J Math Imaging Vis (2018) 60:525–562
https://doi.org/10.1007/s10851-017-0766-9

Spatio-Temporal Scale Selection in Video Data

Tony Lindeberg[1]

Received: 3 March 2017 / Accepted: 25 September 2017 / Published online: 26 October 2017

Abstract This work presents a theory and methodology for simultaneous detection of local spatial and temporal scales in video data. The underlying idea is that if we process video data by spatio-temporal receptive fields at multiple spatial and temporal scales, we would like to generate hypotheses about the spatial extent and the temporal duration of the underlying spatio-temporal image structures that gave rise to the feature responses. For two types of spatio-temporal scale-space representations, (i) a non-causal Gaussian spatio-temporal scale space for offline analysis of pre-recorded video sequences and (ii) a time-causal and time-recursive spatio-temporal scale space for online analysis of real-time video streams, we express sufficient conditions for spatio-temporal feature detectors in terms of spatio-temporal receptive fields to deliver scale-covariant and scale-invariant feature responses. We present an in-depth theoretical analysis of the scale selection properties of eight types of spatio-temporal interest point detectors in terms of either: (i)–(ii) the spatial Laplacian applied to the first- and second-order temporal derivatives, (iii)–(iv) the determinant of the spatial Hessian applied to the first- and second-order temporal derivatives, (v) the determinant of the spatio-temporal Hessian matrix, (vi) the spatio-temporal Laplacian and (vii)–(viii) the first- and second-order temporal derivatives of the determinant of the spatial Hessian matrix. It is shown that seven of these spatio-temporal feature detectors allow for provable scale covariance and scale invariance. Then, we describe a time-causal and time-recursive algorithm for detecting sparse spatio-temporal interest points from video streams and show that it leads to intuitively reasonable results. An experimental quantification of the accuracy of the spatio-temporal scale estimates and the amount of temporal delay obtained from these spatio-temporal interest point detectors is given, showing that: (i) the spatial and temporal scale selection properties predicted by the continuous theory are well preserved in the discrete implementation and (ii) the spatial Laplacian or the determinant of the spatial Hessian applied to the first- and second-order temporal derivatives leads to much shorter temporal delays in a time-causal implementation compared to the determinant of the spatio-temporal Hessian or the first- and second-order temporal derivatives of the determinant of the spatial Hessian matrix.

Keywords Scale · Scale space · Scale selection · Spatial · Temporal · Spatio-temporal · Scale invariance · Scale covariance · Feature detection · Differential invariant · Interest point · Video analysis · Computer vision

1 Introduction

A basic paradigm for video analysis consists of performing the first layers of visual processing based on successive layers of spatio-temporal receptive fields.

From a mathematical viewpoint, such an approach can be motivated from the fact that the measurement of the image intensity at a single point in space–time does in general not carry any meaningful information, since such a measurement is strongly dependent on external factors, such as the usually

The support from the Swedish Research Council (Contract 2014-4083) and Stiftelsen Olle Engkvist Byggmästare (Contract 2015/465) is gratefully acknowledged.

✉ Tony Lindeberg
tony@kth.se

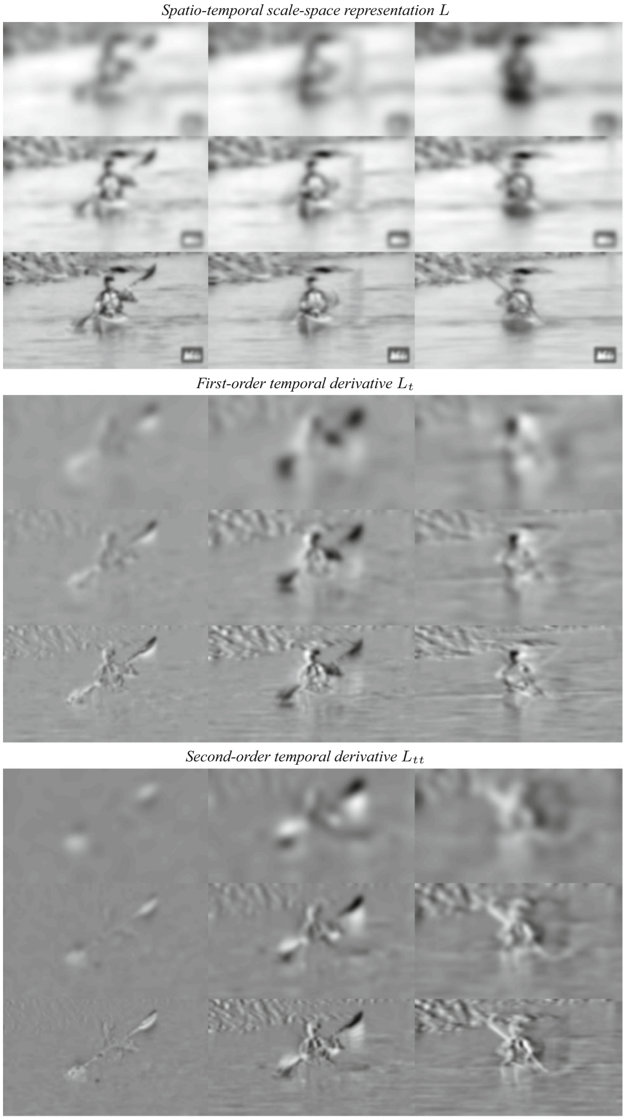
[1] Computational Brain Science Lab, Department of Computational Science and Technology, School of Computer Science and Communication, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden

unknown illumination of the scene. The relevant information is instead carried by the relative relations between the measurements of image intensities at different points over space and time, which implies that it is natural to perform visual processing of video data based on local neighbourhoods over space and time.

From a biological viewpoint, such an approach can also be motivated from the fact that the first layers of mammalian vision can be modelled in terms of spatio-temporal receptive fields over multiple spatial and temporal scales. Cell recordings from neurones in the primary visual cortex have shown that there are spatio-temporal receptive fields tuned to different sizes and orientations in the image domain, to different integration times over the temporal domain as well as to different image velocities in space–time [12,13,32,33]. Interestingly, the shapes of the spatio-temporal receptive field families that have been measured in biological vision can furthermore be explained by normative theories of visual receptive fields [69,71,75,78], whose axiomatic derivation is based on structural properties of the environment in combination with assumptions about the internal structure of an idealized vision system to ensure a consistent treatment of image representations over multiple spatio-temporal scales.

Based on these or related motivations, a large number of computer vision approaches have been developed in which the first layers of image features are computed based on spatio-temporal receptive field responses [3,16,22,35–37,43,48,51,53,93,95,96,98,101–103,108,116–119,121,125].

A general problem when applying the notion of receptive fields in practice, however, is that the types of responses that are obtained in a specific situation can be strongly dependent on the scale levels at which they are computed. Figures 1, 2, 3 and 4 show illustrations of the this problem by showing snapshots of spatio-temporal receptive field responses over multiple spatial and temporal scales for a video sequence and for different types of spatio-temporal features computed from it. Note how qualitatively different types of responses are obtained at different spatio-temporal scales. At some spatio-temporal scales, we get strong responses due to the movements of the paddle or the motion of the paddler in the kayak. At other spatio-temporal scales, we get relatively larger responses because of the movements of the here unstabilized camera. The spatio-temporal texture due to the wave patterns on the water surface does also lead to different type of responses at different spatio-temporal scales. A computer vision system intended to process the visual input from general spatio-temporal scenes does therefore need to decide what responses within the family of spatio-temporal receptive fields over different spatial and temporal scales it should base its analysis on as well as about how the information from different subsets of spatio-temporal scales should be combined.

