions from movies, *in* 'Proc. Computer Vision and Pattern Recognition CVPR'08', pp. 1–8.

Larsen, A. B. L., Darkner, S., Dahl, A. L. & Pedersen, K. S. (2012), Jet-based local image descriptors, *in* 'Proc. Eur. Conf. on Computer Vision ECCV'12', Vol. 7574 of *Springer LNCS*, pp. III:638–650.

Lazebnik, S., Schmid, C. & Ponce, J. (2005), 'A sparse texture representation using local affine regions', *IEEE Trans. Pattern Analysis and Machine Intell.* **27**(8), 1265–1278.

Lifshitz, L. & Pizer, S. (1990), 'A multiresolution hierarchical approach to image segmentation based on intensity extrema', *IEEE Trans. Pattern Analysis and Machine Intell.* **12**(6), 529–541.

Linde, O. & Lindeberg, T. (2004), Object recognition using composed receptive field histograms of higher dimensionality, *in* 'International Conference on Pattern Recognition', Vol. 2, Cambridge, pp. 1–6.

Linde, O. & Lindeberg, T. (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.

Lindeberg, T. (1990), 'Scale-space for discrete signals', *IEEE Trans. Pattern Analysis and Machine Intell.* **12**(3), 234–254.

Lindeberg, T. (1994*a*), 'Scale-space theory: A basic tool for analysing structures at different scales', *Journal of Applied Statistics* **21**(2), 225–270. Also available from http://www.csc.kth.se/∼tony/abstracts/Lin94-SI-abstract.html.

Lindeberg, T. (1994*b*), *Scale-Space Theory in Computer Vision*, The Springer International Series in Engineering and Computer Science, Springer.

Lindeberg, T. (1996), On the axiomatic foundations of linear scale-space, *in* J. Sporring, M. Nielsen, L. Florack & P. Johansen, eds, 'Gaussian Scale-Space Theory: Proc. PhD School on Scale-Space Theory', Springer, Copenhagen, Denmark.

Lindeberg, T. (1997), Linear spatio-temporal scale-space, *in* B. M. ter Haar Romeny, L. M. J. Florack, J. J. Koenderink & M. A. Viergever, eds, 'Scale-Space Theory in Computer Vision: Proc. First Int. Conf. Scale-Space'97', Vol. 1252 of *Lecture Notes in Computer Science*, Springer, Utrecht, The Netherlands, pp. 113–127. Extended version available as technical report ISRN KTH NA/P–01/22–SE from KTH.

Lindeberg, T. (1998), 'Feature detection with automatic scale selection', *Int. J. of Computer Vision* **30**(2), 77–116.

Lindeberg, T. (2002), Time-recursive velocity-adapted spatio-temporal scale-space filters, *in* P. Johansen, ed., 'Proc. ECCV'02', Vol. 2350 of *Lecture Notes in Computer Science*, Springer, Copenhagen, Denmark, pp. 52–67.

Lindeberg, T. (2008), Scale-space, *in* B. Wah, ed., 'Encyclopedia of Computer Science and Engineering', John Wiley and Sons, Hoboken, New Jersey, pp. 2495–2504.

Lindeberg, T. (2011), 'Generalized Gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space', *J. of Mathematical Imaging and Vision* **40**(1), 36–81.

Lindeberg, T. (2013*a*), 'A computational theory of visual receptive fields', *Biological Cybernetics* **107**(6), 589–635. doi:10.1007/s00422-013-0569-z.

Lindeberg, T. (2013*b*), 'Invariance of visual operations at the level of receptive fields', *PLOS One* **8**(7), e66990.

Lindeberg, T. (2013*c*), Scale selection, *in* 'Encyclopedia of Computer Vision', Springer. (in press).

Lindeberg, T., Akbarzadeh, A. & Laptev, I. (2004), Galilean-corrected spatio-temporal interest operators, *in* 'International Conference on Pattern Recognition', Cambridge, pp. I:57–62.

Lindeberg, T. & Fagerström, D. (1996), Scale-space with causal time direction, *in* 'Proc. ECCV'96', Vol. 1064, Springer, Cambridge, UK, pp. 229–240.

Lindeberg, T. & Gårding, J. (1994), Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D structure, *in* J.-O. Eklundh, ed., 'Proc. ECCV'94', Vol. 800 of *Lecture Notes in Computer Science*, Springer-Verlag, Stockholm, Sweden, pp. 389–400.

Lindeberg, T. & 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.

Lowe, D. (2004), 'Distinctive image features from scale-invariant keypoints', *Int. J. of Computer Vision* **60**(2), 91–110.

Mikolajczyk, K. & Schmid, C. (2004), 'Scale and affine invariant interest point detectors', *Int. J. of Computer Vision* **60**(1), 63–86.

Nagel, H. & Gehrke, A. (1998), Spatiotemporal adaptive filtering for estimation and segmentation of optical flow fields, *in* 'Proc. ECCV'98', Springer-Verlag, Freiburg, Germany, pp. 86–102.

