J Math Imaging Vis (2015) 52:3–36
DOI 10.1007/s10851-014-0541-0

Image Matching Using Generalized Scale-Space Interest Points

Tony Lindeberg

Received: 20 August 2013 / Accepted: 5 October 2014 / Published online: 24 October 2014

Abstract The performance of matching and object recognition methods based on interest points depends on both the properties of the underlying interest points and the choice of associated image descriptors. This paper demonstrates advantages of using generalized scale-space interest point detectors in this context for selecting a sparse set of points for computing image descriptors for image-based matching. For detecting interest points at any given scale, we make use of the Laplacian $\nabla^2_{norm} L$, the determinant of the Hessian $\det \mathcal{H}_{norm} L$ and four new unsigned or signed Hessian feature strength measures $\mathcal{D}_{1,norm} L$, $\tilde{\mathcal{D}}_{1,norm} L$, $\mathcal{D}_{2,norm} L$ and $\tilde{\mathcal{D}}_{2,norm} L$, which are defined by generalizing the definitions of the Harris and Shi-and-Tomasi operators from the second moment matrix to the Hessian matrix. Then, feature selection over different scales is performed either by scale selection from local extrema over scale of scale-normalized derivates or by linking features over scale into feature trajectories and computing a significance measure from an integrated measure of normalized feature strength over scale. A theoretical analysis is presented of the robustness of the differential entities underlying these interest points under image deformations, in terms of invariance properties under affine image deformations or approximations thereof. Disregarding the effect of the rotationally symmetric scale-space smoothing operation, the determinant of the Hessian $\det \mathcal{H}_{norm} L$ is a truly affine covariant differential entity and the Hessian feature strength measures $\mathcal{D}_{1,norm} L$ and $\tilde{\mathcal{D}}_{1,norm} L$ have a major contribution from the affine covariant determinant of the Hessian, implying that local extrema of these differential entities will be more robust under affine image deformations than local extrema of the Laplacian operator or the Hessian feature strength measures $\mathcal{D}_{2,norm} L$, $\tilde{\mathcal{D}}_{2,norm} L$. It is shown how these generalized scale-space interest points allow for a higher ratio of correct matches and a lower ratio of false matches compared to previously known interest point detectors within the same class. The best results are obtained using interest points computed with scale linking and with the new Hessian feature strength measures $\mathcal{D}_{1,norm} L$, $\tilde{\mathcal{D}}_{1,norm} L$ and the determinant of the Hessian $\det \mathcal{H}_{norm} L$ being the differential entities that lead to the best matching performance under perspective image transformations with significant foreshortening, and better than the more commonly used Laplacian operator, its difference-of-Gaussians approximation or the Harris–Laplace operator. We propose that these generalized scale-space interest points, when accompanied by associated local scale-invariant image descriptors, should allow for better performance of interest point based methods for image-based matching, object recognition and related visual tasks.

Keywords Feature detection · Interest point · Scale selection · Scale linking · Matching · Object recognition · Scale invariance · Affine invariance · Scale space · Computer vision

1 Introduction

A common approach to image-based matching consists of detecting interest points from the image data, computing associated local image descriptors around the interest points and then establishing a correspondence between the image descriptors (see Fig. 1 for an illustration). Specifically, the SIFT operator (Lowe [119]) and the SURF oper-

T. Lindeberg (✉)
Department of Computational Biology, School of Computer Science and Communication, KTH Royal Institute of Technology,
100 44 Stockholm, Sweden
e-mail: tony@csc.kth.se

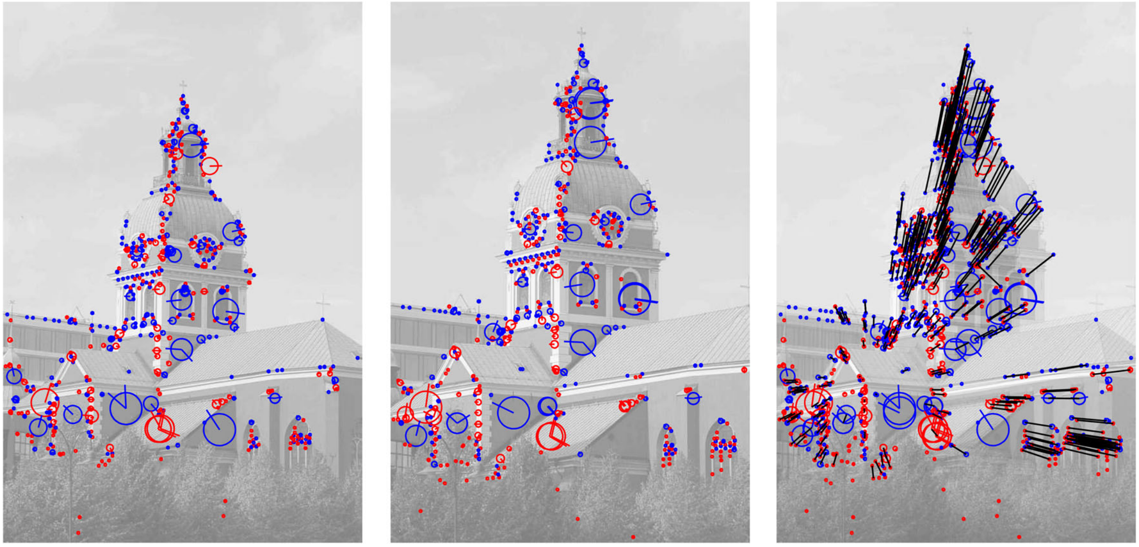


Fig. 1 Illustration of image matching using Laplacian interest points with locally adapted SIFT descriptors computed around each interest point. (*left*) and (*middle*) Two images of a building in downtown Stockholm taken from different 3-D positions with the interest points shown as *circles* overlaid on a bright copy of the original image with the size of each *circle* proportional to the locally adapted scale estimate and with the orientation estimate used for aligning the orientation of the SIFT descriptor drawn as an *angular line* from the center of the interest point. The colour of the *circle* indicates the polarity of the interest point, with *red* corresponding to *bright blobs* and *blue* corresponding to *dark blobs*. (*right*) Matching relations between the interest points drawn as *black lines* on *top* of a superposition of the original grey-level images

ator (Bay et al. [7]) have been demonstrated to be highly useful for this purpose with many successful applications, including multi-view image matching, object recognition, 3-D object and scene modelling, video tracking, gesture recognition, panorama stitching as well as robot localization and mapping. Different generalizations of the SIFT operator in terms of the image descriptor have been presented by Ke and Sukthankar [66], Mikolajczyk and Schmid [125], Burghouts and Geusebroek [24], Toews and Wells [149], van de Sande et al. [138], Tola et al. [150] and Larsen et al. [81].

In the SIFT operator, the initial detection of interest points is based on differences-of-Gaussians from which local extrema over space and scale are computed. Such points are referred to as scale-space extrema. The difference-of-Gaussians operator can be seen as an approximation of the Laplacian operator, and it follows from general results in (Lindeberg [101]) that the scale-space extrema of the scale-normalized Laplacian have scale-invariant properties that can be used for normalizing local image patches or image descriptors with respect to scaling transformations. The SURF operator is on the other hand based on initial detection of image features that can be seen as approximations of the determinant of the Hessian operator with the underlying Gaussian derivatives replaced by an approximation in terms of Haar wavelets. From the general results in (Lindeberg [101]) it follows that scale-space extrema of the determinant of the Hessian do also lead to scale-invariant behaviour, which can be used for explaining the good performance of the SIFT and SURF operators under scaling transformations.

The subject of this article is to show how the performance of image matching can be improved by using a generalized framework for detecting interest points from scale-space features involving new Hessian feature strength measures at a fixed scale and linking of image features over scale into feature trajectories. By replacing the interest points in regular SIFT or SURF by generalized scale-space interest points to be described below, it is possible to define new scale-invariant image descriptors that lead to better matching performance compared to the interest point detection mechanisms used in the regular SIFT or SURF operators.

1.1 Outline of the Presentation and Main Contributions

The paper is organized as follows: Sect. 2 gives an overview of previous work in this area, and Sect. 3 summarizes basic concepts regarding linear (Gaussian) scale-space representation that we build upon. Section 4 describes how interest point detectors can be defined at a fixed scale, by combining Gaussian derivatives computed from a scale-space representation into linear or non-linear differential invariants at every image point followed by local extrema detection. These interest point detectors comprise the previously known Laplacian and determinant of the Hessian interest point detectors and four new Hessian feature strength measures denoted $\mathcal{D}_1 L$, $\tilde{\mathcal{D}}_1 L$, $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$, which are conceptually related to the previously known Harris and Shi-and-Tomasi operators while being defined from the Hessian matrix instead of the second-moment matrix.

Section 5 presents a theoretical analysis of the robustness properties of these interest point detectors under image deformations, by analyzing the covariance properties of the underlying differential entities under affine transformations of the image domain. It is shown that the determinant of the Hessian operator is affine covariant and that the Hessian feature strength measures $\mathcal{D}_1 L$ and $\tilde{\mathcal{D}}_1 L$ have a major contribution that it is affine covariant.

Section 6 outlines how these interest point detectors can be complemented by thresholding operations, including (i) magnitude thresholding with formal relationships between thresholding values for different interest point detectors and (ii) complementary thresholding based on the sign of a complementary differential expression to increase the selective properties of the interest point detector.

Section 7 describes how these interest point detectors can be complemented by scale selection mechanisms, including the previously established scale selection methodology based on local extrema over scale of scale-normalized derivatives and a new methodology where image features at adjacent scales are linked into feature trajectories over scale. For the latter, a new measure of the significance or saliency of an interest point is defined as an integral of the scale-normalized feature strength measure along each feature trajectory. It is argued that such an integrated measure over scale may give a more robust ranking of image features by including their life length over scales and thus their stability in scale space into the significance measure.

Section 8 describes how the resulting enriched family of generalized scale-space interest point detectors can be complemented with local image descriptors, leading to a generalized family of locally adapted and scale-invariant image descriptors. Specifically, we define Gauss-SIFT and Gauss-SURF descriptors in ways analogous to the original SIFT and SURF descriptors, however, with the interest points detectors replaced by our generalized interest point detectors and with the image measurements used for computing the image descriptors defined in terms of Gaussian derivatives instead of a pyramid as done in original SIFT or Haar wavelets as used in original SURF.

Section 9 evaluates the performance of the resulting generalized interest points with their associated Gaussian image descriptors with regard to image matching. It is shown that scale linking may lead to a better selection of image features compared to scale-space extrema detection, and that the new Hessian feature strength measures $\mathcal{D}_1 L$, $\tilde{\mathcal{D}}_1 L$ and the determinant of the Hessian perform better than the Laplacian operator, its difference-of-Gaussians approximation or the Harris–Laplace operator. Section 10 shows how the approach can be extended to illumination invariance. Finally, Sects. 11 and 12 conclude with a summary and discussion.

2 Related Work

In early work, Marr and his collaborator Hildreth [121,122] proposed an early primal sketch representation of image data in terms of edges, bars, blobs and terminations defined from zero-crossings of the Laplacian as the primary type of image feature. Such features or their approximation in terms of zero-crossings of difference-of-Gaussians, however, suffer from inherent problems. If used for edge detection, they may give rise to “false edges” and for curved edges or corners they give rise to a substantial localisation error.

Today, we have access to a much more developed theory for early visual operations, which allows us to formulate a much richer and also more well-defined vocabulary of local image features. A major cornerstone for the development of a well-founded operational theory for detecting robust image features from real-world image data was provided by the framework of representing image data at multiple scales using scale-space representation, as originally proposed by Witkin [157] and Koenderink [69]. Koenderink also proposed to link image features over scales using iso-intensity linking, and this idea was picked up by Lifshitz and Pizer [86] and Gauch and Pizer [48], who developed early systems for coarse-to-fine segmentation of medical images.

A major problem when linking image features over scales based on iso-intensity linking, however, is that the intensity values of local image features are strongly affected by the Gaussian smoothing operation. To avoid such problems, Lindeberg [90] considered the linking of local extrema and saddle points over scales and defined the associated notions of grey-level blobs at any single scale and scale-space blobs over scales. The life length and extent of these structures in scale space were measured, resulting in a representation called the scale-space primal sketch, and the significance of such image structures by the 4-D volume that these linked objects occupy in scale space. Experimentally, it was shown that the resulting scale-space primal sketch allowed for extraction of salient blob-like image structures as well as scale levels for processing these in a purely bottom-up manner.

Closely related notions of linking of image structures over scales for watersheds of the gradient magnitude were used by Olsen [129] for medical image segmentation. Medical applications of the scale-space primal sketch have been developed for analyzing functional brain activation images (Lindeberg et al. [117], Coulon et al. [29], Rosbacke et al. [136], Mangin et al. [120]) and for capturing the folding patterns of the cortical surface (Cachia et al. [25]). More algorithmically based work on building graphs of blob and ridge features at different scales was presented by Crowley and his co-workers [31,32] using difference of low-pass features defined from a pyramid; hence with very close similarities to differences-of-Gaussians operators and thus the Laplacian.

Within the area of local feature detection from image data, both Harris [55] and Förstner and Gülch [45] proposed corner detectors defined from the second-moment matrix. Early applications of the trace or the determinant of the Hessian operators were presented by Beaudet [9] and blob features defined from the Laplacian responses were used as primitives for texture analysis by Voorhees and Poggio [153] and Blostein and Ahuja [15,16]. Corner detectors based on the curvature of level curves with different variations were studied by Kitchen and Rosenfeld [68], Dreschler and Nagel [40], Koenderink and Richards [70], Noble [128], Deriche and Giraudon [39], Blom [14], Brunnström et al. [23] and Lindeberg [91]. In many cases, these operators were combined with a Gaussian smoothing step, sometimes motivated by the need for decreasing the influence of noise. Today, we would refer to these operators as *single scale* feature detectors. The experimental results, however, often revealed a substantial lack of robustness, due to the need for manually choosing the scale levels and the lack of a built-in scale selection mechanism.

The general idea of performing scale selection and detecting image features by computing local maxima with respect to space and scale of γ-normalized derivatives, which leads to theoretically provable scale invariance, was initiated in Lindeberg [93,95] and then refined in Lindeberg [100,101]. Specifically, scale-invariant blob detectors were proposed from scale-space extrema of the Laplacian or the determinant of the Hessian and a scale-invariant corner detector from the rescaled level curve curvature. This approach was applied to scale-invariant feature tracking (Bretzner and Lindeberg [21]), local pattern classification (Wiltschi et al. [156]), image feature extraction for geon-based object recognition (Lindeberg and Li [116]), fingerprint analysis (Almansa and Lindeberg [3]) and real-time gesture recognition (Bretzner et al. [19,20]). Tutorial overviews of parts of the underlying scale-space framework can be found in Lindeberg [94,98,102,103,107,111].

Chomat et al. [28] and Hall et al. [54] made use of scale selection from local maxima over scales of normalized derivatives for computing scale-invariant Gaussian derivative descriptors for object recognition. Lowe [118,119] developed an object recognition system based on local position dependent histograms computed at positions and scales determined from scale-space extrema of differences of Gaussians, thus with very close similarities to scale-invariant blob detection from scale-space extrema of the Laplacian. Closely related object recognition approaches, although with different image descriptors, have been presented by Lazebnik et al. [82] and Ke and Sukthankar [66]. Bay et al. [7,8] developed an alternative approach with image features that instead can be seen as approximations to determinant-of-Hessian features expressed in terms of Haar wavelets. Opelt et al. [130] presented an object recognition approach that combines different types of interest points, specifically differences-of-Gaussians features and Harris points. Kokkinos et al. [74] made use of a related approach based on primal sketch features in terms of scale-invariant edge and ridge features. Kokkinos and Yuille [75] proposed an alternative way of computing scale invariant image descriptors, by performing explicit search in a log-polar domain based on the foveal scale-space model in (Lindeberg and Florack [113]).

A real-time system for gesture recognition based on a combination of scale-invariant Laplacian blobs and scale-invariant ridge features was presented in Bretzner et al. [19,20] based on a method for simultaneous tracking and recognition using scale-invariant features (Laptev and Lindeberg [79]). The underlying theory for real-time scale selection based on a hybrid pyramid representation was then reported in Lindeberg and Bretzner [112]. Parallel developments of real-time implementations of scale-selection have been performed by Crowley and Riff [30] and by Lowe [119].

Due to the scale invariant nature of the scale selection step, all these visual modules become scale invariant, which makes it possible for them to automatically adapt to and handle image structures of different size. Specifically, scale selection based on local extrema over scales of scale-normalized derivatives constitutes the theoretical foundation for scale-invariant object recognition based on SIFT or SURF.

The methodology for scale-invariant ridge detection based on maximisation of γ-normalized measures of ridge strength (Lindeberg [97,100]) was extended to three-dimensional images by Sato et al. [139], Frangi et al. [46] and Krissian et al. [76]; see also Kirbas and Quek [67] for a review of vessel extraction techniques. Closely related works on multi-scale ridge detection have presented by Pizer and his co-workers [134] leading to their notion of M-reps (Pizer et al. [133]).

In Lindeberg and Gårding [114], Gårding and Lindeberg [47] scale invariant blob detection by scale-space extrema was combined with subsequent computation of scale-adaptive second moment matrices to provide image features for deriving cues to local surface shape by shape-from-texture and shape-from-disparity gradients. In Lindeberg and Gårding [115] the notion of affine shape adaptation was proposed and was demonstrated to improve the accuracy of local surface orientation estimates by computing them at affine invariant fixed points in affine scale space (Lindeberg [95, chapter 15]). Baumberg [6] combined the notion of affine shape adaptation with Harris interest points for wide baseline stereo matching. Mikolajczyk and Schmid [124] furthered this notion by developing a computationally more efficient algorithm and evaluating the performance more extensively (Mikolajczyk et al. [126]). Tuytelaars and van Gool [151] showed how affine-adapted features can be used for matching of widely separated views. Lazebnik et al. [83] used affine invariant features for recognizing textured patterns. Rothganger et al. [137] used affine invariant patches from multiple views of an object for building three-dimensional object models. Other combinations of second-moment based texture descriptors with scale selection for object segmentation were used by Belongie et al. [10] and Carson et al. [27].

With these publications, there has been an increasing interest in scale invariant and affine invariant features as reflected in the surveys by Pinz [132], Lew et al. [85], Tuytelaars and Mikolajczyk [152] and Daniilidis and Eklundh [36]. In particular, so-called "bag-of-features" models have become very popular, where the computation of local image descriptors is initiated by either scale invariant or affine invariant interest points (Sivic et al. [145], Nowak et al. [33], Jiang et al. [59]); see also Mikolajczyk et al. [126] for an experimental evaluation of image descriptors at interest points, Moreels and Perona [127] and Aanaes et al. [1] for evaluations of feature detectors and image descriptors on three-dimensional datasets and Kaneva et al. [65] for evaluations using photorealistic virtual worlds. For all of these recognition approaches, the invariance properties of the recognition system rely heavily on the invariance properties of the interest points at which local image descriptors are computed. There are also other approaches to interest point detection not within the Gaussian derivative framework (Kadir and Brady [63,64], Matas et al. [123]) with applications to image based recognition by Fergus et al. [42].

During recent years there has been a growing interest in defining graph-like representations of image features (Shokoufandeh et al. [143,144]; Bretzner and Lindeberg [22], Demirci et al. [38]). Inspired by early theoretical studies by Johansen [60,61] regarding the information content in so-called "top points" in scale space where bifurcations occur, Platel et al. [135], Balmashnova et al. [5], Balmashnova and Florack [4] and Demirci et al. [37] proposed to use such bifurcation events as primitives in graph representations for image matching. Such bifurcations events were also registered in the original scale-space primal sketch concept for intensity data (Lindeberg [90]), in which the bifurcation events delimited the extent of grey-level blobs in the scale direction and provided explicit relations of how neighbouring image features (local extrema with extent) were related across scales. With the generalized notion of a scale-space primal sketch for differential descriptors used here, we obtain a straightforward and general way to compute a richer family of corresponding bifurcation events for any sufficiently well-behaved differential expression $\mathcal{D}L$. More recently, Gu et al. [52] proposed a representation for image matching based on local spatial neighbourhood relations, referred to as critical nets, that possess local stability properties over scale, with close similarities to these ideas.

With regard to the area of image matching and object recognition, Swain and Ballard [148] initiated a direction of research on histogram-based recognition methods by showing how reasonable performance of an object recognition scheme could be obtained by comparing RGB colour histograms. Schiele and Crowley [140] generalized this idea to histograms of receptive fields (Koenderink and van Doorn [72,73]) and computed histograms of either first-order Gaussian derivative operators or the gradient magnitude and the Laplacian operator at three scales, leading to 6-D histograms. Schneiderman and Kanade [141] showed that efficient recognition of faces and cars could be performed from histograms of wavelet coefficients. Linde and Lindeberg [87,88] presented a set of composed histogram descriptors of higher dimensionality that lead to better recognition performance compared to previously used receptive field histograms.

Lowe [119] combined the ideas of feature based and histogram based image descriptors, and defined a scale invariant feature transform, SIFT, which integrates the accumulation of statistics of gradient directions in local neighbourhoods of scale adapted interest points with summarizing information about the spatial layout. Bay et al. [7] presented an alternative approach with SURF features that are instead expressed in terms of Haar wavelets. Dalal and Triggs [34] extended the local SIFT descriptor to the accumulation of regional histograms of gradient directions (HOG) over larger support regions. Other closely related probabilistic methods have been presented by Fergus et al. [41], Lazebnik et al. [82] and Ke and Suktankar [66]. An evaluation and comparison of several spatial recognition methods has been presented by Mikolajczyk and Schmid [125]. Dense local approaches have been investigated by Jurie and Triggs [62], Lazebnik et al. [84], Bosch et al. [18], Agarwal and Triggs [2] and Tola et al. [150]. More recently, Larsen et al. [81] made use of multi-local N-jet descriptors that do not rely on a spatial statistics of receptive field responses as used in the SIFT and SURF descriptors or their analogues. A notable observation from experimental results is that very good performance can be obtained with coarsely quantized even binary image descriptors (Pietikäinen et al. [131], Linde and Lindeberg [88], Calonder et al. [26]). Moreover, Zhang et al. [158] have demonstrated what can be gained in computer vision by considering biologically inspired image descriptors.