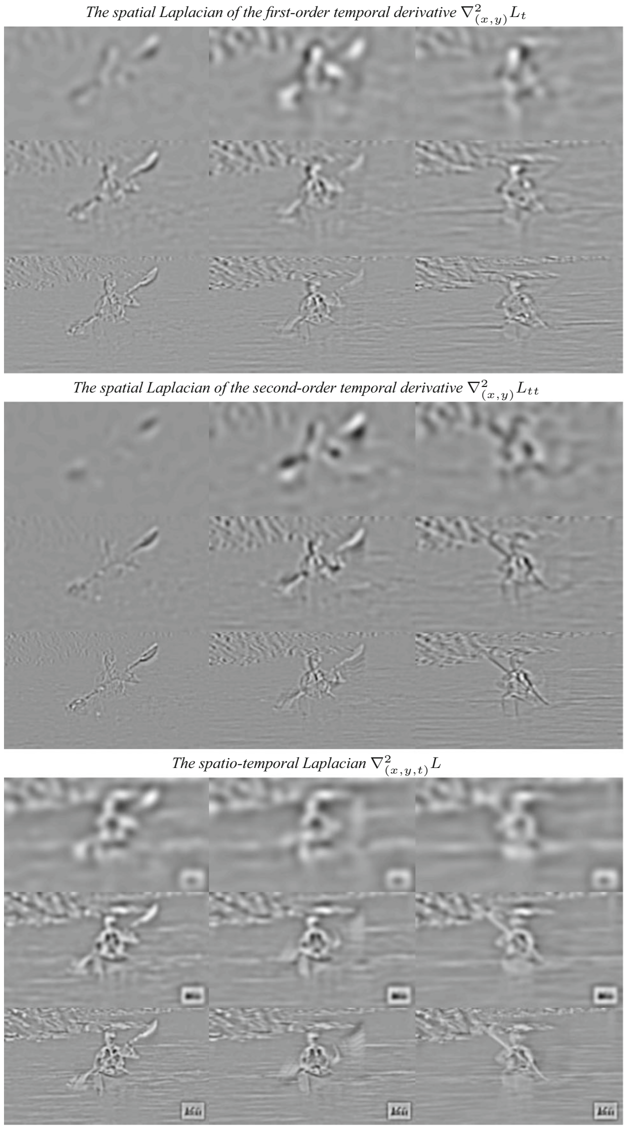

Spatio-temporal scale-space representation L

First-order temporal derivative L_t

Second-order temporal derivative L_{tt}

Fig. 1 Time-causal spatio-temporal scale-space representation $L(x, y, t;\quad s, \tau)$ with its first- and second-order temporal derivatives $L_t(x, y, t;\quad s, \tau)$ and $L_{tt}(x, y, t;\quad s, \tau)$ computed from a video sequence in the UCF-101 dataset (Kayaking_g01_c01.avi) at 3 × 3 combinations of the spatial scales (bottom row) $\sigma_{s,1} = 2$ pixels, (middle row) $\sigma_{s,2} = 4.6$ pixels and (top row) $\sigma_{s,3} = 10.6$ pixels and the temporal scales (left column) $\sigma_{\tau,1} = 40$ ms, (middle column) $\sigma_{\tau,2} = 160$ ms and (right column) $\sigma_{\tau,3} = 640$ ms with the spatial and temporal scale parameters in units of $\sigma_s = \sqrt{s}$ and $\sigma_\tau = \sqrt{\tau}$ and using a logarithmic distribution of the temporal scale levels with distribution parameter $c = 2$ (image size: 320 × 172 pixels of original 320 × 240 pixels; frame 90 of 226 frames at 25 frames/s)

For purely spatial data, the problem of performing spatial scale selection is nowadays rather well understood. Given the spatial Gaussian scale-space concept [24,34,44,46,47,59,60,67,70,106,111,120,123], a general methodology for spatial scale selection has been developed based on local extrema over spatial scales of scale-normalized differential entities [62,64,65,72,73]. This general method-

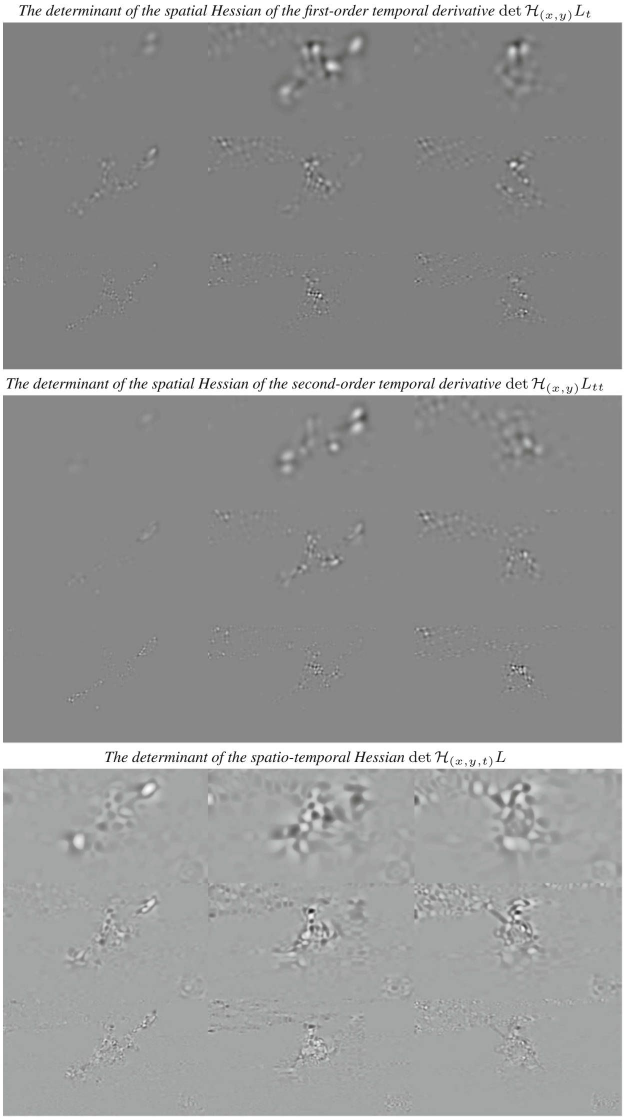


The spatial Laplacian of the first-order temporal derivative $\nabla^2_{(x,y)} L_t$

The spatial Laplacian of the second-order temporal derivative $\nabla^2_{(x,y)} L_{tt}$

The spatio-temporal Laplacian $\nabla^2_{(x,y,t)} L$

Fig. 2 The spatial Laplacian applied to the first- and second-order temporal derivatives $\nabla^2_{(x,y)} L_t$ and $\nabla^2_{(x,y)} L_{tt}$ as well as the spatio-temporal Laplacian $\nabla^2_{(x,y,t)} L$ computed from a video sequence in the UCF-101 dataset (Kayaking_g01_c01.avi) at 3 × 3 combinations of the spatial scales (bottom row) $\sigma_{s,1} = 2$ pixels, (middle row) $\sigma_{s,2} = 4.6$ pixels and (top row) $\sigma_{s,3} = 10.6$ pixels and the temporal scales (left column) $\sigma_{\tau,1} = 40$ ms, (middle column) $\sigma_{\tau,2} = 160$ ms and (right column) $\sigma_{\tau,3} = 640$ ms with the spatial and temporal scale parameters in units of $\sigma_s = \sqrt{s}$ and $\sigma_\tau = \sqrt{\tau}$ and using a time-causal spatio-temporal scale-space representation with a logarithmic distribution of the temporal scale levels for $c = 2$ (image size: 320 × 172 pixels of original 320 × 240 pixels; frame 90 of 226 frames at 25 framesframes/s)



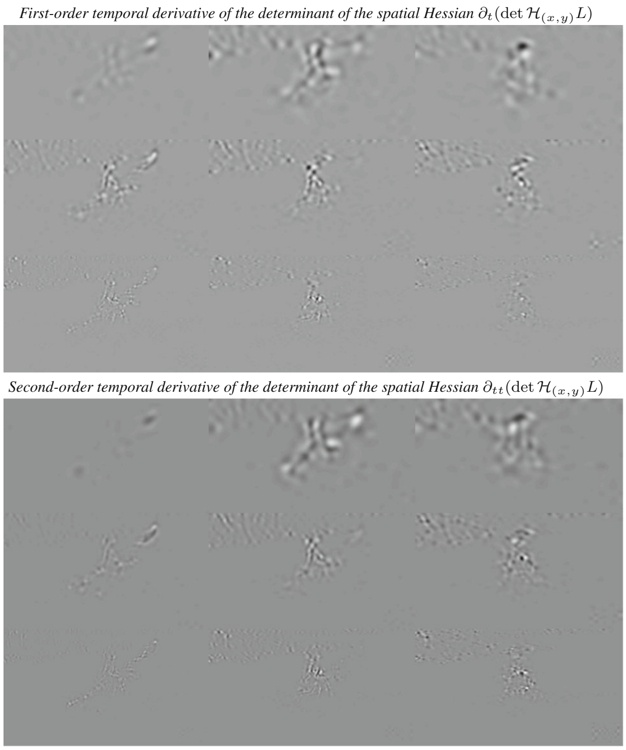
The determinant of the spatial Hessian of the first-order temporal derivative $\det \mathcal{H}_{(x,y)} L_t$

The determinant of the spatial Hessian of the second-order temporal derivative $\det \mathcal{H}_{(x,y)} L_{tt}$

The determinant of the spatio-temporal Hessian $\det \mathcal{H}_{(x,y,t)} L$

Fig. 3 The determinant of the spatial Hessian applied to the first- and second-order temporal derivatives $\det \mathcal{H}_{(x,y)} L_t$ and $\det \mathcal{H}_{(x,y)} L_{tt}$ as well as the determinant of the spatio-temporal Hessian $\det \mathcal{H}_{(x,y,t)} L$ computed from a video sequence in the UCF-101 dataset (Kayaking_g01_c01.avi) at 3 × 3 combinations of the spatial scales (bottom row) $\sigma_{s,1} = 2$ pixels, (middle row) $\sigma_{s,2} = 4.6$ pixels and (top row) $\sigma_{s,3} = 10.6$ pixels and the temporal scales (left column) $\sigma_{\tau,1} = 40$ ms, (middle column) $\sigma_{\tau,2} = 160$ ms and (right column) $\sigma_{\tau,3} = 640$ ms with the spatial and temporal scale parameters in units of $\sigma_s = \sqrt{s}$ and $\sigma_\tau = \sqrt{\tau}$ and using a time-causal spatio-temporal scale-space representation with a logarithmic distribution of the temporal scale levels for $c = 2$. The magnitude values of $\det \mathcal{H}_{(x,y,t)} L$ have been stretched by the monotone function $\phi(z) = (\operatorname{sign} z)\sqrt{|z|}$ (image size: 320 × 172 pixels of original 320 × 240 pixels; frame 90 of 226 frames at 25 frames/s)

ology has in turn been successfully applied to develop robust methods for image-based matching and recognition [5,41,52,68,74,84,86,87,89,90,112–114] that are able to handle large variations of the size of the objects in the image domain and with numerous applications regarding object recognition, object categorization, multi-view geometry, construction of 3-D models from visual input, human–computer interaction, biometrics and robotics. Alternative approaches for spatial scale selection in other problem domains have also been proposed [7,8,10,19,28,29,31,38–40,54,55,66,82,83,85,91,92,105,109,115].