Pauwels, E. J., Fiddelaers, P., Moons, T. & van Gool, L. J. (1995), 'An extended class of scale-invariant and recursive scale-space filters', *IEEE Trans. Pattern Analysis and Machine Intell.* **17**(7), 691–701.

Pazy, A. (1983), *Semi-groups of Linear Operators and Applications to Partical Differential Equations*, Applied Mathematical Sciences, Springer-Verlag.

Perona, P. (1992), 'Steerable-scalable kernels for edge detection and junction analysis', *Image and Vision Computing* **10**, 663–672.

Rothganger, F., Lazebnik, S., Schmid, C. & Ponce, J. (2006), '3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints', *Int. J. of Computer Vision* **66**(3), 231–259.

Sato, K.-I. (1999), *Lévy Processes and Infinitely Divisible Distributions*, Cambridge Studies in Advanced Mathematics, Cambridge University Press.

Schaffalitzky, F. & Zisserman, A. (2001), Viewpoint invariant texture matching and wide baseline stereo, *in* 'Proc. 8th Int. Conf. on Computer Vision', Vancouver, Canada, pp. II:636–643.

Schiele, B. & Crowley, J. (2000), 'Recognition without correspondence using multidimensional receptive field histograms', *Int. J. of Computer Vision* **36**(1), 31–50.

Schoenberg, I. J. (1950), 'On Pòlya frequency functions. ii. Variation-diminishing integral operators of the convolution type', *Acta Sci. Math. (Szeged)* **12**, 97–106.

Schoenberg, I. J. (1953), 'On smoothing operations and their generating functions', *Bull. Amer. Math. Soc.* **59**, 199–230.

Shao, L. & Mattivi, R. (2010), Feature detector and descriptor evaluation in human action recognition, *in* 'Proc. ACM International Conference on Image and Video Retrieval CIVR'10', Xian, China, pp. 477–484.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H. & Heeger, D. J. (1992), 'Shiftable multi-scale transforms', *IEEE Trans. Information Theory* **38**(2).

Sporring, J., Nielsen, M., Florack, L. & Johansen, P., eds (1996), *Gaussian Scale-Space Theory: Proc. PhD School on Scale-Space Theory*, Series in Mathematical Imaging and Vision, Springer, Copenhagen, Denmark.

ter Haar Romeny, B. (2003), *Front-End Vision and Multi-Scale Image Analysis*, Springer.

ter Haar Romeny, B., ed. (1994), *Geometry-Driven Diffusion in Computer Vision*, Series in Mathematical Imaging and Vision, Springer.

ter Haar Romeny, B., Florack, L. & Nielsen, M. (2001), Scale-time kernels and models, *in* 'Scale-Space and Morphology: Proc. Scale-Space'01', Lecture Notes in Computer Science, Springer, Vancouver, Canada.

Tola, E., Lepetit, V. & Fua, P. (2010), 'Daisy: An efficient dense descriptor applied to wide baseline stereo', *IEEE Trans. Pattern Analysis and Machine Intell.* **32**(5), 815–830.

van de Sande, K. E. A., Gevers, T. & Snoek, C. G. M. (2010), 'Evaluating color descriptors for object and scene recognition', *IEEE Trans. Pattern Analysis and Machine Intell.* **32**(9), 1582–1596.

Wang, H., Ullah, M. M., Kläser, A., Laptev, I. & Schmid, C. (2009), Evaluation of local spatio-temporal features for action recognition, *in* 'Proc. British Machine Vision Conference', London, U.K.

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

Weickert, J., Ishikawa, S. & Imiya, A. (1999), 'Linear scale-space has first been proposed in Japan', *J. of Mathematical Imaging and Vision* **10**(3), 237–252.

Willems, G., Tuytelaars, T. & van Gool, L. (2008), An efficient dense and scale-invariant spatio-temporal interest point detector, *in* 'Proc. ECCV'08', Vol. 5303 of *Lecture Notes in Computer Science*, Springer, Marseille, France, pp. 650–663.

Witkin, A. P. (1983), Scale-space filtering, *in* 'Proc. 8th Int. Joint Conf. Art. Intell.', Karlsruhe, Germany, pp. 1019–1022.

Young, R. A. (1987), 'The Gaussian derivative model for spatial vision: I. Retinal mechanisms', *Spatial Vision* **2**, 273–293.

Young, R. A., Lesperance, R. M. & Meyer, W. W. (2001), 'The Gaussian derivative model for spatio-temporal vision: I. Cortical model', *Spatial Vision* **14**(3, 4), 261–319.

Yuille, A. L. & Poggio, T. A. (1986), 'Scaling theorems for zero-crossings', *IEEE Trans. Pattern Analysis and Machine Intell.* **8**, 15–25.

Zhang, J., Barhomi, Y. & Serre, T. (2012), A new biologically inspired image descriptor, *in* 'Proc. Eur. Conf. on Computer Vision ECCV'12', Vol. 7576 of *Springer LNCS*, pp. III:312–324.