View-based methods for image matching and object recognition have been extended to colour images by several authors. Slater and Healey [146] presented histogram-like descriptors that combine spatial moments with colour information. Gevers and Smeulders [50] investigated the sensitivity of different zero-order colour spaces for histogram-based recognition. Geusebroek et al. [49] proposed a set of differential colour invariants that are invariant to illumination based on a reflectance model and the Gaussian colour model proposed by Koenderink. Hall et al. [54] computed partial derivatives of colour-opponent channels, leading to an N-jet representation up to order one. Linde and Lindeberg [87,88] extended this idea by showing that highly discriminative image descriptors for object recognition can be obtained from histograms of spatio-chromatic differential invariants up to order two defined from colour-opponent channels. Burghouts and Geusebroek [24] showed that the performance of the SIFT descriptor can be improved by complementing it with a set of colour invariants. More recently, van de Sande et al. [138] have presented an evaluation of different colour-based image descriptors for recognition.

A general theoretical framework for how local receptive field responses, as used in the SIFT and SURF descriptors and their extensions or analogues to colour images and spatio-temporal image data, can constitute the basis for computing inherent properties of objects to support invariant recognition under natural image transformations is presented in (Lindeberg [106,109]) including relations to receptive fields in biological vision.

3 Scale-Space Representation

The context we consider is that we for any two-dimensional image $f\colon \mathbb{R}^2 \to \mathbb{R}$ define a Gaussian scale-space representation $L\colon \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$ according to Iijima [57], Witkin [157], Koenderink [69], Koenderink and van Doorn [72,73], Lindeberg [94,95,103,105,107], Sporring et al. [147], Florack [43], ter Haar Romeny [53]:

$$L(x, y;\ t) = \int_{(u,v)\in\mathbb{R}^2} f(x-u, y-v)\, g(u, v;\ t)\, du\, dv \tag{1}$$

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

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

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

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

with initial condition $L(\cdot, \cdot;\ 0) = f$. From this representation, *Gaussian derivatives* are defined by

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

where α and $\beta \in \mathbb{Z}_+$. From such a scale-space representation, we can at any level of scale compute different types of features, typically by combining the Gaussian derivatives into different types of (linear or non-linear) *differential invariants* (preferably rotationally invariant).

When comparing derivative responses at different scales, it is natural to introduce the notion of *γ-normalized derivatives* according to

$$\partial_\xi = t^{\gamma/2}\partial_x \qquad \partial_\eta = t^{\gamma/2}\partial_y \tag{5}$$

where $\gamma \in [0, 1]$ is a free parameter that may be set from specific context information for a particular feature detector (Lindeberg [100,101]). This type of scale normalization makes it possible to define local derivatives with respect to the current level of scale and allows us to compensate for an otherwise general overall decrease in the magnitude of regular (unnormalized) Gaussian derivatives over scale.

4 Differential Entities for Detecting Interest Points

Basic requirements on the interest points on which image matching is to be performed are that they should [110]:

(i) have a clear, preferably mathematically well-founded, *definition*,
(ii) have a well-defined *position* in image space,
(iii) have local image structures around the interest point that are *rich in information content* such that the interest points carry important information to later stages,
(iv) be stable under local and global deformations of the image domain, including perspective image deformations and illumination variations such that the interest points can be reliably computed with a high degree of *repeatability* (Mikolajczyk et al. [126]) and
(v) be sufficiently *distinct*, such that interest points corresponding to physically different points can be kept separate (Lowe [119]).

Preferably, the interest points should also have an attribute of *scale*, to make it possible to compute reliable interest points from real-world image data, including scale variations in the image domain. Specifically, the interest points should preferably be *scale-invariant* to make it possible to match corresponding image patches under scale variations, e.g., corresponding to objects of different size in the world or objects seen from different distances between the camera and the object.

In this section, we shall describe a set of differential entities that can be used for defining interest points at a fixed scale, including four previously known operators and four new ones. The Laplacian operator and the determinant of the Hessian operators have been previously used in the literature, where we here also emphasize how the polarities of these differential entities allow for a finer classification of the type of interest points, including saddle-like interest points detected by the determinant of the Hessian operator. By generalizing the constructions by which the Harris and the Shi-and-Tomasi operators have been previously defined from the second-moment matrix (structure tensor), we will also introduce a set of four new differential entities defined from the Hessian matrix, termed the Hessian feature strength measures $\mathcal{D}_{1,norm}L$, $\tilde{\mathcal{D}}_{1,norm}L$, $\mathcal{D}_{2,norm}L$ and $\tilde{\mathcal{D}}_{2,norm}L$.

4.1 The Laplacian Operator

Among the class of differential detectors that can be defined from combinations of Gaussian derivative operators, the *Laplacian operator*

$$\nabla^2 L = L_{xx} + L_{yy} = \lambda_1 + \lambda_2 \tag{6}$$

is the presumably simplest choice and corresponds to the sum of the eigenvalues λ_1 and λ_2 of the Hessian matrix.

Specifically, with regard to feature detection, we may regard a spatial extremum of $\nabla^2 L$ as a blob response, where

$$\begin{aligned} &\nabla^2 L > 0 \text{ (holds for positive definite } \mathcal{H}L) \Rightarrow \text{ dark blob} \\ &\nabla^2 L < 0 \text{ (holds for negative definite } \mathcal{H}L) \Rightarrow \text{ bright blob} \end{aligned} \tag{7}$$

Figure 2 shows an example of applying this operator to an image at a given scale as well as a number of other differential operators to be presented next.

4.2 The Determinant of the Hessian

Within the degrees of freedom available from the second-order structure of a two-dimensional image, we can obtain two functionally independent differential descriptors that are invariant to rotations in the image domain. If we choose the Laplacian as one of these operators, the *determinant of the Hessian* is a natural complement, corresponding to the product of the eigenvalues λ_1 and λ_2 of the Hessian matrix:

$$\det \mathcal{H}L = L_{xx}L_{yy} - L_{xy}^2 = \lambda_1\lambda_2. \tag{8}$$

In a similar way as for the Laplacian, we can regard local maxima and minima of the determinant of the Hessian as natural indicators of blobs. At image points where the determinant of the Hessian is positive, the Hessian matrix will be either positive or negative definite, depending on the sign of the Laplacian. At points where the determinant of the Hessian is negative, we have that the Hessian matrix is indefinite and it is natural to refer to such points as *saddle-like interest points*. To summarize, it is therefore natural to classify local maxima and minima of $\det \mathcal{H}L$ as follows:

$$\begin{aligned} &\det \mathcal{H}L > 0 \text{ and } \mathcal{H}L \text{ positive definite} \Rightarrow \text{dark blob} \\ &\det \mathcal{H}L > 0 \text{ and } \mathcal{H}L \text{ negative definite} \Rightarrow \text{bright blob} \\ &\det \mathcal{H}L < 0 \qquad \Rightarrow \text{saddle-like response} \end{aligned} \tag{9}$$

Compared to the Laplacian operator, the determinant of the Hessian will only respond if the local image pattern contains significant variations along any two ortogonal directions. Therefore, this operator implies a more restrictive condition and is in this sense a better candidate for detecting interest points compared to the Laplacian.

By comparing the results of applying the Laplacian and the determinant of the Hessian operators to the image in Fig. 2, we can first note that whereas the Laplacian operator gives a large number of responses to the oblique elongated ridge structure in the lower part of the image as well as a rather large number of responses outside the edges of the match boxes and the horse, the determinant of the Hessian does not give any responses to such one-dimensional structures. Hence, the determinant of the Hessian is more selective to corners than the Laplacian, in a agreement with the theoretical prediction. By detailed inspection of the responses at corner like structures, such as at the top of the horse or the intervening space between the legs of the horse, when one leg occludes the other, we can also see that the determinant of the Hessian operator leads to responses with better localization at corners compared to the Laplacian.

4.3 The Harris and Shi-and-Tomasi Measures

The Harris, Förstner and Shi-and-Tomasi operators are all defined from the *second-moment matrix*, or *structure tensor*

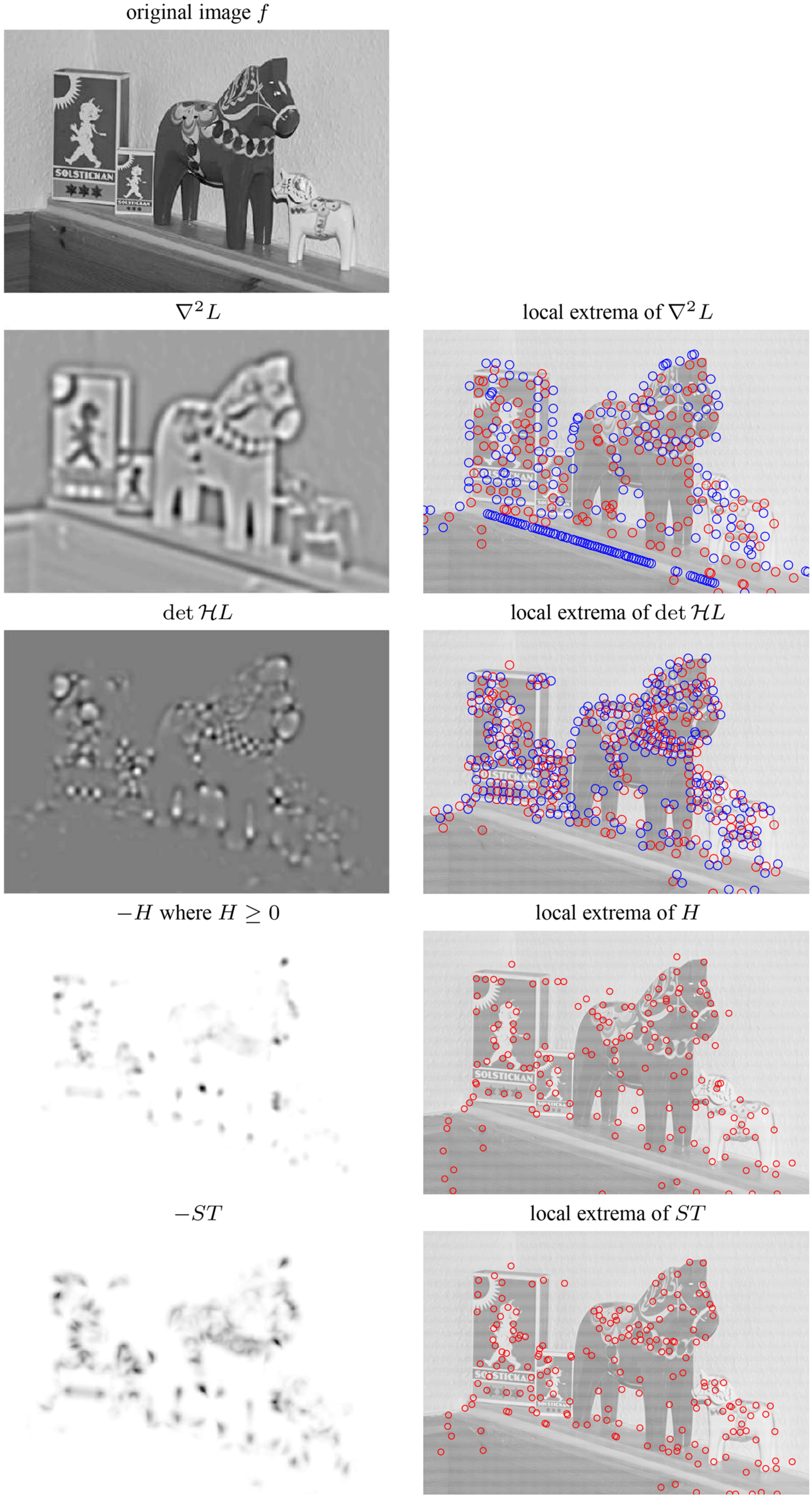
$$\mu(x, y;\ t, s) = \int_{(u,v)\in\mathbb{R}^2} \begin{pmatrix} L_x^2 & L_x L_y \\ L_x L_y & L_y^2 \end{pmatrix} g(x-u, y-v;\ s)\, du\, dv, \tag{10}$$

where the partial derivatives L_x and L_y are computed at local scale t and evaluated at position (u, v), whereas s is an integration scale parameter that can be coupled to the local scale t according to $s = r^2 t$ with $r \geq 1$. In the experiments in this paper, we use $r = 1$ motivated by that it gave the best repeatability properties of the interest points under affine image transformations among $r \in \{1, \sqrt{2}, 2\}$.

The *Harris corner detector* [55] implies that corners are detected from positive spatial maxima of the entity

$$H = \det \mu - k \operatorname{trace}^2 \mu \tag{11}$$

where k is a constant required to be in the interval $k \in]0, 0.25[$ and usually set to $k \approx 0.04$. This operator responds only if the eigenvalues of the second-moment matrix are sufficiently similar and thus only if the local image pattern contains variations along two orthogonal directions. This means that responses will be obtained at corners or near the centers of blob-like structures, whereas responses along one-dimensional edge structures will be suppressed.



original image f

$\nabla^2 L$

local extrema of $\nabla^2 L$

$\det \mathcal{H}L$

local extrema of $\det \mathcal{H}L$

$-H$ where $H \geq 0$

local extrema of H

$-ST$

local extrema of ST

Fig. 2 Differential interest point detectors at a fixed scale defined from Laplacian $\nabla^2 L$, determinant of the Hessian $\det \mathcal{H}L$, Harris H and Shi-and-Tomasi ST responses at scale $t = 32$ with corresponding features obtained by detecting local extrema at a fixed scale, with thresholding on the magnitude of the response with $C_{\nabla^2 L} = 10$, $C_{\det \mathcal{H}L} = 10^2/4 = 25$, $C_H = 10^4/4096 \approx 2.44$ and $C_{ST} = 10^2/64 \approx 1.56$. Note that the Laplacian operator responds to one-dimensional structures, whereas the other operators do not. (Image size: 512 × 350 pixels. *Red circles* denote local maxima of the operator response, while *blue circles* represent local minima.)

Shi and Tomasi [142] proposed to instead use local maxima of the minimum eigenvalue ν_1 of μ as image features

$$ST = \min(\nu_1, \nu_2) = \nu_1$$
$$= \frac{1}{2}\left(\mu_{11} + \mu_{22} - \sqrt{(\mu_{11} - \mu_{22})^2 + 4\mu_{12}^2}\right). \qquad (12)$$

Figure 2 shows an example of computing Harris corners and Shi-and-Tomasi corners at a fixed scale.

Förstner and Gülch [45] have defined other closely related measures of feature strength from the second-moment matrix. Bigün [12] has used a complex-valued generalized structure tensor for detecting different types of local symmetries in image data; see also Bigün and Granlund [13], Jähne et al. [58], Lindeberg [95], Granlund and Knutsson [51], Gårding and Lindeberg [47] and Weickert [154] for other applications of the second moment matrix/structure tensor for computing local features from image data.

4.4 Similarities Between the Hessian Matrix and the Second-Moment Matrix

Under an affine transformation

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{where} \quad A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \qquad (13)$$

the Hessian matrix $\mathcal{H}f$ transforms according to

$$(\mathcal{H}f')(x', y') = A^{-T}\,(\mathcal{H}f)(x, y)\,A^{-1} \qquad (14)$$

and provided that the notion of window function in (10) is properly defined, the second-moment matrix transforms according to Lindeberg [95], Lindeberg and Gårding [115]

$$\mu' = A^{-T}\,\mu\,A^{-1}. \qquad (15)$$

Moreover, if the Hessian matrix $\mathcal{H}L$ at a point (x_0, y_0) is either positive or negative definite, then it defines an either positive or negative definite quadratic form

$$Q_{\mathcal{H}L}(x, y) = \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}^T \begin{pmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{pmatrix}^{-1} \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} \qquad (16)$$

in a similar way as the second-moment matrix μ computed at (x_0, y_0) does

$$Q_{\mu}(x, y) = \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}^T \begin{pmatrix} \mu_{11} & \mu_{12} \\ \mu_{12} & \mu_{22} \end{pmatrix}^{-1} \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix}. \qquad (17)$$

From these two analogies, we can conclude that provided the Hessian matrix is either positive or negative definite, these two types of descriptors should have strong qualitative similarities. Förstner [44] has also shown that the second-moment matrix corresponds to the Hessian matrix of the autocorrelation function. That relation can be directly understood in the Fourier domain, since with vector notation $\omega = (\omega_x, \omega_y)^T$ both the second-moment matrix and the Hessian matrix of the autocorrelation function have a Fourier transform of the form $\omega\omega^T|\hat{L}(\omega)|^2$.

4.5 New Feature Strength Measures From the Hessian Matrix

Inspired by above mentioned similarities between the Hessian matrix and the second-moment matrix and the previous definitions of the Harris and Shi-and-Tomasi cornerness measures, we will in this section define four new interest point operators from the Hessian matrix.

Let us initially consider the following differential entity as a measure of feature strength of the Hessian matrix:

$$\mathcal{D}_1 L = \det \mathcal{H}L - k\ \text{trace}^2\,\mathcal{H}L$$
$$= L_{xx}L_{yy} - L_{xy}^2 - k\,(L_{xx} + L_{yy})^2$$
$$= \lambda_1\lambda_2 - k\,(\lambda_1 + \lambda_2)^2 \qquad (18)$$

where $k \in\,]0, 0.25[$ and λ_1 and λ_2 denote the eigenvalues of $\mathcal{H}L$. Let us then define an interest point operator by detecting *positive local maxima* of this differential entity.

To analyse the properties of this operator, let us first observe that if the Hessian matrix is indefinite, then $\det \mathcal{H}L < 0$ and it follows by necessity that $\mathcal{D}_1 L < 0$ and such points will not be detected. Hence, this operator cannot respond to saddle-like features and will only generate responses if $\mathcal{H}L$ is either positive or negative definite. Without loss of generality, let us henceforth assume that $\mathcal{H}L$ is positive definite (if not, we just change the polarity of the image and replace L by $-L$). Then, $\mathcal{D}_1 L$ will only respond if

$$\lambda_1\lambda_2 - k(\lambda_1 + \lambda_2)^2 > 0. \qquad (19)$$

Let us next assume that the eigenvalues are ordered such that $0 < \lambda_1 \leq \lambda_2$. Then, we can divide Eq. (18) by $\lambda_2^2 \neq 0$ to conclude that the operator $\mathcal{D}_1 L$ will only respond by positive maxima if

$$\frac{\lambda_1}{\lambda_2} - k\left(1 + \frac{\lambda_1}{\lambda_2}\right)^2 > 0. \qquad (20)$$

According to the assumptions, the ratio $\lambda_1/\lambda_2 \in [0, 1]$ and the constant k is assumed to be positive. Since the left hand side in (20) becomes negative if λ_1/λ_2 is close to zero, this inequality cannot be satisfied if the eigenvalues differ too much in magnitude. Thus, the criterion $\mathcal{D}_1 L > 0$ can only be satisfied if the ratio of the eigenvalues λ_1/λ_2 of $\mathcal{H}L$ is sufficiently close to one:

$$\frac{2k}{1-2k+\sqrt{1-4k}} \leq \frac{\lambda_1}{\lambda_2} \leq 1 \tag{21}$$

in other words only if the local image pattern contains second-order information along two orthogonal directions. The parameter k makes it possible to vary the selectivity of this operator where reasonable values of k can roughly be obtained in the interval $k \in [0.04, 0.10]$ (and we have here used $k = 0.06$ for the experiments in this paper).

Compared to the determinant of the Hessian operator, however, a main difference is that the operator $\mathcal{D}_1 L$ does not at all respond to saddle-like features. If we are interested in such features, we can define an alternative *signed* operator:

$$\tilde{\mathcal{D}}_1 L = \begin{cases} \det \mathcal{H}L - k \text{ trace}^2 \mathcal{H}L \\ \quad \text{if } \det \mathcal{H}L - k \text{ trace}^2 \mathcal{H}L > 0 \\ \det \mathcal{H}L + k \text{ trace}^2 \mathcal{H}L \\ \quad \text{if } \det \mathcal{H}L + k \text{ trace}^2 \mathcal{H}L < 0 \\ 0 \quad \text{otherwise} \end{cases} \tag{22}$$

At points where $\mathcal{D}_1 L > 0$ it follows that the signed operator $\tilde{\mathcal{D}}_1 L = \mathcal{D}_1 L$ and for such points the signed operator $\tilde{\mathcal{D}}_1 L$ will have similar properties as the unsigned operator $\mathcal{D}_1 L$. In practice, these points may for example correspond to bright or dark blobs. For saddle-like points, where $\det \mathcal{H}L < 0$, it follows that this operator will only generate a non-zero response if both of the principal curvatures λ_1 and λ_2 (with $|\lambda_1| \leq |\lambda_2|$) are sufficiently different from zero, *i.e.*, if the ratio between their absolute values is sufficiently close to one:

$$\frac{2k}{1-2k+\sqrt{1-4k}} \leq \frac{|\lambda_1|}{|\lambda_2|} \leq 1. \tag{23}$$

Hence, the signed operator $\tilde{\mathcal{D}}_1 L$ can be seen as a generalization of the unsigned operator $\mathcal{D}_1 L$ to make it possible to detect local saddle-like features.