First-order temporal derivative of the determinant of the spatial Hessian $\partial_t(\det \mathcal{H}_{(x,y)}L)$

Second-order temporal derivative of the determinant of the spatial Hessian $\partial_{tt}(\det \mathcal{H}_{(x,y)}L)$

Fig. 4 The first- and second-order temporal derivatives of the determinant of the spatial Hessian $\partial_t(\det \mathcal{H}_{(x,y)}L)$ and $\partial_{tt}(\det \mathcal{H}_{(x,y)}L)$ computed from a video sequence in the UCF-101 dataset (Kayaking_g01_c01.avi) at 3 × 3 combinations of the spatial scales (bottom row) $\sigma_{s,1} = 2$ pixels, (middle row) $\sigma_{s,2} = 4.6$ pixels and (top row) $\sigma_{s,3} = 10.6$ pixels and the temporal scales (left column) $\sigma_{\tau,1} = 40$ ms, (middle column) $\sigma_{\tau,2} = 160$ ms and (right column) $\sigma_{\tau,3} = 640$ ms with the spatial and temporal scale parameters in units of $\sigma_s = \sqrt{s}$ and $\sigma_\tau = \sqrt{\tau}$ and using a time-causal spatio-temporal scale-space representation with a logarithmic distribution of the temporal scale levels for $c = 2$. The magnitude values of $\det \mathcal{H}_{(x,y,t)}L$ have been stretched by the monotone function $\phi(z) = (\operatorname{sign} z)\sqrt{|z|}$ (image size: 320 × 172 pixels of original 320 × 240 pixels; frame 90 of 226 frames at 25 frames/s)

Much less research has, however, been performed on developing methods for choosing locally appropriate temporal scales for spatio-temporal analysis of video data. While some methods for temporal scale selection have been developed [49,63,122], the earliest methods suffer from either theoretical or practical limitations: the initial work on time-causal temporal scale selection in Lindeberg [63] is primarily developed over the discrete temporal Poisson scale space, which possesses a semi-group property over temporal scales and therefore leads to unnecessarily long temporal delays for reasons explained in Lindeberg [77, Appendix A]. The spatio-temporal scale selection method in Laptev and Lindeberg [49] is based on a spatio-temporal Laplacian operator that is not scale covariant under independent relative scaling transformations of the spatial versus the temporal domains (see Sect. 4.8), which implies that the spatial and temporal scale estimates will not be robust under independent variations of the spatial and temporal scales in video data as arise, for example, when viewing the same scene with two cameras having different sensor characteristics in terms of spatial resolution or temporal frame rate. The spatio-temporal scale selection method for the determinant of the spatio-temporal Hessian in Willems et al. [122] does not make use of the full flexibility of the notion of γ-normalized derivative operators (see Sect. 4.5) and has not been previously developed over a time-causal and time-recursive spatio-temporal domain as is necessary for processing real-time image streams with requirements of short temporal latencies of the feature responses for time-critical applications and complementary requirements about only small compact buffers of past information.

The subject of this article is to develop an extended theory for performing spatio-temporal scale selection in video data, to generate hypotheses about local characteristic spatial and temporal scales in the video data before recognizing the objects or the spatio-temporal events in the scene that the camera is observing. For this domain, we can consider two basic use cases: For offline analysis of pre-recorded video, one may take the liberty of accessing the virtual future in relation to any pre-recorded time moment and make use of symmetric filtering over the temporal domain based on the non-causal Gaussian spatio-temporal scale-space theory [61,67,70]. For online analysis of real-time video streams on the other hand, the future cannot be accessed and we will base the analysis on a fully time-causal and time-recursive spatio-temporal scale-space concept for real-time image streams that only requires access to information from the present moment and a very compact buffer of what has occurred in the past [75] and which constitutes an extension of previous temporal scale-space and multi-scale models [23,27,45,81,110]. Specifically, for performing spatio-temporal feature detection in the latter time-causal scenario, we will build upon a recently developed theory for temporal scale selection in a time-causal scale-space representation [77] and extend that theory to spatio-temporal scale selection for features that are computed based on a time-causal spatio-temporal scale-space representation. The resulting theory that we will arrive at can be seen as an extension of the previously developed spatial scale selection methodology [64,65,73] from spatial images to spatio-temporal video and real-time image streams.

To begin, we will start developing our theory for spatio-temporal scale selection with respect to the problem of detecting sparse spatio-temporal interest points [6,9,11,14,18,20, 21,30,49,88,94,97,99,100,107,122,124,126,127], which may be regarded as a conceptually simplest problem domain because of the sparsity of spatio-temporal interest points and the close connection between this problem domain and the detection of spatial interest points for which there exists a theoretically well-founded and empirically tested framework regarding scale selection over the spatial domain [1,4,5,15,17,25,42,65,72,74,84,89,90,112]. Specifically, using a non-causal Gaussian spatio-temporal scale-space model, we will perform a theoretical analysis of the spatio-temporal scale selection properties of eight different types of spatio-temporal interest point detectors and show that seven of them: (i) the spatial Laplacian of the first-order temporal derivative, (ii) the spatial Laplacian of the second-order temporal derivative, (iii) the determinant of the spatial Hessian of the first-order temporal derivative, (iv) the determinant of the spatial Hessian of the second-order temporal derivative, (v) the determinant of the spatio-temporal Hessian matrix, (vi) the first-order temporal derivative of the determinant of the spatial Hessian matrix and (vii) the second-order temporal derivative of the determinant of the spatial Hessian matrix, do all lead to fully scale-covariant spatio-temporal scale estimates and scale-invariant feature responses under independent scaling transformations of the spatial and the temporal domains. For (viii) the spatio-temporal Laplacian, it is on the other hand not possible to achieve scale covariance or scale invariance, which explains the poor robustness of the spatio-temporal interest points computed from the spatio-temporal Harris operator with scale selection based on the spatio-temporal Laplacian [49] on video data in which there are large independent variations in the spatial and temporal scales of the underlying spatio-temporal image structures.

Then, we will show how this theory can be transferred to an implementation based on fully time-causal spatio-temporal receptive fields to enable the detection of spatio-temporal features from real-time image streams in which the future cannot be accessed. Specifically, since any time-causal image measurement at a nonzero temporal scale will be associated with a nonzero temporal delay, we will introduce an additional parameter q to enable scale calibration of the spatio-temporal interest point detectors to deliver a temporal scale estimate at temporal scale $\hat{\sigma}_\tau = q\,\hat{\sigma}_{\tau,0}$ for $q \leq 1$ as opposed to the over the spatial domain more common choice of $\hat{\sigma}_s = \hat{\sigma}_{s,0}$ to enable shorter temporal delays and therefore the ability to respond faster in time-critical real-time scenarios, motivated by the general observation that the temporal delay can be expected to be proportional to the temporal scale level when expressed in units of the temporal standard deviation of the temporal scale-space kernel.

Whereas the explicit algorithms and experiments in this paper are restricted to spatio-temporal scale selection at sparse interest points over space and time, in a companion paper [76] we develop complementary methods for computing dense maps of spatial and temporal scale estimates in video data based on a structurally similar theory.

1.1 Structure of this Article

As conceptual background to the work, Sect. 2 starts by describing the theoretical model for spatio-temporal receptive fields and the resulting scale-space concepts that we build upon for computing image and video representations over multiple spatial and temporal scales.

When to develop a theory for spatio-temporal scale selection, main questions regarding the foundations concern what properties the scale selection method should possess and how the scale estimates should be computed. In Sect. 3, we show how it is possible to construct a well-founded theory for simultaneous selection of spatial and temporal scales in video data, by detecting local extrema over spatial and temporal scales of appropriately scale-normalized spatio-temporal derivative responses. This theory is generally valid for a large class of homogeneous spatio-temporal differential invariants and beyond the more explicit examples of spatio-temporal feature detectors considered in more detail in later sections. This theory specifically includes a general statement about scale-covariant properties of the resulting spatio-temporal scale estimates, which implies that the scale estimates are guaranteed to adaptively follow variabilities in spatial and temporal scale levels in the data. This theory also comprises scale-invariant properties of the resulting spatio-temporal features and their magnitude strength measures, which imply that similar types of spatio-temporal image features, while at different scales, will be computed, if the data in video sequence are subject to independent scaling transformations of the spatial and the temporal domains. In these respects, the proposed theory obeys the desirable properties of a spatio-temporal scale selection methodology.

The theory presented so far, does, however, comprise two free parameters, a spatial scale normalization power γ_s and a temporal scale normalization power γ_τ. To understand the behaviour of spatio-temporal feature detectors over multiple scales in more specific situations, Sect. 4 does then show how the scale selection properties of spatio-temporal feature detectors can be analysed by calculating their feature responses at multiple spatio-temporal scales in closed form to determine the scale normalization powers γ_s and γ_τ.

Specifically, we present an in-depth analysis of the theoretical scale selection properties of eight spatio-temporal derivative expressions that may be considered as candidates for defining spatio-temporal interest point detectors, when applied to idealized model patterns in the form of Gaussian blinks or Gaussian onset blobs of different spatial extent and of different temporal duration. By requiring that the selected spatial and temporal scales should reflect the spatial extent and the temporal duration of the input pattern, we show that seven of these spatio-temporal derivative expressions: (i)–(ii) the spatial Laplacian of the first- and second-order temporal derivatives, (iii)–(iv) the determinant of the spatial

Hessian of the first- and second-order temporal derivatives, (v) the determinant of the spatio-temporal Hessian matrix and (vi)–(vii) the first- and second-order temporal derivatives of the determinant of the spatial Hessian matrix, can be scale calibrated to reflect the spatial extent and the temporal duration of the underlying spatio-temporal image structures that gave rise to the filter responses. For one of these expressions, an attempt to define a spatio-temporal Laplacian operator, the lack of scale covariance under independent scaling transformations of the spatial and temporal domains, corresponding scale-invariant scale calibration cannot, however, be done for that operator. That in turn implies that applying the spatio-temporal Laplacian to video data in which there are unknown spatio-temporal scale variations can be expected to lead to undesirable artefacts.

In Sect. 5, we then present a general algorithm for detecting spatio-temporal interest points from spatio-temporal scale-space extrema of scale-normalized spatio-temporal expressions. Specifically, we present a detailed algorithm for detecting such image features based on a time-causal and time-recursive spatio-temporal scale-space representation. Compared to a corresponding algorithm expressed over a non-causal spatio-temporal scale space, as for the case of using a Gaussian spatio-temporal scale space for analysing pre-recorded video sequences, our time-causal algorithm does never access information from the past and can therefore be applied in real-time settings on video streams. Additionally, by the time-recursive formulation, the requirements about temporal buffering of past information are much lower and do also imply the need for less computations, thus improving the computational efficiency, also if applied in a non-causal setting for analysing pre-recorded video sequences.

As a verification of whether the proposed theory and methods do what they are supposed to do, Sect. 6 presents an experimental quantification of the numerical accuracy of the spatio-temporal scale estimates as well as the amount of temporal delay for the different types of spatio-temporal interest point detectors considered in this work, when applied to idealized spatio-temporal model patterns with ground truth and in the context of a time-causal spatio-temporal scale-space representation. The results do first of all show that the theoretical properties of spatio-temporal feature detectors responding at spatial and temporal scales corresponding to the spatial extent and the temporal duration do with very good approximation transfer to the proposed discrete implementation. Secondly, it is shown that the interest point detectors defined from applying either the spatial Laplacian or the determinant of the spatial Hessian to the first- or second-order temporal derivatives lead to significantly shorter temporal delays compared to the interest point detectors defined from the determinant of the spatio-temporal Hessian or the first- and second-order temporal derivatives of the determinant of the spatial Hessian. For time-critical applications, this implies that the temporal response properties from the first set of spatio-temporal feature detectors will be faster than for those from the other set and therefore the ability of an autonomous agent to react faster. Finally, Sect. 7 concludes with a summary and discussion.

1.2 Relations to Previous Contributions

This paper constitutes a substantially extended version of a shorter conference paper presented at the SSVM 2017 conference [79] and with substantial additions concerning:

- the motivations underlying the developments of this theory and the relations to previous work (Sect. 1),
- more details concerning the underlying spatio-temporal receptive field model (Sect. 2),
- a more extensive description about the proposed general methodology for spatio-temporal scale selection including: (i) its formulation based on temporal scale normalization by L_p-normalization of the temporal derivative operators, (ii) the theory for scale-invariant and scale-covariant properties of the resulting spatio-temporal features with their spatio-temporal scale estimates as well as (iii) spatio-temporal scale selection based on spatio-temporal differential invariants expressed in terms of local gauge coordinates that guarantee rotational invariance and which could not be included in the conference paper because of lack of space (Sect. 3),
- the treatment of two additional spatio-temporal differential invariants, the first- and second-order temporal derivatives of the determinant of the spatial Hessian matrix,
- the detailed theoretical analysis of the scale selection properties of the eight different spatio-temporal differential invariants treated in this paper and showing the explicit derivations of how the spatial and temporal scale normalization γ_s and γ_τ should be determined by scale calibration for each feature detector (Sect. 4),
- more complete details about the composed algorithm for detecting spatio-temporal interest points with spatio-temporal scale selection based on time-causal and time-recursive spatio-temporal receptive fields and including a change of order between the spatial and the temporal smoothing operations that substantially reduces the amount of computations (Sect. 5),
- an experimental quantification of the accuracy of the scale estimates and the temporal delays for the different types of spatio-temporal feature detectors when applied to idealized spatio-temporal model patterns (Sect. 6) and
- a detailed description of the corresponding spatial scale-space extrema algorithm on which the spatio-temporal scale-space extrema algorithm is based (“Appendix A”).

In relation to the SSVM 2017 paper, this paper therefore gives a more complete treatment of the subject, including more details about the spatio-temporal scale selection theory, much more complete algorithmic details when applying spatio-temporal scale selection in practice as well as a numerical quantification of the accuracy of the spatio-temporal scale estimates and the temporal responses properties (the temporal latencies in a time-causal setting).

2 Spatio-Temporal Receptive Field Model

For processing video data at multiple spatial and temporal scales, we follow the approach with idealized models of spatio-temporal receptive fields of the form

$$\begin{aligned} & T(x_1, x_2, t;\ s, \tau;\ v, \Sigma) \\ & \quad = g(x_1 - v_1 t, x_2 - v_2 t;\ s, \Sigma)\, h(t;\ \tau) \end{aligned} \tag{1}$$

as previously derived, proposed and studied in Lindeberg [67,69,75,78], where

- $x = (x_1, x_2)^{\mathrm{T}}$ denotes the image coordinates,
- t denotes time,
- s denotes the spatial scale,
- τ denotes the temporal scale,
- $v = (v_1, v_2)^{\mathrm{T}}$ denotes a local image velocity,
- Σ denotes a spatial covariance matrix determining the spatial shape of a spatial affine Gaussian kernel

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

- $g(x_1 - v_1 t, x_2 - v_2 t;\ s, \Sigma)$ denotes a spatial affine Gaussian kernel that moves with image velocity $v = (v_1, v_2)$ in space–time and
- $h(t;\ \tau)$ is a temporal smoothing kernel over time,

and we specifically here choose as temporal smoothing kernel over time either: (i) the non-causal Gaussian kernel

$$h(t;\ \tau) = g(t;\ \tau) = \frac{1}{\sqrt{2\pi\tau}} \mathrm{e}^{-t^2/2\tau} \tag{3}$$

or (ii) the time-causal limit kernel [75, Equation (38)]

$$h(t;\ \tau) = \Psi(t;\ \tau, c) \tag{4}$$

defined via its Fourier transform of the form

$$\hat{\Psi}(\omega;\ \tau, c) = \prod_{k=1}^{\infty} \frac{1}{1 + i\, c^{-k} \sqrt{c^2 - 1} \sqrt{\tau}\, \omega} \tag{5}$$

and corresponding to an infinite cascade of truncated exponential kernels

$$h_{\exp}(t;\ \mu_i) = \begin{cases} \frac{1}{\mu_i} \mathrm{e}^{-t/\mu_i} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{6}$$

with logarithmically distributed temporal scale levels

$$\tau_k = \sum_{k=-\infty}^{k} \mu_i^2 = c^{2k} \tau_0 \tag{7}$$

that cluster infinitely dense near $\tau \downarrow 0^+$ [75].

Based on this spatio-temporal receptive field model, we define a spatio-temporal scale-space representation of the form [67,69,75]

$$\begin{aligned} & L(x_1, x_2, t;\ s, \tau;\ v, \Sigma) \\ & \quad = (T(\cdot, \cdot, \cdot;\ s, \tau;\ v, \Sigma) * f(\cdot, \cdot, \cdot))\, (x_1, x_2, t;\ s, \tau;\ v, \Sigma). \end{aligned} \tag{8}$$

When using a one-dimensional Gaussian kernel (3) for smoothing over the temporal domain, we obtain a non-causal Gaussian spatio-temporal scale space. When using the time-causal limit kernel (4) for temporal smoothing, we obtain a time-causal and time-recursive spatio-temporal scale space.

For simplicity, we shall in this treatment henceforth restrict ourselves to space–time separable receptive fields obtained by setting the image velocity to zero $v = (v_1, v_2)^{\mathrm{T}} = (0, 0)^{\mathrm{T}}$ and to receptive fields that are based on rotationally symmetric Gaussian kernels over the spatial domain by setting the spatial covariance matrix to a unit matrix $\Sigma = I$.

Figures 5 and 6 show examples of such space–time separable receptive fields over a 1+1-D space time, for the main cases when the temporal smoothing is performed using either the non-causal Gaussian kernel or the time-causal limit kernel.

An alternative model for time-causal temporal smoothing could be to instead use Koenderink's scale-time kernels [45], which correspond to Gaussian smoothing on a logarithmically transformed temporal domain. For reasons described in detail in Lindeberg [77, Section 2.2], in particular the lack of a known time-recursive formulation for Koenderink's scale-time kernels, which in turn implies a need for larger temporal buffers and more computational work for the temporal smoothing operation compared to using a time-recursive implementation of the time-causal limit kernel based on a set of recursive filters coupled in cascade [75, Section 6], we use the time-causal limit kernel for modelling the time-causal temporal smoothing operation in this work. As described in Lindeberg [75, Appendix 2], it is also possible to establish an approximate mapping between the parameters of the time-causal limit kernel and Koenderink's scale-time kernel based