In analogy with the Shi and Tomasi corner detector, we can also define an operator based on the minimum absolute eigenvalue of the Hessian matrix

$$\mathcal{D}_2 L = \min(|L_{pp}|, |L_{qq}|) \tag{24}$$

where L_{pp} and L_{qq} denote the eigenvalues of the Hessian matrix ordered such that $L_{pp} \leq L_{qq}$ (see Lindeberg [100, 103] for explicit expressions).

In analogy with the previous treatment of signed or unsigned versions of the $\mathcal{D}_1 L$ operator, we can also define a signed version of the $\mathcal{D}_2 L$ operator according to

$$\tilde{\mathcal{D}}_2 L = \begin{cases} L_{pp} & \text{if } |L_{pp}| < |L_{qq}| \\ L_{qq} & \text{if } |L_{qq}| < |L_{pp}| \\ (L_{pp} + L_{qq})/2 & \text{otherwise} \end{cases} \tag{25}$$

Figure 3 shows examples of computing these four types of interest points from a grey-level image. By comparing these results to the results in Fig. 2, we can first of all note that compared to the Laplacian operator, the new differential interest point detectors $\mathcal{D}_1 L$, $\tilde{\mathcal{D}}_1 L$, $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$ do not respond to elongated ridge structures and they do not give rise to responses outside the edges of the objects either. As for the determinant of the Hessian operator, this is a consequence of the $\mathcal{D}_1 L$, $\tilde{\mathcal{D}}_1 L$, $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$ operators requiring strong second-order responses in the two orthogonal eigendirections of the Hessian matrix.

The responses from Hessian feature strength measures $\mathcal{D}_1 L$ and $\tilde{\mathcal{D}}_1 L$ are rather similar to the determinant of the Hessian $\det \mathcal{H}L$, with the difference that the responses of $\mathcal{D}_1 L$ and $\tilde{\mathcal{D}}_1 L$ are more selective to corner like structures with more similar contributions from the two orthogonal directions and that the unsigned $\mathcal{D}_1 L$ operator does not respond to saddle-like image structures.

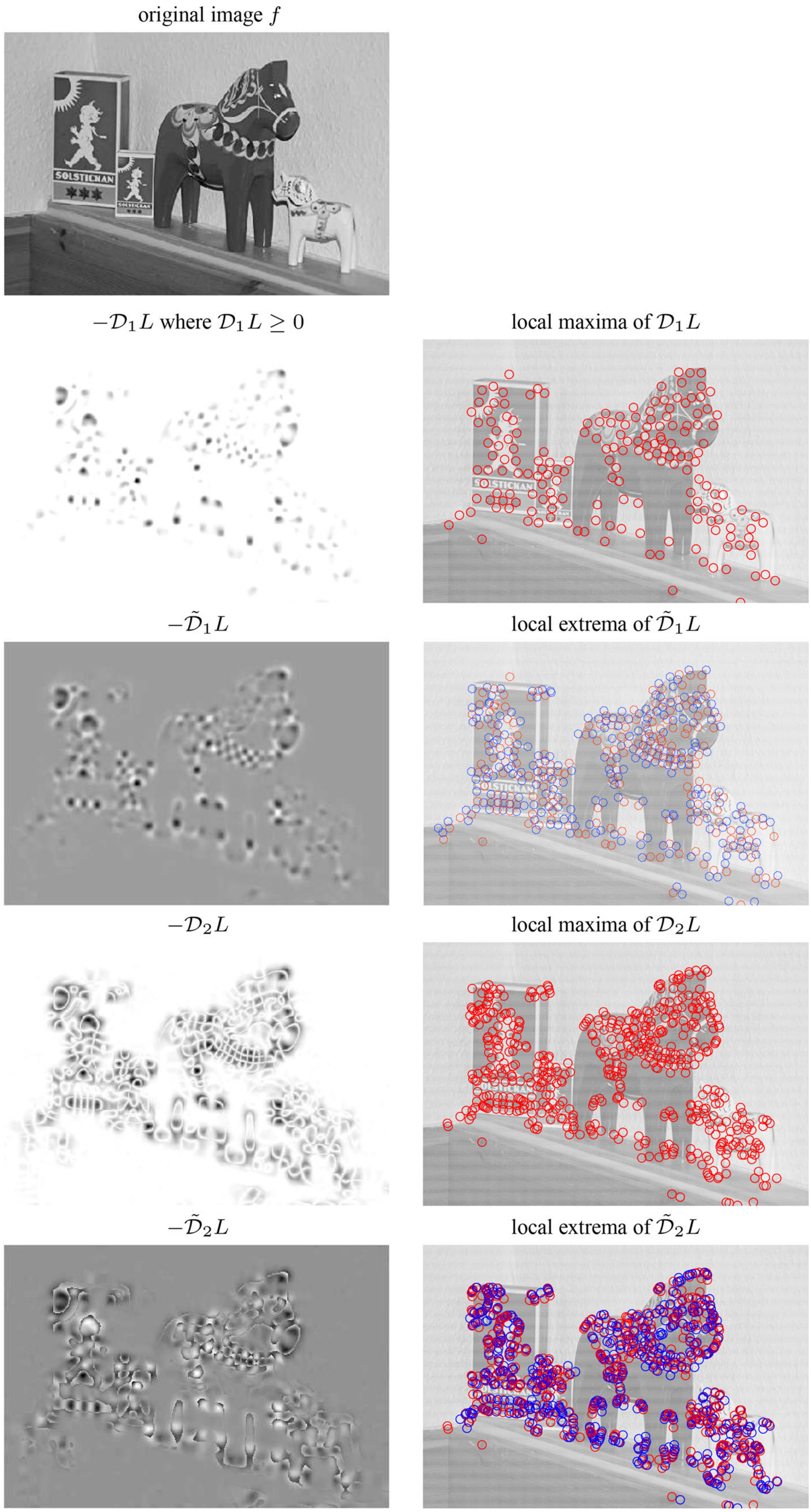
When used alone, the Hessian feature strength measures $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$ may lead to rather dense distributions of interest points in regions containing second-order image structures. When combined with complementary thresholding on either $\mathcal{D}_1 L > 0$ or $\tilde{\mathcal{D}}_1 L > 0$, these operators, in particular the signed $\tilde{\mathcal{D}}_2 L$ operator, can however lead to sparse sets of high quality interest points (see Sect. 6.2; Fig. 4).

Experimentally, the new differential interest point detectors $\mathcal{D}_1 L$, $\tilde{\mathcal{D}}_1 L$, $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$ can be shown to perform very well and to allow for image features with better repeatability properties under affine and perspective transformations than the more traditional Laplacian, difference-of-Gaussians or Harris–Laplace operators.

5 Behaviour of Interest Point Detectors Under Affine Image Transformations

In this section, we shall analyse the theoretical properties of the above mentioned feature detectors under affine transformations of the image domain. (Initially, we disregard the effect of the rotationally symmetric Gaussian smoothing operation and focus on the transformation properties of the differential expressions used for defining the different interest operators).

Consider an image $f(x, y)$ and define an affine transformed image pattern $f'(x', y')$ according to $f'(x', y') = f(x, y)$ for $(x', y')^T = A\,(x, y)^T$. Then, it follows from (14) that the *determinant of the Hessian* transforms according to



original image f

$-\mathcal{D}_1 L$ where $\mathcal{D}_1 L \geq 0$

local maxima of $\mathcal{D}_1 L$

$-\tilde{\mathcal{D}}_1 L$

local extrema of $\tilde{\mathcal{D}}_1 L$

$-\mathcal{D}_2 L$

local maxima of $\mathcal{D}_2 L$

$-\tilde{\mathcal{D}}_2 L$

local extrema of $\tilde{\mathcal{D}}_2 L$

Fig. 3 Differential interest point detectors at a fixed scale defined from the Hessian feature strength measures $\mathcal{D}_1 L$, $\tilde{\mathcal{D}}_1 L$, $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$ computed at scale $t = 32$ with corresponding features obtained by detecting local maxima at a fixed scale, with thresholding on the magnitude of the response with $C_{\mathcal{D}_1 L} = 10^2/4 = 25$ and $C_{\mathcal{D}_2 L} = 10/2 = 5$. (Image size: 512×350 pixels. *Red circles* denote local maxima of the operator response, whereas *blue circles* represent local minima.)

$$\det(\mathcal{H}f')(x', y') = \frac{1}{(\det A)^2} \det(\mathcal{H}f)(x, y) \tag{26}$$

implying that local extrema of this entity are preserved under affine transformations. In this respect, the determinant of the Hessian operator is *affine covariant*.

With regard to interest point detection, the affine covariant property specifically implies that the determinant of the Hessian can be expected to give *qualitatively similar responses for corner structures with different opening angles* as well as for image structures that are deformed in different ways under perspective mappings corresponding to *an object viewed from different viewing directions*.

For the *Laplacian* operator the corresponding transformation property is, however, much more complex

$$f'_{x'x'} + f'_{y'y'} = \frac{1}{(\det A)^2} \Big((a_{12}^2 + a_{22}^2) f_{xx} - 2(a_{11}a_{12} + a_{21}a_{22}) f_{xy} + (a_{11}^2 + a_{21}^2) f_{yy} \Big) \tag{27}$$

implying that we cannot in general assume that local maxima or minima of the Laplacian are to be preserved under general affine transformations, only for the specific similarity subgroup consisting of combined rotations and uniform scaling transformations for which $a_{11} = a_{22}$ and $a_{12} = -a_{21}$. This operator is therefore *not* affine covariant.

With a proper definition of window functions for the definition of the second-moment matrix in equation (10), it follows from equation (15) that the *determinant of the second-moment matrix* transforms as

$$\det \mu' = \frac{1}{(\det A)^2} \det \mu \tag{28}$$

which means that we can expect that local maxima of this operator will be preserved under affine transformations, and this operator is also *affine covariant*.

The transformation property of trace μ is, however, more complex, in analogy with the Laplacian operator. In this respect, the *Harris cornerness measures H* and the *Shi-and-Tomasi measure ST* do *not* possess theoretical affine covariance properties. Since the parameter k in the Harris operator is small, however, a *major contribution* of the Harris operator originates from the affine covariant operator det μ.

Regarding the new *Hessian feature strength measures* $\mathcal{D}_1 L$, $\tilde{\mathcal{D}}_1 L$, $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$, it follows from the fact that $\mathcal{D}_1 L$ and $\tilde{\mathcal{D}}_1 L$ contain a combination of $\det \mathcal{H}L$ and trace $\mathcal{H}L$, where $\det \mathcal{H}f$ is affine covariant while trace $\mathcal{H}f$ is not, that these operator will *not* be affine covariant. Similar conclusions can be drawn for the operators $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$. For the operators $\mathcal{D}_1 L$ and $\tilde{\mathcal{D}}_1 L$, however, a *major contribution* comes from the affine covariant operator $\det \mathcal{H}L$.

To conclude, the determinant of the Hessian $\det \mathcal{H}L$ and the determinant of the second-moment matrix det μ do both possess affine covariant properties disregarding the effect of rotationally symmetric Gaussian smoothing. For the Hessian feature strength measures $\mathcal{D}_1 L$ and $\tilde{\mathcal{D}}_1 L$ as well as for the Harris measure H, a major contribution originates from an affine covariant differential entity although there is also a sometimes non-negligible contribution from another differential entity that is not affine covariant. The Laplacian, the Shi-and-Tomasi and the Hessian feature strength measures $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$ are, however, not affine covariant.

All these differential entities are *scale covariant* in the sense that they transform according to a self-similar scaling law

$$\mathcal{D}f' = \frac{1}{s^{M_{\mathcal{D}}}} \mathcal{D}f \tag{29}$$

for any uniform scaling transformation $f'(x', y') = f(x, y)$ of the image domain $(x', y')^T = s\,(x, y)^T$ by a factor $s > 0$.

Notes regarding the affine covariant properties. For this analysis, it should be noted that when applied to real-world data, these differential geometric feature detectors are to be computed from a linear scale-space representation based on rotationally symmetric Gaussian filters. This scale-space concept is not closed under general affine transformations, only under similarity transformations consisting of combinations of rotations and uniform scaling transformations. For this reason, the affine covariant properties will not hold for the entire chain of image operations. Nevertheless, and as we will show experimentally in Sect. 9, the interest point detectors that possess theoretical covariance properties under general affine transformations will also lead to better repeatability properties than those without when combined with a rotationally symmetric Gaussian smoothing operation.

If full affine covariance properties are desired, this can be accomplished by replacing the rotationally symmetric Gaussian scale space by an *affine scale space* (Lindeberg [95,105,107,115]). Then, it will be possible to achieve full closedness and covariance properties of the feature responses under general (non-degenerate) affine transformations.

6 Thresholding

6.1 Magnitude Thresholding on Scale-Normalized Response

When detecting interest points using the above mentioned differential descriptors, it is natural to complement the detection of positive local maxima and negative local minima of the differential entities by thresholding on the magnitude of the response. When expressing threshold values for such thresholding, it is furthermore natural to express the magni-

Table 1 Relationships between *scale-normalized thresholds* for the different types of scale-invariant interest point detectors $\mathcal{D}L = \nabla^2 L$, $\det \mathcal{H}L$, $\mathcal{D}_1 L$, $\tilde{\mathcal{D}}_1 L$, $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$ as derived in (Lindeberg [110]) using scale-normalized derivatives with $\gamma = 1$, where we will here throughout use C between 5 and 10 for image data in the range $f \in [0, 255]$

Feature detector	$\mathcal{D}L$	$C_{\mathcal{D}L}$
Laplacian	$\nabla^2 L_{norm} = t\,(L_{xx} + L_{yy})$	$C_{\nabla^2 L} = C$
determinant of the Hessian	$\det \mathcal{H}_{norm} L = t^2\,(L_{xx} L_{yy} - L_{xy}^2)$	$C_{\det \mathcal{H} L} = C^2/4$
Hessian feature strength I	$\mathcal{D}_{1,norm} L = t^2\,(L_{xx} L_{yy} - L_{xy}^2 - k\,(L_{xx} + L_{yy})^2)$	$C_{\mathcal{D}_1 L} = (1 - 4k)\,C^2/4$
Hessian feature strength $\tilde{\mathrm{I}}$	$\tilde{\mathcal{D}}_{1,norm} L = t^2\,(L_{xx} L_{yy} - L_{xy}^2 \pm k\,(L_{xx} + L_{yy})^2)$	$C_{\tilde{\mathcal{D}}_1 L} = (1 - 4k)\,C^2/4$
Hessian feature strength II	$\mathcal{D}_{2,norm} = t\,\min(\lvert L_{pp}\rvert, \lvert L_{qq}\rvert)$	$C_{\mathcal{D}_2 L} = C/2$
Hessian feature strength $\widetilde{\mathrm{II}}$	$\tilde{\mathcal{D}}_{2,norm} L = t\,(L_{pp} \text{ or } L_{qq})$	$C_{\tilde{\mathcal{D}}_2 L} = C/2$
Harris-Laplace	$H_{norm} = t^2\,(\det \mu - k\,\mathrm{trace}^2\,\mu)$	$C_H = (1 - 4k)\,C^4/256$
Shi and Tomasi	$ST_{norm} = t\,\min(\nu_1, \nu_2)$	$C_{ST} = C^2/16$

The expressions for the Harris–Laplace operator and the Shi-and-Tomasi operator are based on the assumption of a relative integration scale of $r = 1$

tude of the response in terms of scale-normalized derivatives according to Eq. (5) with $\gamma = 1$, to be able to compare magnitude values at different scales.

For feature detectors that are defined in terms of *pointwise differential expressions* (*i.e.*, $\nabla^2 L$, $\det \mathcal{H}L$, $\mathcal{D}_1 L$, $\tilde{\mathcal{D}}_1 L$, $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$), we therefore perform thresholding on the magnitude of the response according to

$$|\mathcal{D}_{norm} L| \geq C_{\mathcal{D}L}. \tag{30}$$

For feature detectors that are defined in terms of *integrated differential expressions* from the second-moment matrix (*i.e.* the Harris, $\det \mu$ and Shi and Tomasi operators), we express the first-order partial derivatives L_x and L_y in terms of scale-normalized derivatives with $\gamma = 1$ to form scale-normalized feature strength measures for the Harris measure H, the determinant of the second moment matrix $\det \mu$ or the Shi and Tomasi measure ST. It can be shown [104] that this normalization is sufficient to be able to compare entities derived from the second-moment matrix at different scales.

Since the different feature detectors are of different dimensionality in terms of powers of the intensity and orders as well as powers of differentiation, it follows that the threshold value $C_{\mathcal{D}L}$ for a specific feature detector $\mathcal{D}_{norm} L$ must depend on the type of feature detector. By studying the scale-normalized responses of the different types of feature detectors to a Gaussian blob, theoretical relationships between thresholding values can be derived between the different interest point detectors as shown in Table 1. Such thresholds with $C = 10$ (for image data in the range $f \in [0, 255]$) were used for generating the illustrations in Figs. 2 and 3. When performing image-based matching, we have often found it valuable to decrease the parameter C somewhat to $C = 5$ or $C = 7.5$.

6.2 Complementary Thresholding on Other Measures of Feature Strength

In addition to thresholding on the magnitude of the scale-normalized response, a method for detecting interest points may also benefit from further selection criteria. For example, when Lowe [119] used local extrema of differences of Gaussians as basic features for his system for image based recognition, he noted that undesired feature responses may occur near edges and proposed to filter these away by analyzing the eigenvalues of the Hessian matrix. Since the difference-of-Gaussians operator can be seen as an approximation of the Laplacian operator (see Appendix A), a similar effect occurs for Laplacian responses.

Inspired by this idea, and given the definition of the feature strength measure $\mathcal{D}_1 L$ (18) from the Hessian matrix, we will use the criterion

$$\mathcal{D}_1 L = L_{xx} L_{yy} - L_{xy}^2 - k\,(L_{xx} + L_{yy})^2 \geq 0 \tag{31}$$

as a complementary thresholding criterion for Laplacian features. Motivated by experimental results to be presented later, we will also apply such thresholding on $\mathcal{D}_1 L \geq 0$ alternatively thresholding on $\tilde{D}_1 L \geq 0$ to determinant of the Hessian features $\det \mathcal{H}L$, and features from the Hessian strength measures $\mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$. Such *complementary thresholding*[1] can significantly improve the repeatability properties of interest point detectors.

Figure 4 illustrates the effect of performing such thresholding for some of the previously detected image features.

[1] This terminology means that if feature detection is performed using a differential feature detector $\mathcal{D}_A$ and if these responses are thresholded using *another* differential expression, say $\mathcal{D}_B \geq 0$, then the feature detector $\mathcal{D}_A$ is complemented by *complementary thresholding* on $\mathcal{D}_B \geq 0$.

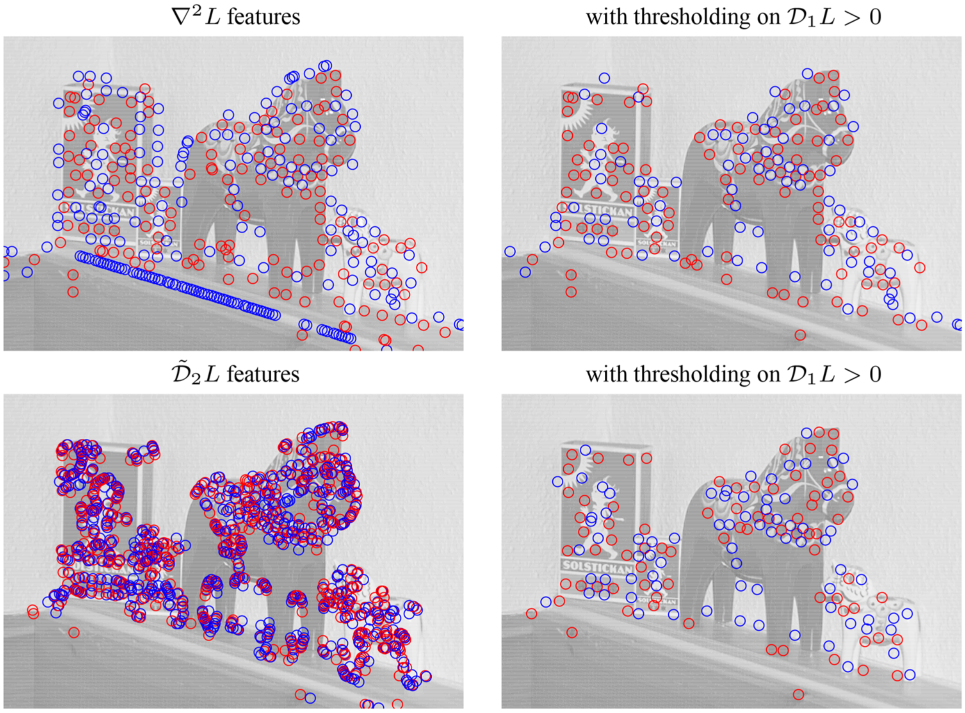

$\nabla^2 L$ features
with thresholding on $\mathcal{D}_1 L > 0$
$\tilde{\mathcal{D}}_2 L$ features
with thresholding on $\mathcal{D}_1 L > 0$

Fig. 4 Illustration of performing complementary thresholding on Laplacian $\nabla^2 L$ and Hessian feature strength $\tilde{\mathcal{D}}_2 L$ features using the sign of the Hessian feature strength measure $\mathcal{D}_1 L$. For these fixed scale feature detectors, operating at the scale $t = 32$, thresholding has also been performed on the magnitude of the response with $C_{\nabla^2 L} = 10$ and $C_{\tilde{\mathcal{D}}_2 L} = 10/2 = 5$. (Image size: 512 × 350 pixels. *Red circles* denote local maxima of the operator response, while *blue circles* represent local minima.)

As can be seen from the results, thresholding on $\mathcal{D}_1 L > 0$ suppresses Laplacian features along elongated structures. After complementary thresholding there are no longer any responses to the oblique ridge structure in the lower part of the image, and much fewer responses outside the edges of the objects. For the Hessian interest feature strength operator $\tilde{\mathcal{D}}_2 L$, the selective properties increase substantially by complementary thresholding on $\mathcal{D}_1 L > 0$. The set of remaining interest points after thresholding is much sparser.

7 Scale Selection Mechanisms

7.1 Scale Selection From γ-normalized derivatives

In Lindeberg [93,95,100,101] a general framework for automatic scale selection was proposed based on the idea of detecting local extrema over scale of *γ-normalized derivatives* according to (5). It was shown that *local extrema over scale* of homogeneous polynomial differential invariants $\mathcal{D}_{norm} L$ expressed in terms of γ-normalized Gaussian derivatives are transformed in a *scale-covariant* way:

> If some scale-normalized differential invariant $\mathcal{D}_{norm} L$ assumes a local extremum over scale at scale t_0 in scale space, then under a uniform rescaling of the input pattern by a factor s there will be a local extremum over scale in the scale space of the transformed signal at scale $s^2 t_0$.

By performing simultaneous scale and spatial selection, by detecting *scale-space extrema*, where the scale-normalized differential expression $\mathcal{D}_{norm} L$ assumes local extrema with respect to both space and scale, constitutes a general framework for detecting *scale-invariant interest points*. Such scale-space extrema are characterized by the first-order derivatives with respect to space and scale being zero

$$\nabla(\mathcal{D}_{norm} L) = 0 \quad \text{and} \quad \partial_t(\mathcal{D}_{norm} L) = 0 \tag{32}$$

and the composed Hessian matrix over both space and scale

$$\mathcal{H}_{(x,y;\,t)}(\mathcal{D}_{norm} L) = \begin{pmatrix} \partial_{xx} & \partial_{xy} & \partial_{xt} \\ \partial_{xy} & \partial_{yy} & \partial_{yt} \\ \partial_{xt} & \partial_{yt} & \partial_{tt} \end{pmatrix} (\mathcal{D}_{norm} L) \tag{33}$$

being either positive or negative definite.

This scale selection method also provides a way of ranking image features on significance by the magnitude of the scale-normalized response $|\mathcal{D}_{norm} L|$ at the scale-space extremum. These magnitude values as well as the associated significance ranking are scale invariant if $\gamma = 1$.

In Lindeberg [95,101] scale-space extrema of the Laplacian and scale-space extrema of the determinant of the Hessian were proposed as general purpose blob detectors/interest point detectors. Here, we complement these interest point detectors by complementary thresholding on either of the Hessian feature strength measures $\mathcal{D}_1 L > 0$ or $\tilde{\mathcal{D}}_1 L > 0$ and additionally emphasize the possibility of

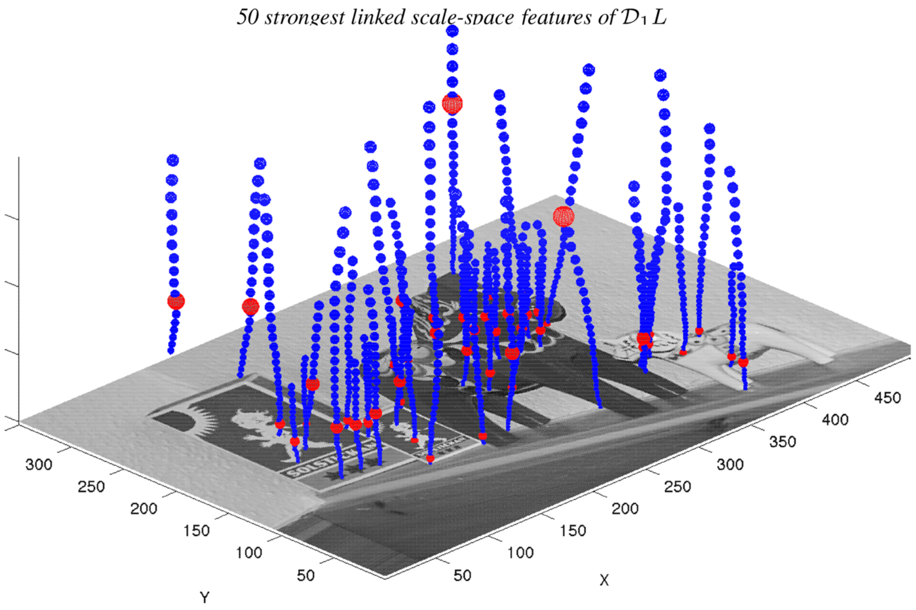


Fig. 5 3-D illustration of *feature trajectories over scale* (in *blue*) and selected scales from global extrema of the scale-normalized response along each feature trajectory (in *red*) when computing a scale-space primal sketch for the Hessian feature strength measure $\mathcal{D}_1 L$. Note how coarser scales are selected for the larger size objects than for the corresponding smaller size objects. Only the 50 strongest interest points are shown. Each feature trajectory is delimited from above and below by bifurcation events (not shown here), which may generically be of either of the types: annihilation, merge, split or creation. (Image size: 512 × 350 pixels. The vertical dimension represents the scale level in scale space in units of $\sigma = \sqrt{t}$, and translated along the vertical direction such that the minimum scale $t_{min} = 2$ is mapped to the level where the underlying image is shown.)

using negative responses of the determinant of the Hessian for detecting saddle-like interest points. We also apply scale-space extrema detection to the new Hessian feature strength measures $\mathcal{D}_1 L, \tilde{\mathcal{D}}_1 L, \mathcal{D}_2 L$ and $\tilde{\mathcal{D}}_2 L$ leading to four new interest point detectors based on scale-space extrema detection.

The Harris operator has been previously combined with scale selection using local extrema over scale of the scale-normalized Laplacian, leading to the Harris–Laplace operator (Mikolajczyk and Schmid [124]). Here, we will also combine the Harris operator with scale selection using local extrema over scale of the determinant of the Hessian, leading to a new Harris–detHessian operator.

7.2 Scale Linking

We extend this scale selection approach by *linking* image features at different scales into *feature trajectories* over scale and performing scale selection by either the strongest response over each feature trajectory or weighted averaging of scale values along the feature trajectory, with each feature trajectory delimited by a minimum scale t_{min} and a maximum scale t_{max} where bifurcation events occur (Fig. 5).

7.2.1 Feature Trajectories Over Scale

A rationale for performing scale linking is that if we detect some image feature at a position $(x_0, y_0)^T$ and scale t_0 in scale space, then it will generically be possible to detect corresponding image features at slightly coarser or finer scales. Formally, such a construction can be justified by the implicit function theorem. For our interest points detectors at a fixed scale, defined from local spatial extrema of either of the differential expressions

$$\mathcal{D}L \in \left\{ \nabla^2 L, \det \mathcal{H}L, \mathcal{D}_1 L, \tilde{\mathcal{D}}_1 L, \mathcal{D}_2 L, \tilde{\mathcal{D}}_2 L \right\} \tag{34}$$

the presence of an image feature at a position $(x_0, y_0;\ t_0)$ is defined by

$$\nabla(\mathcal{D}L)|_{(x_0, y_0;\ t_0)} = \left. \begin{pmatrix} \partial_x(\mathcal{D}L) \\ \partial_y(\mathcal{D}L) \end{pmatrix} \right|_{(x_0, y_0;\ t_0)} = 0 \tag{35}$$

with the additional condition that the Hessian matrix of $\mathcal{D}L$ should be either positive or negative definite. The implicit function theorem then ensures that there exists some smooth function $w_0(t) = (x_0(t), y_0(t))^T$ in some neighbourhood I_{t_0} of t_0 such that the point $(x_0(t), y_0(t);\ t_0)$ is a critical point for the mapping $(x, y)^T \to (\mathcal{D}L)(x, y;\ t)$ and the type of critical point remains the same as long as the Hessian matrix $\mathcal{H}(\mathcal{D}L)$ is non-singular. Specifically, the local *drift velocity* at $(x_0, y_0;\ t_0)$ is given by Lindeberg [89,95]:

$$\begin{aligned} & w'(x_0, y_0;\ t_0) \\ & \quad = -\,(\mathcal{H}(\mathcal{D}L))|^{-1}_{(x_0, y_0;\ t_0)}\ \partial_t(\nabla(\mathcal{D}L))|_{(x_0, y_0;\ t_0)} \end{aligned} \tag{36}$$

(see also Kuijper and Florack [77] for a more detailed study of drift velocities restricted to critical points of the raw image intensity). In other words, if $(x_0, y_0;\ t_0)$ is a local maximum (minimum) point of the differential descriptor $\mathcal{D}L$ then there exists a curve over scales through this point, such that every point on this curve is also a local maximum (minimum) of $\mathcal{D}L$ at that scale. This curve is delimited by two scale levels t_{min} and t_{max} where the Hessian matrix of $\mathcal{D}L$ degenerates (except the boundary cases $t_{min} = 0$ and $t_{max} = \infty$) and where bifurcation events[2] occur. Such a curve $w_0\colon]t_{min}, t_{max}[$ is called an *extremum path* of $\mathcal{D}L$ or a *feature trajectory*.

Using the theoretical and algorithmic framework developed in Lindeberg [104] explicit scale linking of pointwise image features into feature trajectories is performed (see Fig. 5) and bifurcation events are registered, leading to a scale-space primal sketch for differential descriptors, which we use as basis for detecting interest points in this work.

7.2.2 Scale Selection for Feature Trajectories

Along each feature trajectory T, scale selection can be performed either (i) by detecting *the strongest response over scales*

$$\hat{\tau}_T = \operatorname{argmax}_{\tau \in T} |(\mathcal{D}_{norm} L)(p(\tau);\ \tau)| \tag{37}$$

or (ii) by performing *weighted averaging of scale values* along the feature trajectory over scale according to

$$\hat{\tau}_T = \frac{\int_{\tau \in T} \tau\, \psi((\mathcal{D}_{norm} L)(p(\tau);\ \tau))\, d\tau}{\int_{\tau \in T} \psi((\mathcal{D}_{norm} L)(p(\tau);\ \tau))\, d\tau}. \tag{38}$$

Here, the integral is expressed in terms of *effective scale* [92]

$$\tau = \log t \tag{39}$$

to give a scale covariant construction of the corresponding scale estimates

$$\hat{t}_T = \exp \hat{\tau}_T \tag{40}$$

such that the resulting image features will be truly scale-invariant. For each feature trajectory an associated significance measure[3] W_T is defined as the integral of the scale-normalized feature responses along the feature trajectory [104]

$$W_T = \int_{\tau \in T} \psi(|(\mathcal{D}_{norm} L)(p(\tau);\ \tau)|)\, d\tau \tag{41}$$

where

$$\psi(|\mathcal{D}_{norm} L|) = w_{\mathcal{D}L}\, |\mathcal{D}_{norm} L|^a \tag{42}$$

represents a monotonically increasing self-similar transformation and

$$w_{\mathcal{D}L} = \frac{L_{\xi\xi}^2 + 2L_{\xi\eta}^2 + L_{\eta\eta}^2}{A(L_\xi^2 + L_\eta^2) + L_{\xi\xi}^2 + 2L_{\xi\eta}^2 + L_{\eta\eta}^2 + \varepsilon^2} \tag{43}$$

with $A = 4/e$ representing the relative feature weighting function between first- and second-order derivatives [80,99] and with $\varepsilon \approx 0.1$ representing an estimated noise level for image data in the range $[0, 255]$.

The motivation for performing scale selection by weighted averaging of scale-normalized differential responses over scale is analogous to the motivation for scale selection from local extrema over scale in the sense that interesting characteristic scale levels for further analysis should be obtained from the scales at which the differential operator assumes its strongest scale-normalized magnitude values over scale. Contrary to scale selection based on local extrema over scale, however, scale selection by weighted averaging over scale implies that the scale estimate will not only be obtained from the behaviour around the local extremum over scale, but also including the responses from all scales along a feature trajectory over scale. The intention behind this choice is that the scale estimates may therefore be less sensitive to local image perturbations.

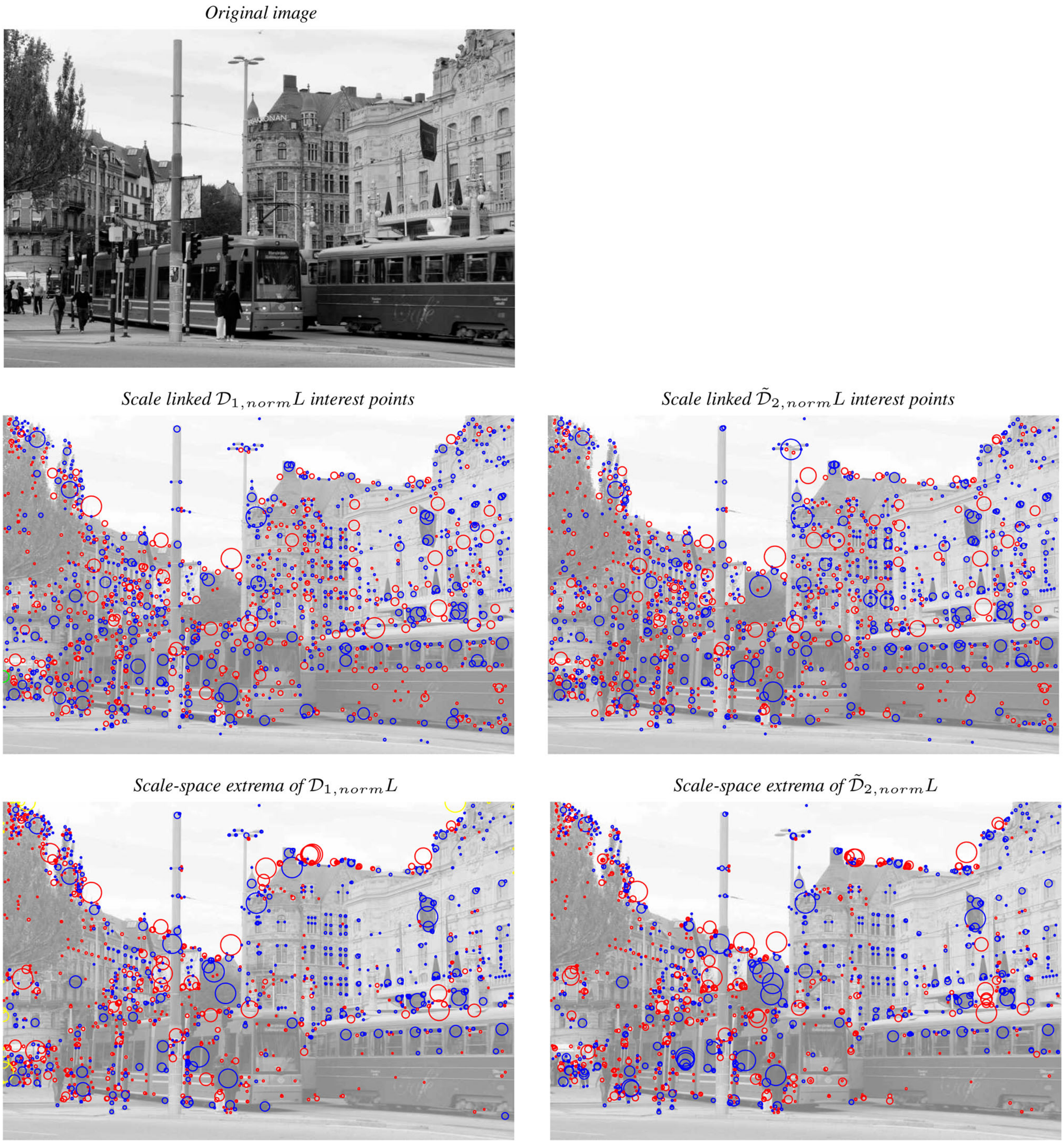
Figure 6 shows the result of detecting interest points in this way by applying either scale linking or scale-space extrema detection to the Hessian strength measures $\mathcal{D}_{1,norm} L$ and $\tilde{\mathcal{D}}_{2,norm} L$ (see also Fig. 11 for results from another scene with strong illumination variations). By comparing these and other results, it can be seen that interest point detection by scale-space extrema detection may give a relatively higher emphasis to image features with locally high and sharp contrasts, whereas interest point detection by scale linking may lead to a comparably higher ranking of image features that stand out from their local surroundings and do therefore get a longer life length in scale space.

7.2.3 Post-smoothing of Differential Entities

In the scale linking algorithm, an additional step *post-smoothing* of the differential expression $\mathcal{D}_{norm} L$ is per-

[2] These bifurcation events can be seen as a generalization of the notion of “top points” (Johansen [60,61]) or bifurcation events (Koenderink and van Doorn [71], Lindeberg [89], Damon [35], Kuijper and Florack [78]) from events between critical points of the smoothed image intensities L to bifurcation events between the critical points of any sufficiently well-behaved differential invariant $\mathcal{D}L$.

[3] An intuitive motivation for defining the significance measure in terms of an integral of scale-normalized feature responses over scale is a heuristic principle that image features that are stable over large ranges of scales should be more likely to be significant than image features that only exist over a shorter life length in scale space (Lindeberg [90, assumption 1 in section 3 on page 296]).



Original image

Scale linked $\mathcal{D}_{1,norm}L$ interest points

Scale linked $\tilde{\mathcal{D}}_{2,norm}L$ interest points

Scale-space extrema of $\mathcal{D}_{1,norm}L$

Scale-space extrema of $\tilde{\mathcal{D}}_{2,norm}L$

Fig. 6 Scale-invariant interest points obtained by (*middle row*) performing scale linking versus (*bottom row*) detecting scale-space extrema for the Hessian feature strength measures $\mathcal{D}_{1,norm}L$ and $\tilde{\mathcal{D}}_{2,norm}L$. The 1,400 strongest responses of each operator are shown. By comparing these and other experimental results, it can be seen that interest point detection by scale-space extrema detection may give a relatively higher emphasis to image features with locally high and sharp contrasts, whereas interest point detection by scale linking may lead to a comparably higher ranking of image features that stand out from their local surroundings and do therefore get a longer life length in scale space. The use of scale linking may also reduce multiple responses to the same underlying image structure. (Scale range: $t \in [2, 256]$. Image size: 800 × 600 pixels. The size of each *circle* represents the detection scale of the interest point. *Red circles* indicate that the Hessian matrix is negative definite (bright features), while *blue circles* that the Hessian matrix is positive definite (*dark features*).)

formed prior to the detection of local extrema over space or scale

$$\overline{(\mathcal{D}_{norm} L)}(x, y;\ t) = \int_{(u,v)\in\mathbb{R}^2} (\mathcal{D}_{norm} L)(x-u, y-u;\ t)\, g(u, v;\ c^2 t)\, du\, dv \tag{44}$$

with integration (post-smoothing) scale $t_{post} = c^2 t$ proportional to the differentiation scale t, where we have used $c = 3/8$ for all experiments in this article. A motivation for using such a post-smoothing step when linking image structures over scale is given in [110, Appendix A.1] and a detailed analysis of its properties in [110, Sects. 3.2–4.2].

7.2.4 Scale Selection Properties for a Gaussian Blob

In (Lindeberg [110, Sect. 3.1]) it is theoretically shown that when applied to a rotationally symmetric Gaussian blob model $f(x, y) = g(x, y;\ t_0)$ both scale-space extrema detection and weighed scale selection lead to similar scale estimates $\hat{t} = t_0$ for interest point detection based on the Laplacian $\nabla^2_{norm} L$, the determinant of the Hessian $\det \mathcal{H}_{norm} L$ and the Hessian feature strength measures $\mathcal{D}_{1,norm} L$, $\tilde{\mathcal{D}}_{1,norm} L$, $\mathcal{D}_{2,norm} L$ and $\tilde{\mathcal{D}}_{2,norm} L$. In this respect, all these interest point detectors are interchangeable.

When subjected to non-uniform affine image deformations outside the similarity group, the determinant of the Hessian $\det \mathcal{H}_{norm} L$ and the Hessian feature strength measures $\mathcal{D}_{1,norm} L$ and $\tilde{\mathcal{D}}_{1,norm} L$ do, however, have theoretical advantages in terms of affine covariance of the scale estimates or approximations thereof [110, Sect. 5.2.2].

8 Scale-Invariant Image Descriptors for Matching

In the following, we shall combine the above mentioned generalized scale-space interest points with local image descriptors. For each interest point, we will compute a complementary image descriptor in analogous ways as done in the SIFT and SURF operators, with the difference that the feature vectors are computed from Gaussian derivative responses in a scale-space representation instead of using a pyramid as done in the original SIFT operator (Lowe [119]) or a Haar wavelet basis as used in the SURF operator (Bay et al. [7]). A major reason for choosing a Gaussian derivative basis instead of a pyramid or Haar wavelets is to emphasize the underlying computational mechanisms of the image descriptors by disregarding as much as possible effects of discrete spatial subsampling. Another reason is to make it possible to combine different types of interest points with similar image descriptors for the purpose of comparison.

Since each one of the generalized scale-space interest point detectors is scale invariant, it follows that also the associated local image descriptors will be scale invariant, provided that these image descriptors are computed at scale levels proportional to the detection scales $\hat{t}$ of the generalized interest points and using window functions of radius proportional to the scale estimate in dimension length $\hat{\sigma} = \sqrt{\hat{t}}$.