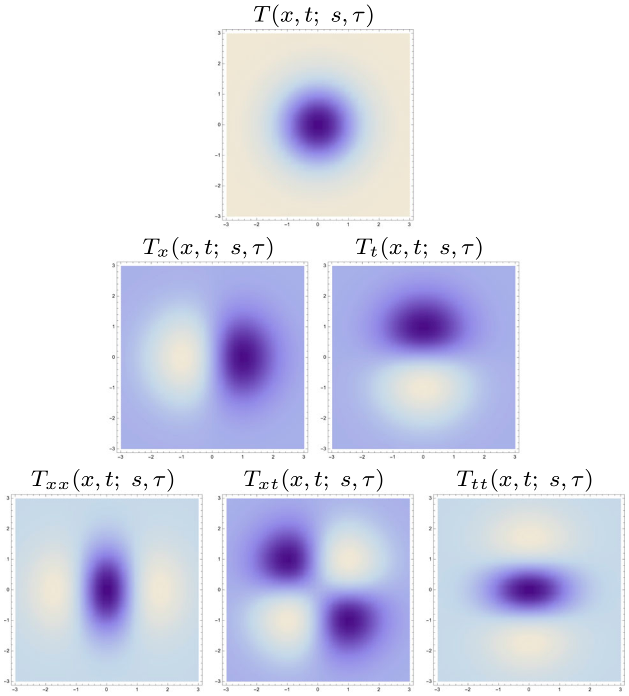


Fig. 5 *Space–time separable kernels* $T_{x^m t^n}(x, t;\ s, \tau) = \partial_{x^m t^n}(g(x;\ s)\, h(t;\ \tau))$ up to order two obtained as the composition of Gaussian kernels over the spatial domain x and the non-causal Gaussian kernel over the temporal domain ($s = 1, \tau = 1$) (horizontal axis: space $x \in [-3, 3]$; vertical axis: time $t \in [-3, 3]$)

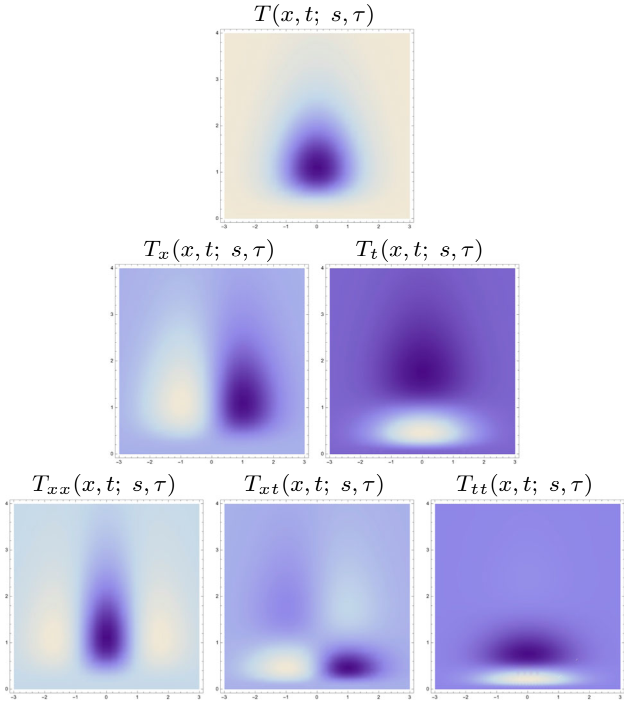


Fig. 6 *Space–time separable kernels* $T_{x^m t^n}(x, t;\ s, \tau) = \partial_{x^m t^n}(g(x;\ s)\, h(t;\ \tau))$ up to order two obtained as the composition of Gaussian kernels over the spatial domain x and the time-causal limit kernel over the temporal domain ($s = 1, \tau = 1, c = 2$) (horizontal axis: space $x \in [-3, 3]$; vertical axis: time $t \in [0, 4]$)

on the requirement that the zero-, first- and second-order temporal moments of the kernels in the two families should be equal [75, Equation (161)] and leading to qualitatively similar while not identical temporal receptive fields based on temporal derivatives of the time-causal scale-space kernels from the two families [75, Figure 11].

While yet a third type of ad hoc model for time-causal smoothing could possibly also be formulated based on truncated and time-delayed Gaussian kernels, with the temporal delay determined such that the truncation effects in some sense could be regarded as sufficiently small, we will not develop such an approach here because: (i) such a model could be expected to lead to significantly longer temporal delays and (ii) require significantly larger temporal buffers and more computational work compared to our family of time-causal and time-recursive scale-space kernels. For time-critical applications, where the temporal response properties of the vision system need to be as fast as possible, it should in general be much better to base the temporal processing on an inherently time-causal temporal scale-space concept.

2.1 Scale-Normalized Spatio-Temporal Derivatives

Specifically, a natural way of normalizing the spatio-temporal derivative operators within this space–time separable spatio-temporal scale-space concept

$$\begin{aligned} & L(x_1, x_2, t;\ s, \tau) \\ & \quad = (T(\cdot, \cdot, \cdot;\ s, \tau) * f(\cdot, \cdot, \cdot))\, (x_1, x_2, t;\ s, \tau) \end{aligned} \tag{9}$$

with respect to the spatial and temporal scale parameters is by introducing scale-normalized derivative operators according to Lindeberg [65,75]

$$\partial_\xi = \partial_{x,\text{norm}} = s^{\gamma_s/2}\, \partial_x, \tag{10}$$

$$\partial_\eta = \partial_{y,\text{norm}} = s^{\gamma_s/2}\, \partial_y, \tag{11}$$

$$\partial_\zeta = \partial_{t,\text{norm}} = \alpha_n(\tau)\, \partial_t, \tag{12}$$

and studying scale-normalized partial derivates of the form [75, Equation (108)]

$$L_{x_1^{m_1} x_2^{m_2} t^n,\text{norm}} = s^{(m_1+m_2)\gamma_s/2}\, \alpha_n(\tau)\, L_{x_1^{m_1} x_2^{m_2} t^n}, \tag{13}$$

where the factor $s^{(m_1+m_2)\gamma_s/2}$ transforms the regular spatial partial derivatives to corresponding scale-normalized spatial derivatives with γ_s denoting the spatial scale normalization parameter [65] and the factor $\alpha_n(\tau)$ is the scale normalization factor for scale-normalized temporal derivatives determined according to either: (i) variance-based normalization [75, Equation (74)]

$$\alpha_n(\tau) = \tau^{n\gamma_\tau/2} \tag{14}$$

or (ii) L_p-normalization [75, Equation (76)]

$$\alpha_n(\tau) = \frac{\|g_{\xi^n}(\cdot;\ \tau)\|_p}{\|h_{t^n}(\cdot;\ \tau)\|_p} = \frac{G_{n,\gamma_\tau}}{\|h_{t^n}(\cdot;\ \tau)\|_p} \tag{15}$$

with G_{n,γ_τ} denoting the L_p-norm of the non-causal temporal Gaussian derivative kernel for the γ_τ-value for which this L_p-norm becomes constant over temporal scales (see [75, Equations (80)–(83)]).

2.2 Temporal Delays

For the non-causal temporal scale-space concept given by convolution with symmetric temporal Gaussian kernels of the form (3), the temporal delay is always zero. When using time-causal temporal scale-space kernels, there will on the other hand always be a nonzero temporal delay δ. Unfortunately, because of the lack of compact closed-form expression for the time-causal limit kernel (4) over the temporal domain, it is non-trivial to derive an compact closed-form expression for its exact temporal delay. Based on a scale-time approximation of the time-causal limit kernel, it is, however, possible to derive the following approximate expression for the temporal maximum of the temporal smoothing kernel [75, Equation (172)][1]

$$\delta \approx \frac{(c+1)^2\sqrt{\tau}}{2\sqrt{2}\sqrt{(c-1)\,c^3}}. \tag{16}$$

From this expression, we can see that the temporal delay δ increases linearly with the temporal scale $\sigma_\tau = \sqrt{\tau}$ in units of the standard deviation of the temporal smoothing kernel. Additionally, the temporal delay depends on the distribution parameter c of the time-causal limit kernel in such a way that larger values of $c > 1$ lead to shorter temporal delays at the cost of a sparser temporal scale sampling.

3 General Spatial-Temporal Scale Selection Methodology

In this section, we will describe a general spatio-temporal scale selection methodology for simultaneous computation of local characteristic spatial and temporal scale estimates from video data, which for appropriate choices of spatio-temporal derivative expressions for feature detection may reflect the spatial extent and the temporal duration of the underlying spatio-temporal image structures that gave rise to the feature responses.

3.1 Homogeneous Spatio-Temporal Differential Expressions

An essential property of the definition of scale-normalized spatio-temporal derivative operators according to (13) is that they will lead to scale-covariant spatio-temporal image features, if the spatial smoothing performed by a spatial Gaussian kernel (2) and if the temporal smoothing is performed with either a non-causal temporal Gaussian kernel (3) or the time-causal limit kernel (4), provided that the underlying spatio-temporal expression $\mathcal{D}_{\text{norm}}L$ used for defining the spatio-temporal features is covariant under independent scaling transformations of the spatial and temporal domains.

To express this property compactly, let us introduce multi-index notation for spatio-temporal derivatives

$$L_{x^\alpha t^\beta} = L_{x_1^{\alpha_1} x_2^{\alpha_2} t^\beta}, \tag{17}$$

where $x = (x_1, x_2)$, $\alpha = (\alpha_1, \alpha_2)$ and $|\alpha| = \alpha_1 + \alpha_2$. Then, consider a spatio-temporal differential expression of the form

$$\mathcal{D}L = \sum_{i=1}^{I} \prod_{j=1}^{J} c_i\, L_{x^{\alpha_{ij}} t^{\beta_{ij}}} = \sum_{i=1}^{I} \prod_{j=1}^{J} c_i\, L_{x_1^{\alpha_{1ij}} x_2^{\alpha_{2ij}} t^{\beta_{ij}}}, \tag{18}$$

where the sum of the orders of spatial and temporal differentiation in a certain term

$$\sum_{j=1}^{J} |\alpha_{ij}| = \sum_{j=1}^{J} \alpha_{1ij} + \alpha_{2ij} = M \tag{19}$$

$$\sum_{j=1}^{J} \beta_{ij} = N \tag{20}$$

does not depend on the index i of that term. Such a differential expression is referred to as homogeneous.