8.1 Gauss-SIFT

For our SIFT-like image descriptor *Gauss-SIFT*, we compute image gradients ∇L at the detection scale $\hat{t}$ of the interest point. An orientation estimate is computed in a similar way as by Lowe [119], by accumulating a histogram of gradient directions $\arg \nabla L$ quantized into 36 bins with the area of the accumulation window proportional to the detection scale $\hat{t}$, and then detecting peaks in the smoothed orientation histograms. Multiple peaks are accepted if the height of the secondary peak(s) are above 80 % of the highest peak. Then, for each point on a 4 × 4 grid with the grid spacing proportional to the detection scale measured in units of $\hat{\sigma} = \sqrt{\hat{t}}$, a weighed local histogram of gradient directions $\arg \nabla L$ quantized into 8 bins is accumulated around each grid point, with the weights proportional to the gradient magnitude $|\nabla L|$ and a Gaussian window function with its area proportional to the detection scale $\hat{t}$ (see Fig. 7). To increase the accuracy of the local histograms, the local histograms are accumulated with the image measurements sampled at twice the spatial resolution of the image using bicubic interpolation and with trilinear interpolation for distributing the weighted increments for the sampled image measurements into adjacent histogram bins. The resulting 128-dimensional descriptor is normalized to unit sum to achieve contrast invariance, with the relative contribution of a single bin limited to a maximum value of 0.20.

8.2 Gauss-SURF

For our SURF-like image descriptor *Gauss-SURF*, we compute the following sums of derivative responses $\sum L_x$, $\sum |L_x|$, $\sum L_y$, $\sum |L_y|$ at the scale $\hat{t}$ of the interest point, for each one of 4 × 4 subwindows around the interest point as Bay et al. [7] and with similar orientation normalization as for the SIFT operator. The resulting 64-D descriptor is then normalized to unit length for contrast invariance.

9 Matching Properties Under Perspective Transformations

To evaluate the quality of the interest points with their associated local image descriptors, we apply bi-directional nearest-neighbour matching of the image descriptors in Euclidean

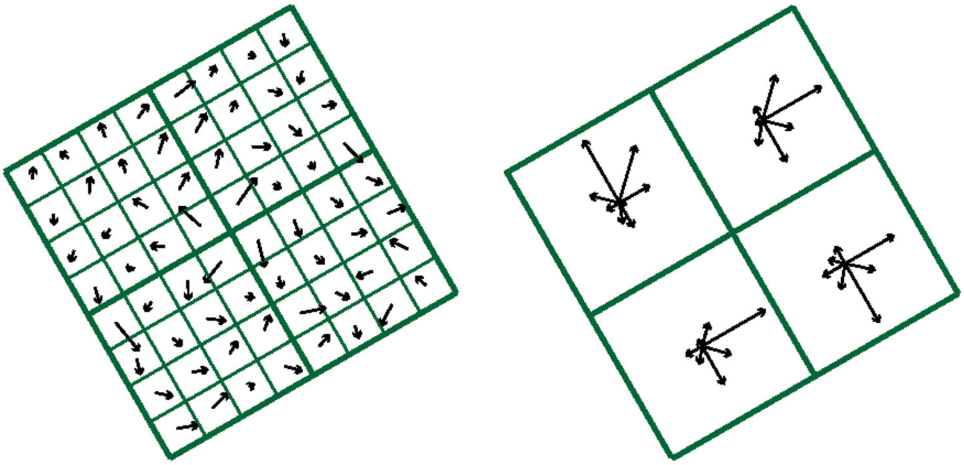


Fig. 7 Our Gauss-SIFT descriptor is defined in an analogous way as Lowe [119] defined his SIFT descriptor, by first computing an overall orientation of the interest point and then computing a 4 × 4 position-dependent histogram of gradient directions quantized into 8 bins, with the differences that (i) the Gauss-SIFT descriptor is defined from Gaussian derivatives instead of difference approximations in a pyramid and that (ii) we use our family of generalized interest points as initial keypoints instead of difference-of-Gaussian features. In this schematic illustration, a 2 × 2 grid is shown instead of 4 × 4 grid. With a 4 × 4 spatial grid and 8 bins for the gradient directions, one obtains a 128-D descriptor

norm. In other words, given a pair of images f_A and f_B with corresponding sets of interest points $A = \{A_i\}$ and $B = \{B_j\}$, a match between the pair of interest points (A_i, B_j) is accepted only if:

(i) A_i is the best match for B_j in relation to all the other points in A and, in addition,
(ii) B_j is the best match for A_i in relation to all the other points in B.

To suppress matching candidates for which the correspondence may be regarded as ambiguous, we furthermore require the ratio between the distances to the nearest and the next nearest image descriptor to be less than $r = 0.9$.

Next, we will evaluate the matching performance of such interest points with local image descriptors over a dataset of poster images with calibrated homographies over different amounts of perspective scaling and foreshortening.

9.1 Poster Image Dataset

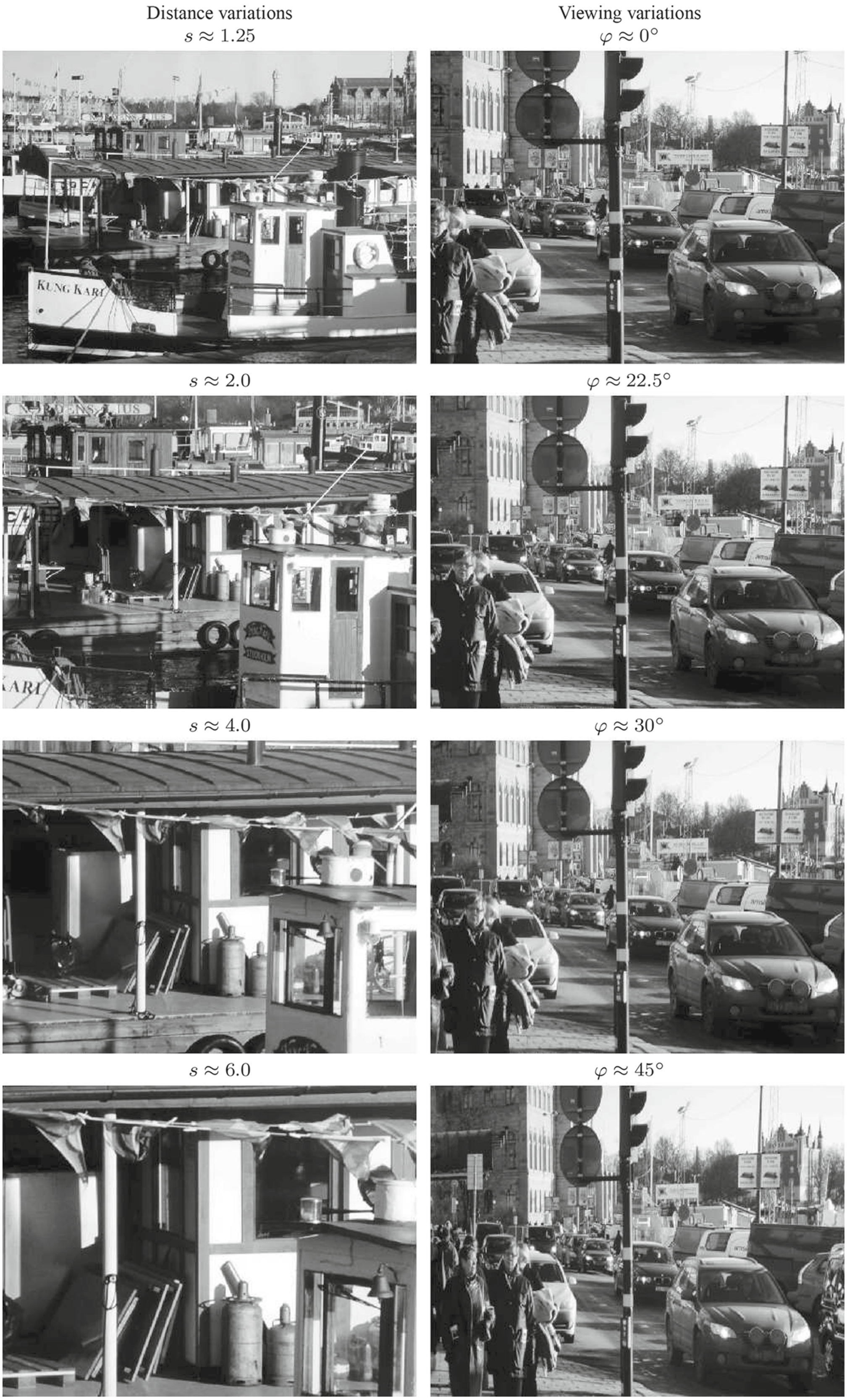
High-resolution photographs of approximately 4900 × 3200 pixels were taken of 12 outdoor and indoor scenes in natural city and office environments, from which poster printouts of size 100 × 70 cm were produced by a professional laboratory. Each such poster was then photographed from 14 different positions (see Fig. 8 for examples):

(i) 11 normal views leading to approximate scaling transformations with relative scale factors s approximately equal to 1.25, 1.5, 1.75, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0 and 6.0, and
(ii) 3 additional oblique views leading to foreshortening transformations with slant angles of about 22.5°, 30° and 45° relative to the frontal view with $s \approx 2.0$.

For the 11 normal views of each objects, homographies were computed between each pair of images using the ESM method (Benhimane and Malis [11]) with initial estimates of the relative scaling factors obtained from manual measurements of the distance between the poster surface and the camera. For the oblique views, for which the ESM method did not produce sufficiently accurate results, homographies were computed by first manually marking correspondences between the four images of each poster, computing an initial estimate of the homography using the linear method in Hartley and Zisserman [56, Algorithm 3.2, Page 92] and then computing a refined estimate by minimizing the Sampson approximation of the geometric error (Hartley and Zisserman [56, Algorithm 3.3, Page 98]).

The motivations for using such a poster image dataset for evaluation are that:

(i) the use of poster images from natural city and office environments should lead to a representative selection of image structures from natural scenes,
(ii) the use of planar posters implies that ground truth can be defined by homographies and calibration which may not otherwise be easy to achieve for natural 3-D scenes without 3-D reconstruction,



Distance variations

Viewing variations

$s \approx 1.25$

$\varphi \approx 0°$

$s \approx 2.0$

$\varphi \approx 22.5°$

$s \approx 4.0$

$\varphi \approx 30°$

$s \approx 6.0$

$\varphi \approx 45°$

Fig. 8 Illustration of images of posters from multiple views (*left column*) by varying the distance between the camera and the object for different frontal views, and (*right column*) by varying the viewing direction relative to the direction of the surface normal. (Image size: 768 × 576 pixels.)

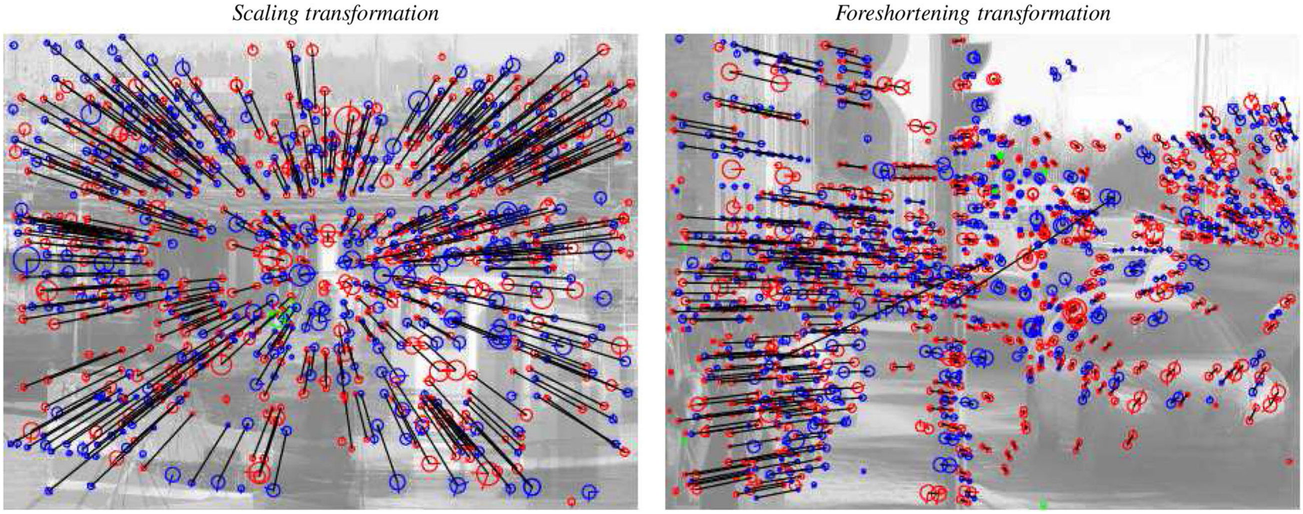


Fig. 9 Illustration of matching relations obtained by bi-directional matching of Gauss-SIFT descriptors computed at interest points of the signed Hessian feature strength measure $\tilde{\mathcal{D}}_{1,norm} L$ for (*left*) a scaling transformation and (*right*) a foreshortening transformation between pairs of poster images of the harbour and city scenes shown in Fig. 8. These illustrations have been generated by first superimposing bright copies of the two images to be matched by adding them. Then, the interest points detected in the two domains have been overlaid on the image data, and a *black line* has been drawn between each pair of image points that has been matched. *Red circles* indicate that the Hessian matrix is negative definite (*bright features*), *blue circles* that the Hessian matrix is positive definite (*dark features*), whereas *green circles* indicate that the Hessian matrix is indefinite (*saddle-like features*)

(iii) the use of a large range of variations in scale (up to a factor of 6) should provide a thorough test of the scale invariant properties of the interest points under scaling transformations,

(iv) the use of a multiple slant angles in the range between 22.5° and 45° should make it possible to investigate robustness of the interest point detectors to image deformations caused by moderate variations of the viewing direction relative to the object, and

(v) including image data for a sufficiently large number of scenes to enable statistical comparisons from the experimental results.

Specifically, the motivation for focusing the experimental evaluation on the robustness to scaling transformations and oblique perspective views corresponding to different amounts of foreshortening is that if we consider a perspective camera that views a regional surface patch of an object and linearize the non-linear perspective transformation locally by computing its derivative, we then around any image point (x_0, y_0) obtain a local affine transformation matrix A that can be decomposed into the form [96]

$$A = R_1 \, \text{diag}(\sigma_1, \sigma_2) \, R_2^{-1} \qquad (45)$$

where R_1 and R_2 can be forced to be rotation matrices, if we relax the requirement of non-negative entries in the diagonal elements σ_1 and σ_2 of a regular singular value decomposition. With this model, the geometric average of the absolute values of the diagonal entries

$$\sigma_{uniform} = \sqrt{|\sigma_1 \, \sigma_2|} \qquad (46)$$

corresponds to the amount of scaling, whereas the ratio $|\sigma_2/\sigma_1| = \cos\theta$ corresponds to the amount of foreshortening with θ denoting the slant angle. By studying the robustness to uniform scaling transformations and perspective foreshortening transformations, we do therefore investigate the sensitivity to the two harder components in the decomposition (45) of a locally linearized perspective image deformation.

9.2 Matching Criteria and Performance Measures

Figure 9 shows an illustration of point matches obtained between two pairs of images corresponding to a scaling transformation and a foreshortening transformation based on interest points detected using the $\tilde{\mathcal{D}}_{1,norm} L$ operator.

To make a judgement of whether two image features A_i and B_j matched in this way should be regarded as belonging to the same feature or not, we associate a scale dependent circle C_A and C_B to each feature, with the radius of each circle equal to the detection scale of the corresponding feature measured in units of the standard deviation $\sigma = \sqrt{t}$. Then, each such feature is transformed to the other image domain, using the homography and with the scale value transformed by a scale factor of the homography. The relative amount of overlap between any pair of circles is defined by forming the ratio between the intersection and the union of the two circles in a similar way as Mikolajczyk et al. [126] define a corresponding ratio for ellipses

$$m(C_A, C_B) = \frac{|\bigcap(C_A, C_B)|}{|\bigcup(C_A, C_B)|}. \tag{47}$$

We then accept a match if the relative overlap between a pair of mutually best matches is greater than 0.2. The motivation for representing the interest points by circles in this case is that the interest points are only scale invariant (since no affine shape adaptation process has been included here that would make the interest points affine invariant and thus motivate a representation in terms of ellipses). The motivation for using a liberal criterion on the overlap is that the previous criterion of mutually best pairwise matches implies a strong condition on the matches, so if a nearby match can be found given such a strong criterion it should then also be accepted.

Then, we measure the performance of the interest point detector by:

$$\text{efficiency} = \frac{\#(\text{interest points that lead to accepted matches})}{\#(\text{interest points})}$$

$$\text{1-precision} = \frac{\#(\text{rejected matches})}{\#(\text{accepted matches}) + \#(\text{rejected matches})}$$

The evaluation of the matching score is only performed for image features that are within the image domain for both images before and after the transformation. Moreover, only features within corresponding scale ranges are evaluated. In other words, if the scale range for the image f_A is $[t_{min}, t_{max}]$, then image features are searched for in the transformed image f_B within the scale range $[t'_{min}, t'_{max}] = [s^2\, t_{min}, s^2\, t_{max}]$, where s denotes an overall scaling factor of the homography. In the experiments below, we used $[t_{min}, t_{max}] = [4, 256]$.[4]

9.3 Experimental Results

Tables 2 and 3 show the result of evaluating 2 × 9 different types of scale-space interest point detectors with respect to the problem of establishing point correspondences between pairs of images on the poster dataset. Each interest point detector is applied in two versions (i) detection of scale-space extrema or (ii) using scale linking with scale selection from weighted averaging of scale-normalized feature responses along feature trajectories.

In addition to the 2 × 7 differential interest point detectors described in Sect. 4, we have also included 2 × 2 interest point detectors derived from the Harris operator [55]: (i) the Harris–Laplace operator [124] based on spatial extrema of the Harris measure and scale selection from local extrema over scale of the scale-normalized Laplacian, (ii) a scale-linked version of the Harris–Laplace operator with scale selection by weighted averaging over feature trajectories of Harris features [104], and (iii-iv) two Harris–detHessian operators analogous to the Harris–Laplace operators, with the difference that scale selection is performed based on the scale-normalized determinant of the Hessian instead of the scale-normalized Laplacian [104].

The experiments are based on detecting the $N = 800$ strongest interest points extracted from the first image, regarded as reference image for the homography. To perform the experimental evaluation over an approximate uniform density of interest points under scaling transformations, an adapted number of $N' = N/s^2$ strongest interest points is searched for (i) within the subwindow of the reference image that is mapped to the interior of the transformed image and (ii) in the transformed image, with s denoting the relative scaling factor between the two images.[5]

[4] The reason for prefiltering the interest points by position and scale is to prevent the performance measures from being primarily dominated by geometric parameters of the experimental setup. For example, with a relative scaling factor of $s > 1$ between two images, on average $1 - 1/s^2$ of the points in the first image will fall outside the domain of the transformed image if the image size is kept constant as in these experiments. In a corresponding manner, if an image feature is detected at scale level t_0 in the original image, it would be expected to be detected at scale level $s^2 t_0$ in the transformed image, because of the properties of the scale selection method described in Sect. 7. If non-matching scale ranges would be used for the evaluation, then there would be corresponding geometric limitations on the performance values because of mismatches between the scale ranges. With the used limitations of the spatial domains and the scale ranges, the performance values do therefore report the ratio of image features that have been matched in relation to those who could possibly be matched at all, given the geometry of the experimental setup.

[5] The reason for adapting the number of interest points to the amount of geometric scaling is that with a relative scaling factor $s > 1$ between two images, on average only N/s^2 of the points in the first image will be inside the domain of the second image. In previous experiments regarding repeatability properties of interest points, we have found that the repeatability scores may depend systematically on the number of image features used for the evaluation. If one would ask for the same number of images in the transformed image irrespective of the amount of scaling, that would effectively correspond to asking for a larger number of image features in the central part of the image with increasing amount of scaling. To prevent such geometric factors from dominating the performance values under variations in the amount of scaling, we have chosen to adapt the number of image features to a geometric transformation such that when performing matching between an image at scale factor s_1 and an image at scale factor $s_2 > s_1$, only the $N/(s_2/s_1)^2$ strongest image features are used for computing the performance measure. Thereby, the density of image features will always be the same in relation to the first image. The intention behind this choice is that the performance values should reflect how much harder is to match image features over a relative scale factor of say 5 to a corresponding matching over a relative scale factor of say 2, and not how the repeatability of the interest points depends on the thresholds for interest point detection. The actual selection of a lower number of image features is performed by sorting the interest points in decreasing order of significance, using the scale-normalized magnitude measure $|\mathcal{D}_{norm} L|$ at the scale-space extremum for scale-space extrema or the scale integrated significance measure W_T along feature trajectories (41) for interest points computed by scale linking.

Table 2 Performance measures obtained by matching different types of scale-space interest points with associated Gauss-SIFT image descriptors for the poster image dataset

Interest points		Scaling		Foreshortening		Average	
		Extr	Link	Extr	Link	Extr	Link
Efficiency: Gauss-SIFT image descriptor							
$\nabla^2_{norm} L$	$(\mathcal{D}_1 L > 0)$	0.7484	0.7994	0.7512	0.7574	0.7498	0.7784
$\det \mathcal{H}_{norm} L$	$(\mathcal{D}_1 L > 0)$	0.7721	*0.8225*	0.7635	*0.7932*	0.7678	*0.8079*
$\det \mathcal{H}_{norm} L$	$(\tilde{\mathcal{D}}_1 L > 0)$	0.7691	0.8163	0.7602	0.7841	0.7647	0.8002
$\mathcal{D}_{1,norm} L$		0.7719	**0.8280**	0.7596	**0.7977**	0.7658	**0.8128**
$\tilde{\mathcal{D}}_{1,norm} L$		0.7698	0.8241	0.7578	*0.7916*	0.7638	*0.8079*
$\mathcal{D}_{2,norm} L$	$(\mathcal{D}_1 L > 0)$	0.7203	0.8187	0.7111	0.7776	0.7157	0.7981
$\tilde{\mathcal{D}}_{2,norm} L$	$(\mathcal{D}_1 L > 0)$	0.7204	*0.8261*	0.7113	0.7766	0.7159	0.8014
Harris–Laplace		0.7002	0.7855	0.7046	0.7535	0.7024	0.7695
Harris–detHessian		0.7406	0.7608	0.7561	0.7319	0.7406	0.7463
1-precision: Gauss-SIFT image descriptor							
$\nabla^2_{norm} L$	$(\mathcal{D}_1 L > 0)$	0.0577	0.0336	0.0141	0.0163	0.0359	0.0250
$\det \mathcal{H}_{norm} L$	$(\mathcal{D}_1 L > 0)$	0.0544	*0.0333*	0.0133	**0.0127**	0.0339	*0.0230*
$\det \mathcal{H}_{norm} L$	$(\tilde{\mathcal{D}}_1 L > 0)$	0.0537	**0.0315**	0.0133	*0.0132*	0.0335	**0.0224**
$\mathcal{D}_{1,norm} L$		0.0543	0.0340	0.0135	*0.0133*	0.0339	*0.0236*
$\tilde{\mathcal{D}}_{1,norm} L$		0.0542	0.0340	0.0134	0.0134	0.0338	0.0237
$\mathcal{D}_{2,norm} L$	$(\mathcal{D}_1 L > 0)$	0.0512	0.0356	0.0174	0.0153	0.0343	0.0255
$\tilde{\mathcal{D}}_{2,norm} L$	$(\mathcal{D}_1 L > 0)$	0.0512	*0.0329*	0.0175	0.0143	0.0343	*0.0236*
Harris–Laplace		0.1272	0.0587	0.0306	0.0215	0.0789	0.0401
Harris–detHessian		0.1232	0.0664	0.0274	0.0264	0.0753	0.0464

The columns show from left to right: (i) the average performance over all pairs of perspective scaling transformations, (ii) the average performance over all pairs of perspective foreshortening transformations and (iii) the average total computed as the mean of the scaling and foreshortening scores. The columns labelled "extr" show results obtained by scale-space extrema detection, whereas the columns labelled "link" show results obtained by scale linking. (Within each type of experimental condition (scaling transformations/foreshortening transformations/combined average of these) the best result over all interest point detectors is shown in bold and the two next best results in italics)

This procedure is repeated for all pairs of images within the groups of distance variations or viewing variations respectively, implying up to 55 image pairs for the scaling transformations and 6 image pairs for the foreshortening transformations, *i.e.* up to 61 matching experiments for each one of the 12 posters, thus up to 732 experiments for each one of 2 × 9 interest point detectors.

As can be seen from the results of matching SIFT-like Gauss-SIFT image descriptors in Table 2, the interest point detectors based on scale linking generally lead to higher efficiency rates and lower 1-precision rates compared to the corresponding interest point detectors based on scale-space extrema detection. Specifically, the highest efficiency rates are obtained with the unsigned Hessian feature strength measure $\mathcal{D}_{1,norm} L$, followed by the signed Hessian feature strength measure $\tilde{\mathcal{D}}_{1,norm} L$ and the determinant of the Hessian operator $\det \mathcal{H}_{norm} L$ with complementary thresholding on $\mathcal{D}_{1,norm} L > 0$.

The lowest and thus the best 1-precision score is obtained with the determinant of the Hessian operator $\det \mathcal{H}_{norm} L$ with complementary thresholding on $\tilde{\mathcal{D}}_{1,norm} L > 0$, followed by the determinant of the Hessian operator $\det \mathcal{H}_{norm} L$ with complementary thresholding on $\mathcal{D}_{1,norm} L > 0$. In this respect, the inclusion of saddle-like image features with $\det \mathcal{H}_{norm} L$ as are accepted by the $\tilde{\mathcal{D}}_{1,norm} L$ operator can contribute to a lower number of rejected matches.

Among the more traditional feature detectors based on scale selection from local extrema over scale, the determinant of the Hessian operator $\det \mathcal{H}_{norm} L$ performs better than both the Laplacian operator $\nabla^2_{norm} L$ and the Harris–Laplace operator. We can also note that the Harris–Laplace operator can be improved by either scale linking or by replacing scale selection based on the scale-normalized Laplacian by scale selection based on the scale-normalized determinant of the Hessian. Specifically, the interest point detectors based on the Hessian feature strength measures $\tilde{\mathcal{D}}_{2,norm} L$ and $\tilde{\mathcal{D}}_{2,norm} L$ are very much improved by scale linking.

Table 3 shows corresponding results for interest point matching based on SURF-like Gauss-SURF descriptors. As can be seen from the results, the highest efficiency scores

Table 3 Performance measures obtained by matching different types of scale-space interest points with associated Gauss-SURF image descriptors for the poster image dataset

Interest points		Scaling		Foreshortening		Average	
		Extr	Link	Extr	Link	Extr	Link
Efficiency: Gauss-SURF image descriptor							
$\nabla^2_{norm} L$	$(\mathcal{D}_1 L > 0)$	0.7424	0.7832	0.7280	0.7140	0.7352	0.7486
$\det \mathcal{H}_{norm} L$	$(\mathcal{D}_1 L > 0)$	0.7656	0.8072	0.7402	*0.7504*	0.7529	*0.7788*
$\det \mathcal{H}_{norm} L$	$(\tilde{\mathcal{D}}_1 L > 0)$	0.7628	0.8015	0.7372	0.7430	0.7500	0.7723
$\mathcal{D}_{1,norm} L$		0.7661	**0.8126**	0.7354	**0.7537**	0.7507	**0.7831**
$\tilde{\mathcal{D}}_{1,norm} L$		0.7640	*0.8081*	0.7334	*0.7478*	0.7487	*0.7779*
$\mathcal{D}_{2,norm} L$	$(\mathcal{D}_1 L > 0)$	0.7157	0.8014	0.6870	0.7284	0.7013	0.7649
$\tilde{\mathcal{D}}_{2,norm} L$	$(\mathcal{D}_1 L > 0)$	0.7158	*0.8100*	0.6873	0.7328	0.7015	0.7714
Harris–Laplace		0.6948	0.7620	0.6724	0.6944	0.6836	0.7282
Harris–detHessian		0.7345	0.7381	0.7192	0.6705	0.7268	0.7043
1-precision: Gauss-SURF image descriptor							
$\nabla^2_{norm} L$	$(\mathcal{D}_1 L > 0)$	0.0611	0.0399	0.0217	0.0287	0.0414	0.0343
$\det \mathcal{H}_{norm} L$	$(\mathcal{D}_1 L > 0)$	0.0572	*0.0373*	*0.0210*	0.0232	0.0391	*0.0303*
$\det \mathcal{H}_{norm} L$	$(\tilde{\mathcal{D}}_1 L > 0)$	0.0566	**0.0356**	0.0214	0.0239	0.0390	**0.0298**
$\mathcal{D}_{1,norm} L$		0.0572	0.0381	**0.0207**	0.0221	0.0389	*0.0301*
$\tilde{\mathcal{D}}_{1,norm} L$		0.0571	0.0385	*0.0210*	0.0230	0.0391	0.0307
$\mathcal{D}_{2,norm} L$	$(\mathcal{D}_1 L > 0)$	0.0549	0.0392	0.0278	0.0282	0.0414	0.0337
$\tilde{\mathcal{D}}_{2,norm} L$	$(\mathcal{D}_1 L > 0)$	0.0549	*0.0365*	0.0279	0.0273	0.0414	0.0319
Harris–Laplace		0.1312	0.0654	0.0458	0.0409	0.0885	0.0532
Harris–detHessian		0.1271	0.0743	0.0406	0.0483	0.0838	0.0613

The columns show from left to right: (i) the average performance over all pairs of perspective scaling transformations, (ii) the average performance over all pairs of perspective foreshortening transformations and (iii) the average total computed as the mean of the scaling and foreshortening scores. The columns labelled "extr" show results obtained by scale-space extrema detection, whereas the columns labelled "link" show results obtained by scale linking. (Within each type of experimental condition (scaling transformations/foreshortening transformations/combined average of these) the best result over all interest point detectors is shown in bold and the two next best results in italics)

are again obtained for the unsigned and signed scale linked Hessian feature strength measures $\mathcal{D}_{1,norm} L$ and $\tilde{\mathcal{D}}_{1,norm} L$ followed by the determinant of the Hessian $\det \mathcal{H}_{norm} L$ with complementary thresholding on $\mathcal{D}_{1,norm} L > 0$. The lowest average 1-precision score is also obtained for the scale linked determinant of the Hessian $\det \mathcal{H}_{norm} L$ with complementary thresholding on $\tilde{\mathcal{D}}_{1,norm} L > 0$, followed by the determinant of the Hessian $\det \mathcal{H}_{norm} L$ with complementary thresholding on $\mathcal{D}_{1,norm} L > 0$, and the Hessian feature strength measure $\mathcal{D}_{1,norm} L$.

When comparing the results obtained for our Gauss-SIFT and Gauss-SURF image descriptors, we can see that the Gauss-SIFT image descriptors lead to both higher efficiency rates and lower 1-precision scores than the Gauss-SURF image descriptors. This qualitative relationship holds over all types of interest point detectors. In this respect, the pure image descriptor in the SIFT operator is clearly better than the pure image descriptor in the SURF operator. Specifically, more reliable image matches can be obtained by replacing the pure image descriptor in the SURF operator by the pure image descriptor in the SIFT operator.

Table 4 lists the five best combinations of interest point detectors and image descriptors in this evaluation as ranked on their efficiency values. For comparison, the results of our corresponding analogues of the SIFT operator with interest point detection from scale-space extrema of the Laplacian and our analogue of the SURF operator based on scale-space extrema of the determinant of the Hessian are also shown. As can be seen from the ranking, the best combinations of generalized points with Gauss-SIFT image descriptors perform better than the corresponding analogues of regular SIFT or regular SURF based on scale-space extrema of the Laplacian in combination with a Gauss-SIFT descriptor or the determinant of the Hessian in combination with a Gauss-SURF descriptor.

Figure 10 shows graphs of how the efficiency rate depends upon the amount of scaling for the scaling transformations and the difference in viewing angle for the foreshortening transformations. As can be seen from the graphs, the interest point detectors $\det \mathcal{H}_{norm} L$, $\mathcal{D}_{1,norm} L$ and $\tilde{\mathcal{D}}_{1,norm} L$ that possess affine covariance properties or approximations thereof (see Sect. 5 and [104,110]) do also have the best

Table 4 The five best combinations of interest points and image descriptors among the $2 \times 2 \times 9 = 36$ combinations considered in this experimental evaluation as ranked on the ratio of interest points that lead to correct matches

Interest points and image descriptors ranked on matching efficiency

Interest points		Scale selection	Descriptor	Efficiency
$\mathcal{D}_{1,norm} L$		link	SIFT	0.8128
$\tilde{\mathcal{D}}_{1,norm} L$		link	SIFT	0.8079
$\det \mathcal{H}_{norm} L$	$(\mathcal{D}_1 L > 0)$	link	SIFT	0.8079
$\tilde{\mathcal{D}}_{2,norm} L$	$(\mathcal{D}_1 L > 0)$	link	SIFT	0.8014
$\det \mathcal{H}_{norm} L$	$(\tilde{\mathcal{D}}_1 L > 0)$	link	SIFT	0.8002
$\vdots$				$\vdots$
$\det \mathcal{H}_{norm} L$	$(\mathcal{D}_1 L > 0)$	extr	SIFT	0.7721
$\det \mathcal{H}_{norm} L$	$(\mathcal{D}_1 L > 0)$	extr	SURF	0.7656
$\nabla^2_{norm} L$	$(\mathcal{D}_1 L > 0)$	extr	SIFT	0.7484
Harris–Laplace		extr	SIFT	0.7002

For comparison, results are also shown for the SIFT descriptor based on scale-space extrema of the Laplacian, the SIFT or SURF descriptors based on scale-space extrema of the determinant of the Hessian and the SIFT descriptor based on Harris–Laplace interest points

matching properties under the foreshortening transformations that involve transformations outside the similarity group.

10 Extension to Illumination Invariance

The treatment so far has been concerned with the detection of interest points under geometric transformations, modelled as local scaling transformations and local affine image deformation representing the essential dimensions in the variability of a local linearization of the perspective mapping from a surface patch in the world to the image plane.

To obtain theoretically well-founded handling of image data under illumination variations, it is natural to represent the image data on a logarithmic luminosity scale

$$f(x, y) \sim \log I(x, y). \tag{48}$$

Specifically, receptive field responses that are computed from such a logarithmic parameterization of the image luminosities can be *interpreted physically* as a superposition of relative variations of surface structure and illumination variations. Let us assume a (i) perspective camera model extended with (ii) a thin circular lens for gathering incoming light from different directions and (iii) a Lambertian illumination model extended with (iv) a spatially varying albedo factor for modelling the light that is reflects from surface patterns in the world. Then, it can be shown (Lindeberg [106, Sect. 2.3]) that a spatial receptive field response

$$L_{x^\alpha y^\beta}(\cdot, \cdot;\ s) = \partial_{x^\alpha y^\beta}\, \mathcal{T}_s\, f \tag{49}$$

of the image data f, where $\mathcal{T}_s$ represents the spatial smoothing operator (here corresponding to a two-dimensional Gaussian kernel (2)) can be expressed as

$$\begin{aligned} L_{x^\alpha y^\beta} = \partial_{x^\alpha y^\beta}\, \mathcal{T}_s \Big(&\log \rho(x, y) + \log i(x, y) \\ &+ \log C_{cam}(\tilde{f}) + V(x, y)\Big) \end{aligned} \tag{50}$$

where

(i) $\rho(x, y)$ is a spatially dependent *albedo factor* that reflects *properties of surfaces of objects* in the environment with the implicit understanding that this entity may in general refer to points on different surfaces in the world depending on the viewing direction and thus the image position (x, y),

(ii) $i(x, y)$ denotes a spatially dependent *illumination field* with the implicit understanding that the amount of incoming light on different surfaces may be different for different points in the world as mapped to corresponding image coordinates (x, y),

(iii) $C_{cam}(\tilde{f}) = \frac{\pi}{4} \frac{d}{f}$ represents *internal camera parameters* with the ratio $\tilde{f} = f/d$ referred to as the *effective f-number*, where d denotes the diameter of the lens and f the focal distance and

(iv) $V(x, y) = -2 \log(1 + x^2 + y^2)$ represents a geometric *natural vignetting* effect corresponding to the factor $\log \cos^4(\phi)$ for a planar image plane, with ϕ denoting the angle between the viewing direction (x, y, f) and the surface normal $(0, 0, 1)$ of the image plane. This vignetting term disappears for a spherical camera model.

From the structure of Eq. (50) we can note that for any non-zero order of differentiation $\alpha > 0$ or $\beta > 0$, the influence of the internal camera parameters in $C_{cam}(\tilde{f})$ will disappear because of the spatial differentiation with respect to x or y, and so will the effects of any other multiplicative exposure control mechanism. Furthermore, for any multiplicative illumination variation $i'(x, y) = C\, i(x, y)$, where C is a scalar constant, the logarithmic luminosity will be transformed as $\log i'(x, y) = \log C + \log i(x, y)$, which implies that the dependency on C will disappear after spatial differentiation.

After a logarithmic transformation of the intensity axis, the Gaussian derivatives that we use for defining the interest point detectors in Sect. 4 will therefore be *invariant under local multiplicative illumination variations and exposure control mechanisms* and thus also the responses of the interest point detectors. Figure 11 gives an illustration of this property by showing interest points detected from a scene with a build-

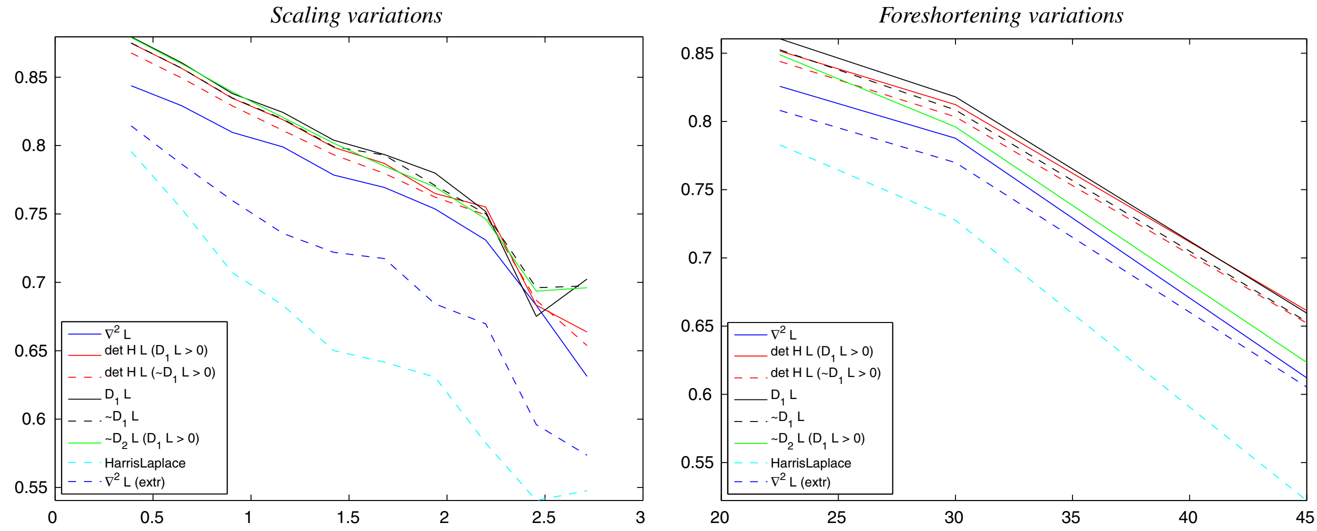


Fig. 10 Graphs showing how the matching efficiency depends upon (*left*) the amount of scaling $s \in [1.25, 6.0]$ for the perspective scaling transformations (with $\log_2 s$ on the horizontal axis) and (*right*) the difference in viewing angle $\varphi \in [22.5°, 45°]$ for the perspective foreshortening transformations for interest point matching based on SIFT-like Gauss-SIFT image descriptors. (The reason why the curve showing the matching efficiency under scaling variations is more jaggy in the rightmost part is that much fewer interest points are used for larger scale factors (N/s^2) thereby affecting the statistical determinacy in the results.)

ing where one wall is strongly sunlit, whereas another wall is in the shadow. When using a linear parameterization of the intensity values as obtained from the camera (which can be assumed to represent a gamma transformation I^γ of local energy measurements I), a dominance of the strongest interest points is obtained from the sunlit parts in the scene. When using a logarithmic transformation of the brightness values (which by the assumption of camera measurements using a gamma transformation can be assumed to represent a logarithmic transformation of local energy measurements $\gamma \log I$), we obtain much more responses from regions in the shadow.

The logarithmic transformation prior to the computation of a scale-space representation and interest point detectors based on Gaussian derivatives does therefore compensate for the subclass of illumination variations that can be modelled by local multiplicative intensity transformations within the support region of the underlying receptive fields that are used for computing the image features.

For this building we could not expect perfectly equal responses from the two walls, since the local 3-D geometry differs somewhat between the walls. The important point, however, is that the invariance of receptive field responses under local multiplicative illumination variations implies that the responses from the interest point detectors will not be affected by the difference in local image contrast that would otherwise be the result on a linear brightness scale.

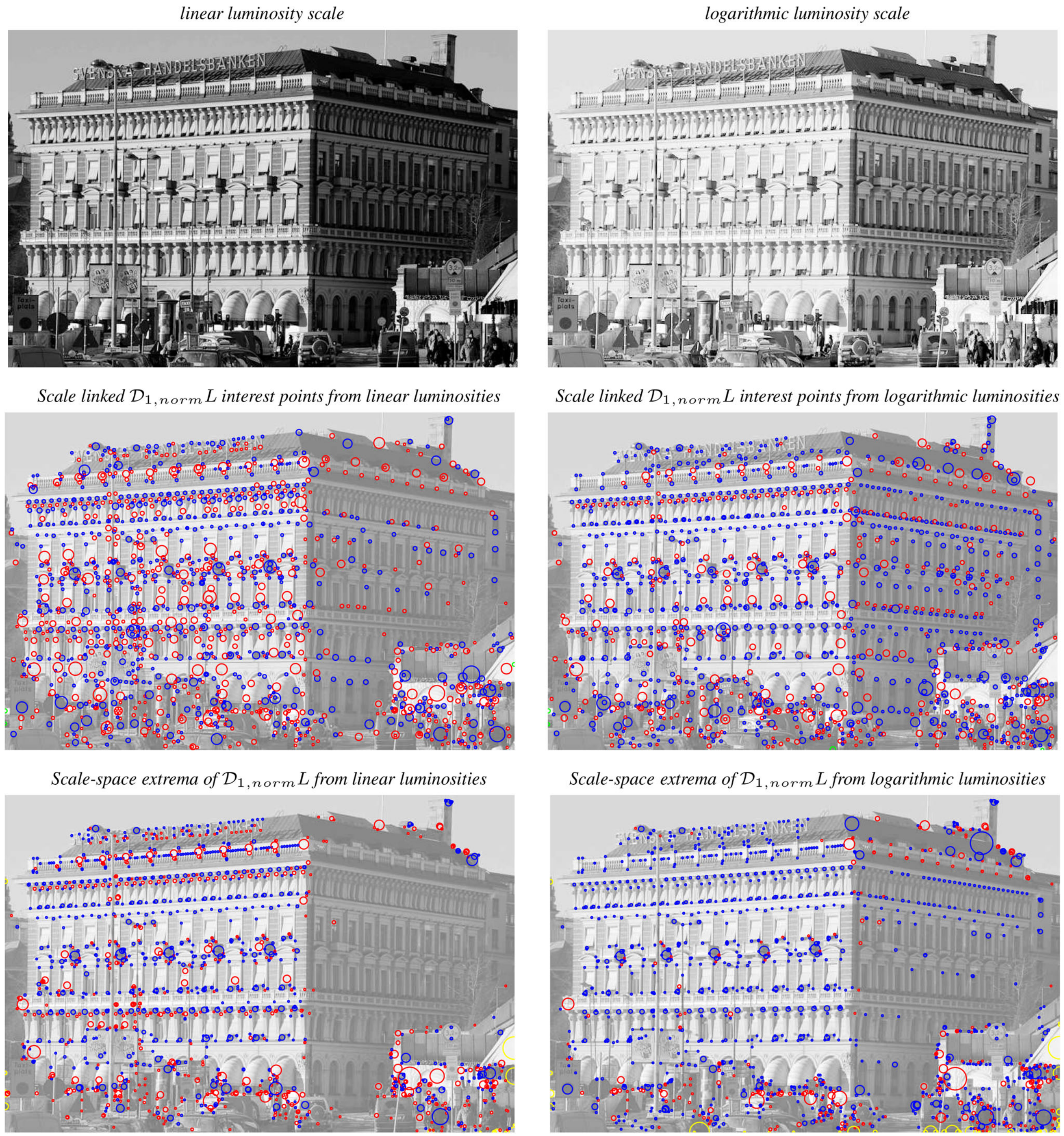
The computation of receptive field responses in terms of spatial derivates over a logarithmic brightness scale does therefore lead to an automatic compensation for illumination variations that can be modelled as local multiplicative intensity transformations.