3.2 Transformation Property Under Independent Scaling Transformations of the Spatial and the Temporal Domains

Consider next an independent scaling transformation of the spatial and the temporal domains of a video sequence

$$f'\left(x_1', x_2', t'\right) = f(x_1, x_2, t) \tag{21}$$

[1] When computing estimates of the temporal delay of the time-causal spatio-temporal scale-space kernel in an actual discrete implementation, we do, however, not make use of the approximate expression (16). Instead, we do for each temporal scale level compute the temporal maximum point of the discrete time-causal scale-space kernel that approximates the continuous time-causal kernel, and do then add additionally half a time step $\Delta t/2$ for each order of temporal differentiation as implemented in terms of backward difference operators over time $\partial_{t^n} L \approx \delta_t^n L/(\Delta t)^n$, where Δt denotes the temporal time step between successive frames.

for

$$\left(x_1', x_2', t'\right) = (S_s\, x_1, S_s\, x_2, S_\tau\, t), \tag{22}$$

where S_s and S_τ denote the spatial and temporal scaling factors, respectively, and define the space–time separable spatio-temporal scale-space representations L and L' of f and f', respectively, according to

$$\begin{aligned} &L(x_1, x_2, t;\ s, \tau)\\ &\quad = (T(\cdot, \cdot, \cdot;\ s, \tau) * f(\cdot, \cdot, \cdot))\,(x_1, x_2, t;\ s, \tau), \qquad (23)\\ &L'\left(x_1', x_2', t';\ s', \tau'\right)\\ &\quad = \left(T(\cdot, \cdot, \cdot;\ s', \tau') * f'(\cdot, \cdot, \cdot)\right)\left(x_1', x_2', t';\ s', \tau'\right). \qquad (24) \end{aligned}$$

These spatio-temporal scale-space representations are closed under independent scaling transformations of the spatial and the temporal domains

$$L'\left(x_1', x_2', t';\ s', \tau'\right) = L(x_1, x_2, t;\ s, \tau) \tag{25}$$

provided that the spatio-temporal scale levels are appropriately matched [67,75]

$$s' = S_s^2\, s, \quad \tau' = S_\tau^2 \tau. \tag{26}$$

For the non-causal Gaussian spatio-temporal scale space having a continuum of both spatial and temporal scale levels, this closedness relation holds for all spatial scaling factors $S_s > 0$ and all temporal scaling factors $S_\tau > 0$. For the time-causal spatio-temporal scale-space representation having a continuum of spatial scale levels, while the temporal scale levels are restricted to be discrete (7), the scaling relation holds for all spatial scaling factors $S_s > 0$, whereas the closedness relation under temporal scaling transformations holds only for temporal scaling factors of the form $S_\tau = c^j\,(j \in \mathbb{Z})$ that correspond to exact mappings between the discrete temporal scale levels (7), where $c > 1$ is the distribution parameter of the time-causal limit kernel (4).

Specifically, a homogeneous spatio-temporal derivative expression of the form (18) with the spatio-temporal derivatives $L_{x_1^{m_1} x_2^{m_2} t^n}$ replaced by scale-normalized spatio-temporal derivatives $L_{x_1^{m_1} x_2^{m_2} t^n,\mathrm{norm}}$ according to (13) transforms according to

$$\mathcal{D}'_{\mathrm{norm}} L' = S_s^{M(\gamma_s - 1)}\, S_\tau^{N(\gamma_\tau - 1)}\, \mathcal{D}_{\mathrm{norm}} L. \tag{27}$$

This result follows from a combination and generalization of Equation (25) in [65], which states that a purely spatial differential expression of the form

$$\mathcal{D}L = \sum_{i=1}^{I} \prod_{j=1}^{J} c_i\, L_{x^{\alpha_{ij}}} \tag{28}$$

when expressed in terms of scale-normalized spatial derivatives transforms according to

$$\mathcal{D}'_{\mathrm{norm}} L' = S_s^{M(\gamma_s - 1)}\, \mathcal{D}_{\mathrm{norm}} L \tag{29}$$

with Equations (10) and (104) in [77], which state that an nth-order temporal derivative transforms according to

$$\partial_{t'^n,\mathrm{norm}} L' = S_\tau^{n(\gamma_\tau - 1)}\, \partial_{t^n,\mathrm{norm}} L. \tag{30}$$

With the temporal smoothing performed by the scale-invariant limit kernel (4), the temporal scaling transformation property does, however, only hold for temporal scaling transformations that correspond to exact mappings between the discrete temporal scale levels $\tau_i = \tau_0\, c^{2i}$ in the time-causal temporal scale-space representation and thus to temporal scaling factors $S_\tau = c^i$ that are integer powers of the distribution parameter c of the time-causal limit kernel.

The scaling property (27) of homogeneous polynomial spatio-temporal differential invariants also extends to homogenous rational expressions of spatio-temporal derivatives, i.e., rational expressions formed by ratios of two homogeneous polynomials of the form (18).

3.3 General Scale-Covariant Property of the Spatio-Temporal Scale Estimates

The scale-covariant property (27) implies that local extrema over spatio-temporal scales are preserved under independent scaling transformations of the spatial and the temporal domains and that local (possibly multi-valued) spatio-temporal scale estimates obtained from local extrema over spatio-temporal scales[2]

$$\begin{aligned} &\{(\hat{s}, \hat{\tau})\}(x, y, t)\\ &\quad = \operatorname{argmaxminlocal}_{s,\tau}\, (\mathcal{D}_{\mathrm{norm}} L)(x, y, t;\ s, \tau) \end{aligned} \tag{31}$$

are guaranteed to transform in a scale-covariant way under independent scaling transformations of the spatial and the temporal domains

$$\left(\hat{s}', \hat{\tau}'\right) = \left(S_s^2\, \hat{s}, S_\tau^2\, \hat{\tau}\right) \tag{32}$$

or in units of the standard deviation $(\sigma_s, \sigma_\tau) = (\sqrt{s}, \sqrt{\tau})$ of the spatio-temporal scale-space kernel

$$\left(\hat{\sigma}_s', \hat{\sigma}_\tau'\right) = \left(S_s\, \hat{\sigma}_s, S_\tau\, \hat{\sigma}_\tau\right) \tag{33}$$

[2] This notation is intended to reflect the fact that a set of multiple spatio-temporal scale estimates $(\hat{s}, \hat{\tau})$ may be obtained at any point (x, y, t) in space–time, corresponding to qualitatively different types of spatio-temporal image structures at different spatio-temporal scales.

provided that the spatial positions (x, y) and the temporal moments t are appropriately matched

$$\left(x_1', x_2', t'\right) = (S_s\, x_1, S_s\, x_2, S_\tau\, t). \tag{34}$$

Specifically, the scale-covariant property (27) implies that if we can detect a spatio-temporal scale level $(\hat{s}, \hat{\tau})$ such that the scale-normalized expression $\mathcal{D}_{\text{norm}} L$ assumes a local extremum over both space–time (x_1, x_2, t) and spatio-temporal scales (s, τ) at some point $(\hat{x}_1, \hat{x}_2, \hat{t};\;\; \hat{s}, \hat{\tau})$ in spatio-temporal scale space, then this local extremum is preserved under independent scaling transformations of the spatial and temporal domains and is transformed in a scale-covariant way

$$(\hat{x}_1, \hat{x}_2, \hat{t};\;\; \hat{s}, \hat{\tau}) \mapsto \left(S_s\, \hat{x}_1, S_s\, \hat{x}_2, S_\tau\, \hat{t};\;\; S_s^2\, \hat{s}, S_\tau^2\, \hat{\tau}\right). \tag{35}$$

The properties (27), (32) and (35), which mean that spatio-temporal scale estimates follow local independent spatial and temporal scaling transformations in video data, constitute a theoretical foundation for scale-covariant spatio-temporal scale selection and scale-invariant feature detection.

3.4 General Scale-Covariant and Scale-Invariant Properties of Feature Responses at Local Extrema Over Spatio-Temporal Scales

Additionally, the magnitude of the feature response $(\mathcal{D}_{\text{norm}} L)_{\text{extr}}$ at the spatio-temporal scale-space extremum over spatial and temporal scales will also transform according to power law

$$\left(\mathcal{D}'_{\text{norm}} L'\right)_{\text{extr}} = S_s^{M(\gamma_s - 1)}\, S_\tau^{N(\gamma_\tau - 1)}\, (\mathcal{D}_{\text{norm}} L)_{\text{extr}}. \tag{36}$$

In the special case when the scale normalization powers $\gamma_s = 1$ and $\gamma_\tau = 1$, the magnitude responses at the scale-space extrema will be equal.

For reasons that will be explained later in Sect. 4, there are, however, situations where it can be highly motivated to use scale normalization powers not equal to one. Then, the important message is that the magnitude estimates are transformed by a power law and can be compensated for by post-normalization of the magnitude responses that also takes the actual spatio-temporal scale levels into account.