For the purely second-order differential entities $\nabla^2_{norm} L$, $\det \mathcal{H}_{norm} L$, $\mathcal{D}_{1,norm} L$, $\tilde{\mathcal{D}}_{1,norm} L$, $\mathcal{D}_{2,norm} L$ and $\tilde{\mathcal{D}}_{2,norm} L$, the differential invariants will also be invariant to local linear illumination gradients of the form

$$f(x, y) \mapsto f(x, y) + A(x - x_0) + B(y - y_0). \tag{51}$$

If we consider local surface markings on a curved object (by a painted surface assumptions) and model the local illumination alternative reflectance variations by illuminating the object from two different directions relative to an object centered frame, alternatively observing the object from two different viewing directions for a non-Lambertian reflectance model, we could therefore expect the responses of the interest point detectors to be invariant to the first-order linear component of such illumination or reflectance variations. Thus, these interest point detectors obey basic robustness properties under illumination variations as well as multiplicative exposure control parameters, provided that the interest point detectors are applied to image intensities represented on a logarithmic brightness scale.

Concerning the subsequent computation of image descriptors at the interest points, it follows from a similar way of reasoning that the measurements of the first-order partial derivatives L_x and L_y underlying the SIFT and SURF descriptors will be invariant to local multiplicative illumination transformations or exposure control mechanisms if the image intensities are represented on a logarithmic brightness scale. By



linear luminosity scale

logarithmic luminosity scale

Scale linked $\mathcal{D}_{1,norm}L$ interest points from linear luminosities

Scale linked $\mathcal{D}_{1,norm}L$ interest points from logarithmic luminosities

Scale-space extrema of $\mathcal{D}_{1,norm}L$ from linear luminosities

Scale-space extrema of $\mathcal{D}_{1,norm}L$ from logarithmic luminosities

Fig. 11 Scale-space interest points computed from a scene with strong illumination variations using (*left column*) a linear and (*right column*) a logarithmic parameterization of the luminosity values. Using a linear parameterization of the intensity values, most of the strongest responses are obtained in the sunlit parts and only a few from the shadowed regions, whereas a logarithmic transformation leads to a more similar treatment of the sunlit versus the shadowed regions. This result is a consequence of the invariance of receptive field responses to local multiplicative illumination transformations and corresponding invariance to multiplicative exposure parameters. For each image, the 1,400 strongest responses have been selected, using scale linking of $\mathcal{D}_{1,norm}L$ in the *top row* and scale-space extrema of $\mathcal{D}_{1,norm}L$ in the *bottom row*. The dominance of repetitive image structures on the two walls of the building indirectly also demonstrate the good repeatability properties of these interest point detectors. (Scale range: $t \in [2, 256]$. Image size: 725 × 480 pixels. The size of each *circle* represents the detection scale of the interest point. *Red circles* indicate that the Hessian matrix is negative definite (*bright features*), while *blue circles* that the Hessian matrix is positive definite (*dark features*).)

being defined in terms of first-order derivatives, the SIFT and SURF descriptors are, however, not invariant to linear illumination gradients of the form (51).

11 Summary and Conclusions

We have presented a set of extensions of the SIFT and SURF operators, by replacing the underlying interest point detectors used for computing the SIFT or SURF descriptors by a family of generalized scale-space interest points.

These generalized scale-space interest points are based on (i) new differential entities for interest point detection at a fixed scale in terms of new Hessian feature strength measures, (ii) linking of image structures into feature trajectories over scale and (iii) performing scale selection by either the strongest response of the responses along a feature trajectory, or by weighted averaging of scale-normalized feature responses along each feature trajectory.

The generalized scale-space interest points are all *scale-invariant* in the sense that (i) the interest points are preserved under scaling transformation and that (ii) the detection scales obtained from the scale selection step are transformed in a scale covariant way. Thereby, the detection scale can be used for defining a local scale normalized reference frame around the interest point [109,111] implying that image descriptors defined relative to such a scale-normalized reference frame will also be provably scale invariant.

By complementing the generalized scale-space interest points with local image descriptors defined in a conceptually similar way as the pure image descriptor parts in regular SIFT or SURF, while being based on image measurements in terms of Gaussian derivatives instead of image pyramids or Haar wavelets, we have shown that the generalized interest points with their associated scale-invariant image descriptors lead to a higher ratio of correct matches and a lower ratio of false matches compared to corresponding results obtained with interest point detectors based on more traditional scale-space extrema of the Laplacian, its difference-of-Gaussians approximation or the Harris–Laplace operator.

In the literature, there has been some debate concerning which one of the SIFT or SURF descriptors leads to the best performance. In our experimental evaluations, we have throughout found that our SIFT-like Gauss-SIFT descriptor based on Gaussian derivatives generally performs much better than our SURF-like Gauss-SURF descriptor, also expressed in terms of Gaussian derivatives. In this respect, the pure image descriptor in the regular SIFT operator can be seen as better than the pure image descriptor in the regular SURF operator, and we can in this respect regard the underlying information content in the SIFT descriptor as allowing for more accurate image matching than the information content underlying the SURF descriptor.

Concerning the underlying interest points, we have on the other hand found that the determinant of the Hessian operator to generally perform better than the Laplacian operator, for both scale-space extrema detection and feature detection by scale linking. Since the difference-of-Gaussians interest point detector in the regular SIFT operator can be seen as an approximation of the scale-normalized Laplacian (see Appendix A), we can therefore regard the underlying interest point detector in the SURF operator as better than the interest point detector in the SIFT operator. Specifically, we could expect an increase in the performance of SIFT by replacing the scale-space extrema of the difference-of-Gaussians operator by scale-space extrema of the determinant of the Hessian.

In addition, the experimental evaluation shows that further improvements are possible by replacing the interest points obtained from scale-space extrema in our Gauss-SIFT and Gauss-SURF operators by generalized scale-space interest points obtained by scale linking, with the best results obtained with the Hessian feature strength measures $\mathcal{D}_{1,norm} L$ and $\tilde{\mathcal{D}}_{1,norm} L$ followed by the determinant of the Hessian $\det \mathcal{H}_{norm} L$ and the Hessian feature strength measure $\tilde{\mathcal{D}}_{2,norm} L$ with complementary thresholding on $\mathcal{D}_{1,norm} L > 0$.

These relative relations between the different differential interest points are good agreement with the theoretical analysis of covariance properties of the underlying differential expressions in Sect. 5 and previous results concerning robustness of scale estimates under affine image deformations [110]. Hence, this demonstrates how the experimental performance of interest point detectors defined within the scale-space framework can be predicted from differential geometric analysis of the underlying differential expressions.

12 Discussion

An overall aim with this work to demonstrate the possibility of using a richer vocabulary of interest point detectors for image-based matching and recognition, beyond Laplacian, difference-of-Gaussians or Harris/Harris–Laplace points, which are the most commonly used features today.

Regarding the choice of differential entities, we have presented both theoretical and experimental support advocating the use of affine covariant differential entities or approximations thereof. Concerning the selection of significant image features, we have advocated for including the behaviour of image structures over scale in significance measures of feature strength, specifically by integrating local evidence over their lifetime across scales. On a poster dataset and for systematic experiments over synthetic affine image deformations (not reported here), this mechanism has been demonstrated to lead to better selection of interest points.

Further work would should however be performed to explore these properties experimentally, preferably on more extensive 3-D datasets from natural scenes, for which however the definition of a proper ground truth may constitute a challenge by itself. In situations with occlusions or other true 3-D effects, it should specifically be investigated if scale selection from local extrema over scale along each feature trajectory is preferable over scale selection by weighted averaging over scale, by being more local and less sensitive to interference with neighbouring image structures. Concerning the definition of a significance measure of interest points, there are also other degrees of freedoms to explore in how feature evidence should be accumulated over scale *e.g.* by statistical measures while respecting scale invariance.

From such a context, the proposed framework for generalized scale-space interest points should be seen as defining a theoretical structure by which richer sets of interest point detectors can be considered and be specifically adapted to different computer vision applications, with additional degrees of freedoms to explore regarding (i) the choice of differential entities for interest point detection, (ii) scale selection mechanisms and (iii) ways of ranking interest points on significance to enable automatic selection of repeatable subsets of sparse interest points for applications in which the number of interest points must be kept low because of the computational complexity of later stage processes.

These generalized scale-space interest point detectors can also be complemented by affine shape adaptation to enable affine invariant interest points and image descriptors. Specifically, we could expect that by initiating the affine shape adaptation process from determinant of the Hessian $\det \mathcal{H}_{norm} L$ or Hessian feature strength measures $\mathcal{D}_{1,norm} L$ or $\tilde{\mathcal{D}}_{1,norm} L$ interest points should make it possible to handle image deformations outside the similarity group in a better manner than *e.g.* Laplacian $\nabla^2_{norm} L$, difference-of-Gaussians or Harris–Laplace interest points.

More generally, we do in a similar way as in [90, 101] argue that qualitative scale information extracted in a bottom-up processing stage, as done by scale invariant feature detection and/or a scale-space primal sketch, may serve as a guide to other visual processing stages and may simplify their tasks. For example, the scale tuning of other early visual processing at scale levels proportional to the detection scale of scale-invariant image features constitutes one such domain of applications [111]. Since all the interest point detectors proposed in this work are scale invariant, it follows that the associated scale estimates obtained from these can be used for normalizing other visual operations with respect to scale or size variations, and that the corresponding derived visual representations will therefore also be scale invariant.

In a similar way as the SIFT descriptor has been extended to colour images by several authors (Bosch et al. [17], van de Weijer and Schmid [155], Burghouts and Geusebroek [24], van de Sande et al. [138]), we propose that the generalized interest points presented here can be integrated with colour extensions of the SIFT descriptor or other image descriptors to increase their discriminative properties.

Acknowledgments An earlier version of this work was presented at the SSVM 2013 conference [108]. I would like to thank Lars Bretzner for his help when preparing the poster image dataset and Oskar Linde for sharing his code for local image descriptors. The support from the Swedish Research Council (contract 2010-4766) and from the Royal Swedish Academy of Sciences as well as the Knut and Alice Wallenberg Foundation is gratefully acknowledged

Appendix: A Relationship Between Laplacian and difference-of-Gaussians Interest Points

Since the difference-of-Gaussians interest point detector in the regular SIFT operator (Lowe [119]) can be seen as an approximation of the scale-normalized Laplacian (Lindeberg [101])

$$\frac{1}{2}\nabla^2 L(x;\ t) = \partial_t L(x;\ t) \approx \frac{L(x;\ t+\Delta t)-L(x;\ t)}{\Delta t} = \frac{DOG(x;\ t,\Delta t)}{\Delta t} \tag{52}$$

with $\Delta t = (k^2 - 1)\,t$ due to the self-similar scale sampling $\sigma_{i+1} = k\,\sigma_i$ corresponding to $t_{i+1} = k^2\,t_i$, thus implying

$$DOG(x, y; t) \approx \frac{(k^2 - 1)}{2}\,\nabla^2_{norm} L(x, y; t), \tag{53}$$

we can regard interest points obtained from scale-space extrema of difference-of-Gaussians as approximations of interest points obtained from scale-space extrema of the Laplacian.

References

1. Aanaes, H., Lindbjerg-Dahl, A., Pedersen, K.S.: Interesting interest points: a comparative study of interest point performance on a unique data set. Int. J. Comput. Vis. **97**(1), 18–35 (2012)
2. Agarwal, A., Triggs, B.: Multilevel image coding with hyperfeatures. Int. J. Comput. Vis. **78**(1), 15–27 (2008)
3. Almansa, A., Lindeberg, T.: Fingerprint enhancement by shape adaptation of scale-space operators with automatic scale-selection. IEEE Trans. Image Process. **9**(12), 2027–2042 (2000)

4. Balmashnova, E., Florack, L.M.J.: Novel similarity measures for differential invariant descriptors for generic object retrieval. J. Math. Imaging Vis. **31**(2–3), 121–132 (2008)
5. Balmashnova, E.G., Platel, B., Florack, L., ter Haar Romeny, B.M.: Object matching in the presence of non-rigid deformations close to similarities. In: Proceedings of International Conference on Computer Vision (ICCV 2007), pp. 2591–2598. Rio de Janeiro, Brazil (2007)
6. Baumberg, A.: Reliable feature matching across widely separated views. In: Proceedings of Computer Vision and Pattern Recognition (CVPR'00), pp. I:1774–1781. Hilton Head, SC (2000)
7. Bay, H., Ess, A., Tuytelaars, T., van Gool, L.: Speeded up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
8. Bay, H., Tuytelaars, T., van Gool, L.: SURF: speeded up robust features. In: Proceedings European Conference on Computer Vision (ECCV 2006), Lecture Notes in Computer Science, vol. 3951, pp. I:404–417. Springer, Graz, Austria (2006)
9. Beaudet, P.R.: Rotationally invariant image operators. In: Proceedings of 4th International Joint Conference on Pattern Recognition, pp. 579–583. Tokyo, Japan (1978)
10. Belongie, S., Carson, C., Greenspan, H., Malik, J.: Color- and texture-based image segmentation using EM and its application to content-based image retrieval. In: Proceedings of International Conference on Computer Vision (ICCV'98), pp. 675–682. Bombay, India (1998)
11. Benhimane, S., Malis, E.: Real-time image-based tracking of planes using efficient second-order minimization. In: Intelligent Robots and Systems (IROS 2004), pp. 943–948 (2004)
12. Bigun, J.: Vision with Direction. Springer, Berlin (2006)
13. Bigün, J., Granlund, G.H.: Optimal orientation detection of linear symmetry. In: Proceedings of 1st International Conference on Computer Vision (ICCV'87), pp. 433–438. London (1987)
14. Blom, J.: Topological and geometrical aspects of image structure. Ph.D. thesis, Dept. Med. Phys. Physics, Univ. Utrecht, NL-3508 Utrecht, Netherlands (1992)
15. Blostein, D., Ahuja, N.: A multiscale region detector. Comput. Vis. Graph. Image Process. **45**, 22–41 (1989)
16. Blostein, D., Ahuja, N.: Shape from texture: integrating texture element extraction and surface estimation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(12), 1233–1251 (1989)
17. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: Proceedings of European Conference on Computer Vision (ECCV 2006), Lecture Notes in Computer Science, vol. 3954, pp. 517–530. Springer (2006)
18. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: Proceedings of International Conference on Computer Vision (ICCV 2007), pp. 1–8. Rio de Janeiro, Brazil (2007)
19. Bretzner, L., Laptev, I., Lindeberg, T.: Hand-gesture recognition using multi-scale colour features, hierarchical features and particle filtering. In: Proceedings of Face and Gesture, pp. 63–74. Washington DC, USA (2002)
20. Bretzner, L., Laptev, I., Lindeberg, T., Lenman, S., Sundblad, Y.: A prototype system for computer vision based human computer interaction. Report, ISRN KTH/NA/P–01/09–SE, Dept. of Numerical Analysis and Computing Science, KTH (2001)
21. Bretzner, L., Lindeberg, T.: Feature tracking with automatic selection of spatial scales. Comput. Vis. Image Underst. **71**(3), 385–392 (1998)
22. Bretzner, L., Lindeberg, T.: Qualitative multi-scale feature hierarchies for object tracking. J. Vis. Commun. Image Represent. **11**, 115–129 (2000)
23. Brunnström, K., Lindeberg, T., Eklundh, J.O.: Active detection and classification of junctions by foveation with a head-eye system guided by the scale-space primal sketch. In: Sandini, G. (ed.) Proceedings of European Conference on Computer Vision (ECCV'92), Lecture Notes in Computer Science, vol. 588, pp. 701–709. Springer, Santa Margherita Ligure, Italy (1992)
24. Burghouts, G.J., Geusebroek, J.M.: Performance evaluation of local colour invariants. Comput. Vis. Image Underst. **113**(1), 48–62 (2009)
25. Cachia, A., Mangin, J.F., Riviere, D., Kherif, F., Boddaert, N., Andrade, A., Papadopoulos-Orfanos, D., Poline, J.B., Bloch, I., Zilbovicius, M., Sonigo, P., Brunelle, F., Regis, J.: A primal sketch of the cortex mean curvature: a morphogenesis based approach to study the variability of the folding patterns. IEEE Trans. Med. Imaging **22**(6), 754–765 (2003)
26. Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P.: BRIEF: computing a local binary descriptor very fast. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1281–1298 (2012)
27. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Pattern Anal. Mach. Intell. **24**(8), 1026–1038 (2002)
28. Chomat, O., de Verdiere, V., Hall, D., Crowley, J.: Local scale selection for Gaussian based description techniques. In: Proceedings European Conference on Computer Vision (ECCV 2000), Lecture Notes in Computer Science, vol. 1842, pp. I:117–133. Springer-Verlag, Dublin, Ireland (2000)
29. Coulon, O., Mangin, J.F., Poline, J.B., Zilbovicius, M., Roumenov, D., Samson, Y., Frouin, V., Bloch, I.: Structural group analysis of functional activation maps. NeuroImage **11**(6), 767–782 (2000)
30. Crowley, J., Riff, O.: Fast computation of scale normalised receptive fields. In: Griffin, L., Lillholm, M. (eds.) Proceedings Scale-Space Methods in Computer Vision (Scale-Space'03), Lecture Notes in Computer Science, vol. 2695, pp. 584–598. Springer, Isle of Skye, Scotland (2003)
31. Crowley, J.L., Parker, A.C.: A representation for shape based on peaks and ridges in the difference of low-pass transform. IEEE Trans. Pattern Anal. Mach. Intell. **6**(2), 156–170 (1984)
32. Crowley, J.L., Sanderson, A.C.: Multiple resolution representation and probabilistic matching of 2-D gray-scale shape. IEEE Trans. Pattern Anal. Mach. Intell. **9**(1), 113–121 (1987)
33. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision. Prague, Czech Republik (2004)
34. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of Computer Vision and Pattern Recognition vol. 1, pp. 886–893 (2005)
35. Damon, J.: Local Morse theory for solutions to the heat equation and Gaussian blurring. J. Differ. Equ. **115**(2), 386–401 (1995)
36. Daniilidis, K., Eklundh, J.O.: 3-D vision and recognition. In: Siciliano, B., Khatib, O. (eds.) Springer Handbook of Robotics, pp. 543–562. Springer, Berlin (2008)
37. Demirci, M.F., Platel, B., Shokoufandeh, A., Florack, L., Dickinson, S.J.: The representation and matching of images using top points. J. Math. Imaging Vis. **35**(2), 103–116 (2009)
38. Demirci, M.F., Shokoufandeh, A., Keselman, Y., Bretzner, L., Dickinson, S.: Object recognition as many-to-many feature matching. Int. J. Comput. Vis. **69**(2), 203–222 (2006)
39. Deriche, R., Giraudon, G.: Accurate corner detection: an analytical study. In: Proceedings of International Conference on Computer Vision (ICCV'90), pp. 66–70. Osaka, Japan (1990)
40. Dreschler, L., Nagel, H.H.: Volumetric model and 3D-trajectory of a moving car derived from monocular TV-frame sequences of a street scene. Comput. Vis. Graph. Image Process. **20**(3), 199–228 (1982)
41. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proceedings of Com-

puter Vision and Pattern Recognition (CVPR'03), pp. 264–271. Madison, Wisconsin (2003)

42. Fergus, R., Perona, P., Zisserman, A.: Weakly supervised scale-invariant learning of models for visual recognition. Int. J. Comput. Vis. **71**(3), 273–303 (2007)
43. Florack, L.M.J.: Image Structure. Series in Mathematical Imaging and Vision. Springer, Berlin (1997)
44. Förstner, W.: Statistische Verfahren für die automatische Bildanalyse und ihre Bewertung bei der Objekterkennung und -vermessung. Habilitation thesis, Universität Stuttgart (1991)
45. Förstner, W.A., Gülch, E.: A fast operator for detection and precise location of distinct points, corners and centers of circular features. In: Proceedings Intercommission Workshop of the International Society for Photogrammetry and Remote Sensing. Interlaken, Switzerland (1987)
46. Frangi, A.F., NW, J., Hoogeveen, R.M., van Walsum, T., Viergever, M.A.: Model-based quantitation of 3D magnetic resonance angiographic images. IEEE Trans. Med. Imaging **18**(10), 946–956 (2000)
47. Gårding, J., Lindeberg, T.: Direct computation of shape cues using scale-adapted spatial derivative operators. Int. J. Comput. Vis. **17**(2), 163–191 (1996)
48. Gauch, J.M., Pizer, S.M.: Multiresolution analysis of ridges and valleys in grey-scale images. IEEE Trans. Pattern Anal. Mach. Intell. **15**(6), 635–646 (1993)
49. Geusebroek, J.M., van den Boomgaard, R., Smeulders, A.W.M., Geerts, H.: Color invariance. IEEE Trans. Pattern Anal. Mach. Intell. **23**(12), 1338–1350 (2001)
50. Gevers, T., Smeulders, A.W.M.: Color-based object recognition. Pattern Recognit. Lett. **32**, 453–464 (1999)
51. Granlund, G.H., Knutsson, H.: Signal Processing in Computer Vision. Springer, Dordrecht (1995)
52. Gu, S., Zheng, Y., Tomasi, C.: Critical nets and beta-stable features for image matching. In: Proceedings of European Conference on Computer Vision (ECCV 2010), Lecture Notes in Computer Science, vol. 6313, pp. 663–676. Springer (2010)
53. ter Haar Romeny, B.: Front-End Vision and Multi-Scale Image Analysis. Springer, Berlin (2003)
54. Hall, D., de Verdiere, V., Crowley, J.: Object recognition using coloured receptive fields. In: Proceedings of European Conference on Computer Vision (ECCV 2000), Lecture Notes in Computer Science, vol. 1842, pp. I:164–177. Springer, Dublin, Ireland (2000)
55. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, pp. 147–152 (1988)
56. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, vol. 1. Cambridge University Press, New York (2000)
57. Iijima, T.: Observation theory of two-dimensional visual patterns. Technical report,Papers of Technical Group on Automata and Automatic Control, IECE, Japan (1962)
58. Jähne, B.: Spatio-Temporal Image Processing-Theory and Scientific Applications. No. 751 in Lecture Notes in Computer Science. Springer, Berlin (1993)
59. Jiang, Y.G., Ngo, C.W., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. In: Proceedings of 6th ACM International Conference on Image and Video Retrieval, pp. 494–501. Amsterdam, The Netherlands (2007)
60. Johansen, P.: On the classification of toppoints in scale space. J. Math. Imaging Vis. **4**, 57–67 (1994)
61. Johansen, P., Skelboe, S., Grue, K., Andersen, J.D.: Representing signals by their top points in scale-space. In: Proceedings of 8th International Confernece on Pattern Recognition, pp. 215–217. Paris, France (1986)
62. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: Proceedings International Conference on Computer Vision (ICCV 2005), vol. 1, pp. 17–21. Beijing, China (2005)
63. Kadir, T., Brady, M.: Saliency, scale and image description. Int. J. Comput. Vis. **45**(2), 83–105 (2001)
64. Kadir, T., Zisserman, A., Brady, M.: An affine invariant salient region detector. In: Proc. European Conf. on Computer Vision (ECCV 2004), Lecture Notes in Computer Science, vol. 3021, pp. I:228–241. Springer, Prague, Czech Republik (2004)
65. Kaneva, B., Torralba, A., Freeman, W.T.: Evaluating image features using a photorealistic world. In: Proceedings of International Conference on Computer Vision (ICCV 2011), pp. 172–177. Barcelona, Spain (2011)
66. Ke, Y., Sukthankar, R.: PCA-SIFT: a more distinctive representation for local image descriptors. In: Proceedings Computer Vision and Pattern Recognition, pp. II: 506–513. Washington D. C. (2004)
67. Kirbas, C., Quek, F.: A review of vessel extraction techniques and algorithms. ACM Comput. Surv. **36**(2), 81–121 (2004)
68. Kitchen, L., Rosenfeld, A.: Gray-level corner detection. Pattern Recognit. Lett. **1**(2), 95–102 (1982)
69. Koenderink, J.J.: The structure of images. Biol. Cybern. **50**, 363–370 (1984)
70. Koenderink, J.J., Richards, W.: Two-dimensional curvature operators. J. Opt. Soc. Am. **5**(7), 1136–1141 (1988)
71. Koenderink, J.J., van Doorn, A.J.: Dynamic shape. Biol. Cybern. **53**, 383–396 (1986)
72. Koenderink, J.J., van Doorn, A.J.: Representation of local geometry in the visual system. Biol. Cybern. **55**, 367–375 (1987)
73. Koenderink, J.J., van Doorn, A.J.: Generic neighborhood operators. IEEE Trans. Pattern Anal. Mach. Intell. **14**(6), 597–605 (1992)
74. Kokkinos, I., Maragos, P., Yuille, A.: Bottom-up & top-down object detection using primal sketch features and graphical models. In: Proceedings of Computer Vision and Pattern Recognition (CVPR'06), pp. II: 1893–1900. New York (2006)
75. Kokkinos, I., Yuille, A.: Scale invariance without scale selection. In: Proceedings ofComputer Vision and Pattern Recognition (CVPR'08), pp. 1–8 (2008)
76. Krissian, K., Malandain, G., Ayache, N., Vaillant, R., Trousset, Y.: Model-based detection of tubular structures in 3D images. Comput. Vis. Image Underst. **80**(2), 130–171 (2000)
77. Kuijper, A., Florack, L.: Calculations on critical points under gaussian blurring. In: Proceedings of International Conference on Scale-Space Theories in Computer Vision (Scale-Space'99), Lecture Notes in Computer Science, vol. 1682, pp. 318–329. Springer, Corfu, Greece (1999)
78. Kuijper, A., Florack, L.: Using catastrophe theory to derive trees from images. J. Math. Imaging Vis. **23**(3), 219–238 (2005)
79. Laptev, I., Lindeberg, T.: Tracking of multi-state hand models using particle filtering and a hierarchy of multi-scale image features. In: Kerckhove, M. (ed.) Proceedings of International Conference on Scale-Space and Morphology in Computer Vision (Scale-Space'01), Lecture Notes in Computer Science, vol. 2106, pp. 63–74. Springer, Vancouver, Canada (2001)
80. Laptev, I., Lindeberg, T.: A distance measure and a feature likelihood map concept for scale-invariant model matching. Int. J. Comput. Vis. **52**, 97–120 (2003)
81. Larsen, A.B.L., Darkner, S., Dahl, A.L., Pedersen, K.S.: Jet-based local image descriptors. In: Proceedings of European Conference on Computer Vision (ECCV 2012), Lecture Notes in Computer Science, vol. 7574, pp. III:638–650. Springer (2012)
82. Lazebnik, S., Schmid, C., Ponce, J.: Semi-local affine parts for object recognition. In: Proceedings of British Machine Vision Conference on Kingston, UK (2004)

83. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1265–1278 (2005)
84. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of Computer Vision and Pattern Recognition (CVPR'06), pp. 2169–2178. Washington, DC, USA (2006)
85. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: state of the art and challenges. ACM Trans. Multimed. Comput. Commun. Appl. **2**(1), 1–19 (2006)
86. Lifshitz, L., Pizer, S.: A multiresolution hierarchical approach to image segmentation based on intensity extrema. IEEE Trans. Pattern Anal. Mach. Intell. **12**(6), 529–541 (1990)
87. Linde, O., Lindeberg, T.: Object recognition using composed receptive field histograms of higher dimensionality. In: International Conference on Pattern Recognition, vol. 2, pp. 1–6. Cambridge (2004)
88. Linde, O., Lindeberg, T.: Composed complex-cue histograms: an investigation of the information content in receptive field based image descriptors for object recognition. Comput. Vis. Image Underst. **116**, 538–560 (2012)
89. Lindeberg, T.: Scale-space behaviour of local extrema and blobs. J. Math. Imaging Vis. **1**(1), 65–99 (1992)
90. Lindeberg, T.: Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention. Int. J. Comput. Vis. **11**(3), 283–318 (1993)
91. Lindeberg, T.: Discrete derivative approximations with scale-space properties: a basis for low-level feature extraction. J. Math. Imaging Vis. **3**(4), 349–376 (1993)
92. Lindeberg, T.: Effective scale: a natural unit for measuring scale-space lifetime. IEEE Trans. Pattern Anal. Mach. Intell. **15**(10), 1068–1074 (1993)
93. Lindeberg, T.: On scale selection for differential operators. In: Proceedings of 8th Scandinavian Conference on Image Analysis (SCIA'93), pp. 857–866. Norwegian Society for Image Processing and Pattern Recognition, Tromsø Norway (1993)
94. Lindeberg, T.: Scale-space theory: a basic tool for analysing structures at different scales. J. Appl. Stat. **21**(2), 225–270 (1994). Also available from http://www.csc.kth.se/~tony/abstracts/Lin94-SI-abstract.html
95. Lindeberg, T.: Scale-Space Theory in Computer Vision. Springer, Berlin (1994)
96. Lindeberg, T.: Direct estimation of affine deformations of brightness patterns using visual front-end operators with automatic scale selection. In: Proceedings of International Conference on Computer Vision (ICCV'95), pp. 134–141. Cambridge, MA (1995)
97. Lindeberg, T.: Edge detection and ridge detection with automatic scale selection. In: Proceedings of Computer Vision and Pattern Recognition, 1996, pp. 465–470. San Francisco, California (1996)
98. Lindeberg, T.: Scale-space theory: a framework for handling image structures at multiple scales. In: Proceedings of CERN School of Computing, Technical Report CERN 96–08, pp. 27–38. Egmond aan Zee, The Netherlands (1996). Also available from http://www.csc.kth.se/cvap/abstracts/lin96-csc.html
99. Lindeberg, T.: On automatic selection of temporal scales in time-casual scale-space. In: Sommer, G., Koenderink, J.J. (eds.) Proceedings of AFPAC'97: Algebraic Frames for the Perception-Action Cycle, Lecture Notes in Computer Science, vol. 1315, pp. 94–113. Springer, Kiel, Germany (1997)
100. Lindeberg, T.: Edge detection and ridge detection with automatic scale selection. Int. J. Comput. Vis. **30**(2), 117–154 (1998)
101. Lindeberg, T.: Feature detection with automatic scale selection. Int. J. Comput. Vis. **30**(2), 77–116 (1998)
102. Lindeberg, T.: Principles for automatic scale selection. In: Handbook on Computer Vision and Applications, pp. 239–274. Academic Press, Boston (1999). Also available from http://www.csc.kth.se/cvap/abstracts/cvap222.html
103. Lindeberg, T.: Scale-space. In: Wah, B. (ed.) Encyclopedia of Computer Science and Engineering, pp. 2495–2504. Wiley, Hoboken (2008)
104. Lindeberg, T.: Generalized scale-space interest points: scale-space primal sketch for differential descriptors (2010). Int. J. Comput. Vis.
105. Lindeberg, T.: Generalized Gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space. J. Math. Imaging Vis. **40**(1), 36–81 (2011)
106. Lindeberg, T.: A computational theory of visual receptive fields. Biol. Cybern. **107**(6), 589–635 (2013)
107. Lindeberg, T.: Generalized axiomatic scale-space theory. In: Hawkes, P. (ed.) Advances in Imaging and Electron Physics, vol. 178, pp. 1–96. Elsevier, Amsterdam (2013)
108. Lindeberg, T.: Image matching using generalized scale-space interest points. In: Proceedings of International Conference on Scale-Space and Variational Methods for Computer Vision (SSVM 2013), Lecture Notes in Computer Science, vol. 7893, pp. 355–367. Springer (2013)
109. Lindeberg, T.: Invariance of visual operations at the level of receptive fields. PLOS One **8**(7), e66,990 (2013)
110. Lindeberg, T.: Scale selection properties of generalized scale-space interest point detectors. J. Math. Imaging Vis. **46**(2), 177–210 (2013)
111. Lindeberg, T.: Scale selection. In: Ikeuchi, K. (ed.) Computer Vision: A Reference Guide, pp. 701–713. Springer, New York (2014)
112. Lindeberg, T., Bretzner, L.: Real-time scale selection in hybrid multi-scale representations. In: Griffin, L., Lillholm, M. (eds.) Proceedings of Scale-Space Methods in Computer Vision (Scale-Space'03), Lecture Notes in Computer Science, vol. 2695, pp. 148–163. Springer, Isle of Skye, Scotland (2003)
113. Lindeberg, T., Florack, L.: Foveal scale-space and linear increase of receptive field size as a function of eccentricity. report, ISRN KTH/NA/P–94/27–SE, Department of Numerical Analysis and Computing Science, KTH (1994). Available from http://www.csc.kth.se/~tony/abstracts/CVAP166.html
114. Lindeberg, T., Gårding, J.: Shape from texture from a multi-scale perspective. In: Nagel, T.S.H.H.-H., Shirai, Y. (eds.) Proceedings of International Conference on Computer Vision (ICCV'93), pp. 683–691. IEEE Computer Society Press, Berlin, Germany (1993)
115. Lindeberg, T., Gårding, J.: Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D structure. Image Vis. Comput. **15**, 415–434 (1997)
116. Lindeberg, T., Li, M.: Segmentation and classification of edges using minimum description length approximation and complementary junction cues. Comput. Vis. Image Underst. **67**(1), 88–98 (1997)
117. Lindeberg, T., Lidberg, P., Roland, P.: Analysis of brain activation patterns using a 3-D scale-space primal sketch. Hum. Brain Mapp. **7**(3), 166–194 (1999)
118. Lowe, D.: Object recognition from local scale-invariant features. In: Proceedings of International Conference on Computer Vision (ICCV'99), pp. 1150–1157. Corfu, Greece (1999)
119. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
120. Mangin, J.F., Riviere, D., Coulon, O., Poupon, C., Cachia, A., Cointepas, Y., Poline, J.B., Le Bihan, D., Regis, J., Papadopoulos-Orfanos, D.: Coordinate-based versus structural approaches to brain image analysis. Artif. Intell. Med. **30**, 177–197 (2004)
121. Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W.H. Freeman, New York (1982)

122. Marr, D., Hildreth, E.: Theory of edge detection. Proc. Royal Soc. Lond. **207**, 187–217 (1980)
123. Matas, J., Chum, O., Urba, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of British Machine Vision Conference, pp. 384–396 (2002)
124. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. Int. J. Comput. Vis. **60**(1), 63–86 (2004)
125. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005)
126. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., van Gool, L.: A comparison of affine region detectors. Int. J. Comput. Vis. **65**(1–2), 43–72 (2005)
127. Moreels, P., Perona, P.: Evaluation of features detectors and descriptors based on 3D objects. In: Proceedings of International Conference on Computer Vision (ICCV'05), vol. I, pp. 800–807. Beijing, China (2005)
128. Noble, J.A.: Finding corners. Image Vis. Comput. **6**(2), 121–128 (1988)
129. Olsen, O.F.: Multi-scale watershed segmentation. In: Sporring, J., Nielsen, M., Florack, L., Johansen, P. (eds.) Gaussian Scale-Space Theory: Proc. PhD School on Scale-Space Theory, pp. 191–200. Springer, Copenhagen (1997)
130. Opelt, A., Pinz, A., Fussenegger, M., Auer, P.: Generic object recognition with boosting. IEEE Trans. Pattern Anal. Mach. Intell. **28**(3), 416–431 (2005)
131. Pietikäinen, M., Hadid, A., Zhao, G., Ahonen, T.: Computer Vision Using Local Binary Patterns. Springer, Berlin (2011)
132. Pinz, A.: Object categorization. Found. Trends Comput. Graph. Vis. **1**(4), 259–362 (2006)
133. Pizer, S., Joshi, S., Fletcher, T., Styner, M., Tracton, G., Chen, J.: Segmentation of single-figure objects by deformable M-reps. In: Proceedings of 4th International Conference on Medical Image Computing and Computer-Assisted Intervention, Lecture Notes in Computer Science, vol. 2208, pp. 862–871. Springer (2001)
134. Pizer, S.M., Eberly, D., Fritsch, D.S.: Zoom-invariant vision of figural shape: the mathematics of cores. Comput. Vis. Image Underst. **69**(1), 55–71 (1998)
135. Platel, B., Balmashnova, E.G., Florack, L., ter Haar Romeny, B.M.: Top points as interest points for image matching. In: Proceedings of European Conference on Computer Vision (ECCV 2006), vol. 3951, pp. 418–429. Graz, Austria (2006)
136. Rosbacke, M., Roland, P.E., Lindeberg, T.: Evaluation of using absolute vs. relative base level when analyzing brain activation images using the scale-space primal sketch. J. Med. Image Anal. **5**(2), 89–110 (2001)
137. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. Int. J. Comput. Vis. **66**(3), 231–259 (2006)
138. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1582–1596 (2010)
139. Sato, Y., Nakajima, S., Shiraga, N., Atsumi, H., Yoshida, S., Koller, T., Gerig, G., Kikinis, R.: 3D multi-scale line filter for segmentation and visualization of curvilinear structures in medical images. Med. Image Anal. **2**(2), 143–168 (1998)
140. Schiele, B., Crowley, J.: Recognition without correspondence using multidimensional receptive field histograms. Int. J. Comput. Vis. **36**(1), 31–50 (2000)
141. Schneiderman, H., Kanade, T.: A statistical method for 3D object detection applied to faces and cars. In: Proceedings of Computer Vision and Pattern Recognition (CVPR'00), vol. I, pp. 746–751. Hilton Head, SC (2000)
142. Shi, J., Tomasi, C.: Good features to track. In: Proceeedings of Computer Vision and Pattern Recognition, pp. 593–600 (1994)
143. Shokoufandeh, A., Dickinson, S., Jansson, C., Bretzner, L., Lindeberg, T.: On the representation and matching of qualitative shape at multiple scales. In: Sparr, Heyden, Johansen, Nielsen (eds.) Proceedings European Conference on Computer Vision (ECCV 2002), pp. 759–775. Springer, Copenhagen, Denmark (2002)
144. Shokoufandeh, A., Marsic, I., Dickinson, S.: View-based object recognition using saliency maps. Image Vis. Comput. **17**(5/6), 445–460 (1999)
145. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.: Discovering objects and their location in images. In: Proceedings of Computer Vision and Pattern Recognition (CVPR'05), pp. I: 370–377. San Diego (2005)
146. Slater, D., Healey, G.: Combining colour and geometric information for illumination invariant recognition of 3-D objects. In: Proceedings of International Conference on Computer Vision (ICCV'95), pp. 563–568. Cambridge, MA (1995)
147. Sporring, J., Nielsen, M., Florack, L., Johansen, P. (eds.): Gaussian Scale-Space Theory: Proc. PhD School on Scale-Space Theory. Series in Mathematical Imaging and Vision. Springer, Copenhagen (1996)
148. Swain, M., Ballard, D.: Color indexing. Int. J. Comput. Vis. **7**(1), 11–32 (1991)
149. Toews, M., Wells, W.M.: SIFT-Rank: ordinal descriptors for invariant feature correspondence. In: Proceedings of Computer Vision and Pattern Recognition (CVPR'09), pp. 172–177. Miami, Florida (2009)
150. Tola, E., Lepetit, V., Fua, P.: Daisy: an efficient dense descriptor applied to wide baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 815–830 (2010)
151. Tuytelaars, T., van Gool, L.: Matching widely separated views based on affine invariant regions. Int. J. Comput. Vis. **59**(1), 61–85 (2004)
152. Tuytelaars, T., Mikolajczyk, K.: A survey on local invariant features. Found. Trends Comput. Graph. Vis. **3**(3), 177–280 (2008)
153. Voorhees, H., Poggio, T.: Detecting textons and texture boundaries in natural images. In: Proceedings of 1st International Conference on Computer Vision (ICCV'87). London, England (1987)
154. Weickert, J.: Anisotropic Diffusion in Image Processing. Teubner-Verlag, Stuttgart (1998)
155. van de Weijer, J., Schmid, C.: Coloring local feature extraction. In: Procedings of European Conference on Computer Vision (ECCV 2006), Lecture Notes in Computer Science, pp. 334–348. Springer (2006)
156. Wiltschi, K., Pinz, A., Lindeberg, T.: An automatic assessment scheme for steel quality inspection. Mach. Vis. Appl. **12**, 113–128 (2000)
157. Witkin, A.P.: Scale-space filtering. In: Proceedings of 8th International Joint Conference Artificial Intelligence, pp. 1019–1022. Karlsruhe, Germany (1983)
158. Zhang, J., Barhomi, Y., Serre, T.: A new biologically inspired image descriptor. In: Procedings of European Conference on Computer Vision (ECCV 2012), Lecture Notes in Computer Science, vol. 7576, pp. III:312–324. Springer (2012).

Tony Lindeberg is a Professor of Computer Science at KTH Royal Institute of Technology in Stockholm, Sweden. He was born in Stockholm in 1964, received his M.Sc. degree in 1987, his Ph.D. degree in 1991, became docent in 1996, and was appointed professor in 2000. He was a Research Fellow at the Royal Swedish Academy of Sciences between 2000 and 2010. His research interests in computer vision relate to scale-space representation, image features, object recognition, spatio-temporal recognition, focus-of-attention and computational modelling of biological vision. He has developed theories and methodologies for continuous and discrete scale-space representation, visual and auditory receptive fields, detection of salient image structures, automatic scale selection, scale-invariant image features, affine invariant features, affine and Galilean normalization, temporal, spatio-temporal and spectro-temporal scale-space concepts as well as spatial and spatio-temporal image descriptors for image-based recognition. He has also worked on topics in medical image analysis and gesture recognition. He is author of the book Scale-Space Theory in Computer Vision.