3.5 Spatio-Temporal Scale Selection for Homogeneous Spatio-Temporal Differential Invariants in Terms of Gauge Coordinates

Introduce at every point (x_1, x_2, t) in space–time, local orthonormal gauge coordinate systems (u, v, t) and (p, q, t) oriented such that: (i) the v-direction is parallel to the spatial gradient direction of L and the u-direction is orthogonal in image space with the partial derivative in the u-direction being zero $L_u = 0$ and (ii) the p- and q-directions are parallel with the eigendirections of the spatial Hessian matrix $\mathcal{H}_{(x,y)} L$ such that the mixed spatial second-order derivative is zero $L_{pq} = 0$. Then, consider spatio-temporal differential expressions of the forms

$$\mathcal{D} L = \sum_{i=1}^{I} \prod_{j=1}^{J} c_i\, L_{u^{\alpha_{1ij}}\, v^{\alpha_{2ij}}\, t^{\beta_{ij}}} \tag{37}$$

or

$$\mathcal{D} L = \sum_{i=1}^{I} \prod_{j=1}^{J} c_i\, L_{p^{\alpha_{1ij}}\, q^{\alpha_{2ij}}\, t^{\beta_{ij}}} \tag{38}$$

that satisfy the homogeneity requirements

$$\sum_{j=1}^{J} \alpha_{1ij} + \alpha_{2ij} = M \tag{39}$$

$$\sum_{j=1}^{J} \beta_{ij} = N \tag{40}$$

for all $i \in [1, I]$. Then, by the construction from these rotationally invariant gauge coordinates, these spatio-temporal differential expressions are guaranteed to be invariant under global rotations of the spatial domain. Additionally, because of the homogeneity of these expressions in terms of the total orders of spatial and temporal differentiation in each term, simultaneous spatial and temporal scale selection based on corresponding scale-normalized derivatives is guaranteed to lead to scale-covariant scale estimates.

As a consequence, the scale estimates will be guaranteed to be rotationally invariant in the sense that if the spatial domain is globally rotated in image space, then both the spatial and the temporal scale estimates will be rotated in the same way as the spatial image positions. A corresponding rotational invariance property of the spatio-temporal scale estimates does also hold for other types of spatio-temporal differential expressions of the form (18) that are additionally rotationally invariant.

What remains in this theory is to choose appropriate scale-normalized spatio-temporal derivative expressions $\mathcal{D}_{\text{norm}} L$ for different visual tasks and to tune the scale normalization powers γ_s and γ_τ to additional complementary requirements. In next section, we will perform a detailed study of this for eight different spatio-temporal differential invariants with respect to the task of detecting spatio-temporal interest points.

4 Spatio-Temporal Scale Selection in Non-Causal Gaussian Spatio-Temporal Scale Space

In this section, we will perform a closed-form theoretical analysis of the spatial and the temporal scale selection properties that are obtained by detecting simultaneous local extrema over both spatial and temporal scales of different scale-normalized spatio-temporal differential expressions. We will specifically analyse: (i) how the spatial and temporal scale estimates $\hat{s}$ and $\hat{\tau}$ are related to the spatial extent s_0 and the temporal duration τ_0 for different types of spatio-temporal model signals for which closed-form theoretical analysis is possible and (ii) how the resulting scale-normalized magnitude responses of the different differential entities at the selected spatio-temporal scales depend upon the spatial extent s_0 and the temporal duration τ_0 of the underlying image structures as well as upon a complementary parameter q introduced to enable detection of spatio-temporal image features at finer temporal scales than at the temporal scales at which they occur, to in turn enable shorter temporal delays when computing image features based on a time-causal spatio-temporal scale-space concept.

A main goal is to perform *scale calibration*, to determine suitable values of the spatial and temporal scale normalization parameters γ_s and γ_τ for different types of spatio-temporal feature detectors, in such a way that the selected spatial and temporal scale levels reflect the spatial extent and the temporal duration of the original spatio-temporal image structures that gave rise to the feature response. The methodology we shall follow is to calculate scale-space representations in closed form for Gaussian-based spatio-temporal image patterns for which the non-causal spatio-temporal scale-space representation can be obtained from the semi-group property of the Gaussian kernel. Then, given that explicit expressions can be calculated for the scale-normalized spatio-temporal derivatives, we will solve for the local extrema of the spatio-temporal differential invariant $\mathcal{D}_{\text{norm}} L$ over spatio-temporal scales, to define equations that determine the scale normalization powers γ_s and γ_τ from the constraints that the spatio-temporal scale estimates should obey $\hat{s} = s_0$ and $\hat{\tau} = q^2 \, \tau_0$.

The spatial assumption $\hat{s} = s_0$ is similar to the method for scale calibration in the spatial scale selection methodology [64,65,72] and corresponds to detecting the image structure at the same scale as they appear, which should be optimal with regard to signal detection theory. Regarding the temporal assumption $\hat{\tau} = q^2 \, \tau_0$, we do, however, also introduce a parameter $q < 1$ to enforce temporal scale selection at finer temporal scales, to enable shorter temporal delays of the feature responses. As previously described in Sect. 2.2, for the time-causal scale-space representation the temporal delay can be expected to be proportional to the temporal scale in units of the standard deviation of the temporal smoothing kernel $\delta \sim \sigma_\tau = \sqrt{\tau}$. A first-order prediction is therefore that a value of $q < 1$ can be expected to reduce the temporal delay by the order of a corresponding factor, to enable an autonomous agent using these features as input to respond faster in a time-critical real-time situation.

4.1 The Spatial Laplacian of the Second-Order Temporal Derivative

Inspired by the way neurones in the lateral geniculate nucleus (LGN) respond to visual input [12,13], which for many LGN cells can be modelled by idealized operations of the form [69, Equation (108)]

$$h_{\text{LGN}}(x, y, t;\ s, \tau) = \pm(\partial_{xx} + \partial_{yy})\, g(x, y;\ s)\, \partial_{t^n}\, h(t;\ \tau), \tag{41}$$

let us for general values of the spatial and temporal scale normalization parameters γ_s and γ_τ study the scale-normalized spatial Laplacian of the second-order temporal derivative defined according to

$$\begin{aligned} \nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}} &= s^{\gamma_s}\, \tau^{\gamma_\tau}\, \nabla^2_{(x,y)} L_{tt} \\ &= s^{\gamma_s}\, \tau^{\gamma_\tau}\, \left(L_{xxtt} + L_{yytt}\right), \end{aligned} \tag{42}$$

which in turn can be seen as an idealized functional model of a so-called “lagged” LGN neurone (compare with [69, Figure 24, right column]). This operator can be expected to give a strong response when both the spatial Laplacian and the second-order temporal derivative give strong responses, e.g., for blinking blobs.

Consider a spatio-temporal image pattern defined as a Gaussian blink with spatial extent s_0 and temporal duration τ_0:

$$\begin{aligned} f(x, y, t) &= g(x, y;\ s_0)\, g(t;\ \tau_0) \\ &= \frac{1}{(2\pi)^{3/2} s_0 \sqrt{\tau_0}}\, e^{-(x^2+y^2)/2s_0}\, e^{-t^2/2\tau_0}. \end{aligned} \tag{43}$$

By spatial smoothing with the two-dimensional spatial Gaussian kernel and temporal smoothing with the non-causal one-dimensional Gaussian kernel, the resulting spatio-temporal scale-space representation will be of the form

$$L(x, y, t;\ s, \tau) = g(x, y;\ s_0 + s)\, g(t;\ \tau_0 + \tau), \tag{44}$$

for which the scale-normalized Laplacian of the second-order temporal derivative at the origin $(x, y, t) = (0, 0, 0)$ is given by

$$\left.\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}\right|_{(0,0,0)} = \frac{s^{\gamma_s} \tau^{\gamma_\tau}}{\sqrt{2}\pi^{3/2} (s + s_0)^2 (\tau + \tau_0)^{3/2}}. \tag{45}$$

Differentiating this expression with respect to the spatial scale parameter s and the temporal scale parameter τ and setting the derivative to zero implies that the local extremum over spatial and temporal scales is given by

$$\hat{s} = \frac{\gamma_s s_0}{2 - \gamma_s}, \tag{46}$$

$$\hat{\tau} = \frac{2 \gamma_\tau \tau_0}{3 - 2\gamma_\tau}. \tag{47}$$

If we require the spatial and temporal scale estimates to reflect the spatial and temporal extent of the Gaussian blink such that

$$\hat{s} = s_0, \tag{48}$$

$$\hat{\tau} = q^2 \tau_0, \tag{49}$$

then this implies that we should calibrate the scale normalization parameters γ_s and γ_τ according to

$$\gamma_s = 1, \tag{50}$$

$$\gamma_\tau = \frac{3q^2}{2(q^2+1)}, \tag{51}$$

where specifically the choice of $q = 1$ corresponds to $\gamma_\tau = 3/4$. For these values of γ_s and γ_τ, the scale-normalized magnitude expression at the extremum over spatial and temporal scales will be given by

$$\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}\Big|_{(x,y,t)=(0,0,0),s=\hat{s},\tau=\hat{\tau}} = \frac{\left(q^2 \tau_0\right)^{\frac{3q^2}{2\left(q^2+1\right)}}}{4\sqrt{2}\pi^{3/2} s_0 \left(\left(q^2+1\right)\tau_0\right)^{3/2}}, \tag{52}$$

where specifically the choice $q = 1$ corresponds to

$$\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}\Big|_{(x,y,t)=(0,0,0),s=\hat{s},\tau=\hat{\tau}} = \frac{1}{16\pi^{3/2} s_0 \tau_0^{3/4}}. \tag{53}$$

If we additionally renormalize the original Gaussian blink to having maximum value equal to C

$$\begin{aligned} f(x,y,t) &= C\,(2\pi)^{3/2} s_0 \sqrt{\tau_0}\, g(x,y;\ s_0)\, g(t;\ \tau_0) \\ &= C\, \mathrm{e}^{-(x^2+y^2)/2s_0}\, \mathrm{e}^{-t^2/2\tau_0}, \end{aligned} \tag{54}$$

then the magnitude value at the extremum over spatio-temporal scales will instead be given by

$$\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}\Big|_{(x,y,t)=(0,0,0),s=\hat{s},\tau=\hat{\tau}} = \frac{C\sqrt{\tau_0}\left(q^2\tau_0\right)^{\frac{3q^2}{2\left(q^2+1\right)}}}{2\left(\left(q^2+1\right)\tau_0\right)^{3/2}}, \tag{55}$$

where specifically the choice $q = 1$ corresponds to

$$\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}\Big|_{(x,y,t)=(0,0,0),s=\hat{s},\tau=\hat{\tau}} = \frac{C}{4\sqrt{2}\tau_0^{1/4}} \tag{56}$$

and implying that if we want to compare responses between different spatio-temporal scale levels, we should consider the following post-normalized magnitude measure defined to achieve scale-invariant magnitude responses over both spatial and temporal scales

$$\begin{aligned} &\nabla^2_{(x,y),\text{norm}} L_{tt,\text{postnorm}} \\ &\quad = \tau^{\frac{2-q^2}{2(q^2+1)}}\ \nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}\Big|_{\gamma_s=1,\gamma_\tau=\frac{3q^2}{2(q^2+1)}} \\ &\quad = s\tau\left(L_{xxtt} + L_{yytt}\right). \end{aligned} \tag{57}$$

4.2 The Spatial Laplacian of the First-Order Temporal Derivative

For the spatial Laplacian of the first-order temporal derivative, the corresponding scale-normalized expression is for general values of the spatial and temporal scale normalization parameters γ_s and γ_τ given by

$$\begin{aligned} \nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}} &= s^{\gamma_s}\, \tau^{\gamma_\tau/2}\, \nabla^2_{(x,y)} L_t \\ &= s^{\gamma_s}\, \tau^{\gamma_\tau/2}\, \left(L_{xxt} + L_{yyt}\right), \end{aligned} \tag{58}$$

which can be seen as an idealized functional model of a so-called “non-lagged” LGN neurone (compare with [69, Figure 24, left column]). This operator can be expected to give a strong response when both the spatial Laplacian and the first-order temporal derivative give strong responses, e.g., for onset and offset blobs.

Consider a spatio-temporal image pattern defined as a Gaussian onset blob with spatial extent s_0 and temporal duration τ_0:

$$\begin{aligned} &f(x,y,t) \\ &\quad = g(x,y;\ s_0) \int_{u=0}^{t} g(u;\ \tau_0)\,\mathrm{d}u \\ &\quad = \frac{1}{(2\pi)^{3/2} s_0 \sqrt{\tau_0}}\, \mathrm{e}^{-(x^2+y^2)/2s_0} \int_{u=0}^{t} \mathrm{e}^{-u^2/2\tau_0}\,\mathrm{d}u. \end{aligned} \tag{59}$$

By spatial smoothing with the two-dimensional spatial Gaussian kernel and temporal smoothing with the non-causal one-dimensional Gaussian kernel, the resulting spatio-temporal scale-space representation will be of the form

$$L(x, y, t;\ s, \tau) = g(x, y;\ s_0 + s) \int_{u=0}^{t} g(u;\ \tau_0 + \tau)\, \mathrm{d}u, \tag{60}$$

for which the scale-normalized spatial Laplacian of the second-order temporal derivative at the origin $(x, y, t) = (0, 0, 0)$ is given by

$$\nabla^2_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}}\Big|_{(0,0,0)} = -\frac{s^{\gamma_s} \tau^{\gamma_\tau/2}}{\sqrt{2}\pi^{3/2}(s_0 + s)^2 \sqrt{\tau_0 + \tau}}. \tag{61}$$

Differentiating this expression with respect to the spatial scale parameter s and the temporal scale parameter τ and setting the derivative to zero implies that the local extremum over spatial and temporal scales is given by

$$\hat{s} = \frac{\gamma_s s_0}{2 - \gamma_s}, \tag{62}$$

$$\hat{\tau} = \frac{\gamma_\tau \tau_0}{1 - \gamma_\tau}. \tag{63}$$

Requiring the spatial and temporal scale estimates to reflect the spatial and temporal extent of the Gaussian onset blob according to

$$\hat{s} = s_0, \tag{64}$$

$$\hat{\tau} = q^2 \tau_0, \tag{65}$$

implies that we should calibrate the scale normalization parameters γ_s and γ_τ according to

$$\gamma_s = 1, \tag{66}$$

$$\gamma_\tau = \frac{q^2}{q^2 + 1}, \tag{67}$$

where specifically the choice $q = 1$ corresponds to $\gamma_\tau = 1/2$. For these values of γ_s and γ_τ, the scale-normalized magnitude expression at the extremum over spatial and temporal scales will be given by

$$\nabla^2_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}}\Big|_{(x,y,t)=(0,0,0), s=\hat{s}, \tau=\hat{\tau}} = -\frac{\left(q^2 \tau_0\right)^{\frac{q^2}{2q^2+2}}}{4\sqrt{2}\pi^{3/2} s_0 \sqrt{\left(q^2 + 1\right)\tau_0}}, \tag{68}$$

where specifically the case $q = 1$ corresponds to

$$\nabla^2_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}}\Big|_{(x,y,t)=(0,0,0), s=\hat{s}, \tau=\hat{\tau}} = -\frac{1}{8\pi^{3/2} s_0 \sqrt[4]{\tau_0}}. \tag{69}$$

If we additionally renormalize the original Gaussian onset blob to having maximum value equal to C

$$f(x, y, t) = 2\pi\, C\, s_0\, g(x, y;\ s_0)\, g(t;\ \tau_0) = \frac{C}{\sqrt{2\pi}}\, \mathrm{e}^{-(x^2+y^2)/2s_0} \int_{u=0}^{t} \mathrm{e}^{-u^2/2\tau_0}\, du, \tag{70}$$

then the magnitude value at the extremum over spatio-temporal scales will instead be given by

$$\nabla^2_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}}\Big|_{(x,y,t)=(0,0,0), s=\hat{s}, \tau=\hat{\tau}} = \frac{C\ \left(q^2 \tau_0\right)^{\frac{q^2}{2q^2+2}}}{2\sqrt{2\pi}\sqrt{\left(q^2 + 1\right)\tau_0}}, \tag{71}$$

where specifically the case $q = 1$ corresponds to

$$\nabla^2_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}}\Big|_{(x,y,t)=(0,0,0), s=\hat{s}, \tau=\hat{\tau}} = -\frac{C}{4\sqrt{\pi}\sqrt[4]{\tau_0}} \tag{72}$$

and implying that if we want to compare responses between different spatio-temporal scale levels, we should consider the following post-normalized magnitude measure to achieve scale-invariant magnitude responses over both spatial and temporal scales

$$\begin{aligned} &\nabla^2_{(x,y),\mathrm{postnorm}} L_{t,\mathrm{postnorm}} \\ &= \tau^{\frac{1}{2(q^2+1)}}\ \nabla^2_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}}\Big|_{\gamma_s=1, \gamma_\tau=\frac{q^2}{q^2+1}} \\ &= s\sqrt{\tau}\left(L_{xxt} + L_{yyt}\right). \end{aligned} \tag{73}$$

4.3 The Determinant of the Spatial Hessian Matrix Applied to the Second-Order Temporal Derivative

Inspired by the way the determinant of the spatial Hessian matrix constitutes a better spatial interest point detector than the spatial Laplacian operator [74], we consider an extension of the spatial Laplacian of the second-order temporal derivative (42) into the determinant of the spatial Hessian applied to the second-order temporal derivative

$$\begin{aligned} \det \mathcal{H}_{(x,y),\mathrm{norm}} L_{tt,\mathrm{norm}} &= s^{2\gamma_s}\, \tau^{2\gamma_\tau}\ \det \mathcal{H}_{(x,y)} L_{tt} \\ &= s^{2\gamma_s}\, \tau^{2\gamma_\tau}\ \left(L_{xxtt}\, L_{yytt} - L_{xytt}^2\right). \end{aligned} \tag{74}$$

This operator can be expected to give a strong response when both the second-order temporal derivative and the determinant of the spatial Hessian give strong responses, e.g., when there are strong second-order temporal variations in combination with simultaneously strong spatial variations in two orthogonal spatial directions, such as for blinking blobs or corners.

When applied to a Gaussian blink of the form (43) having a spatio-temporal scale-space representation of the form (44), the scale-normalized determinant of the spatio-temporal Hessian at the origin then assumes the form

$$\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}\big|_{(0,0,0)} \frac{s^{2\gamma_s} \tau^{2\gamma_\tau}}{8\pi^3 (s+s_0)^4 (\tau+\tau_0)^3} \quad (75)$$

and assumes its extremum over spatial and temporal scales at

$$\hat{s} = \frac{\gamma_s s_0}{2-\gamma_s}, \quad (76)$$

$$\hat{\tau} = \frac{2\gamma_\tau \tau_0}{3-2\gamma_\tau}. \quad (77)$$

If we require the spatial and temporal scale estimates to reflect the spatial and temporal extent of the Gaussian blink according to $\hat{s} = s_0$ and $\hat{\tau} = q^2 \tau_0$, then this implies that we should calibrate the scale normalization parameters γ_s and γ_τ according to

$$\gamma_s = 1, \quad (78)$$

$$\gamma_\tau = \frac{3q^2}{2\left(q^2+1\right)}, \quad (79)$$

where specifically the choice $q = 1$ corresponds to $\gamma_\tau = 3/4$. For these values of γ_s and γ_τ, the scale-normalized magnitude expression at the extremum over spatial and temporal scales will be given by

$$\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}\big|_{(0,0,0)} = \frac{\left(q^2 \tau_0\right)^{\frac{3q^2}{q^2+1}}}{128\pi^3 \left(q^2+1\right)^3 s_0^2 \tau_0^3}, \quad (80)$$

where specifically the choice $q = 1$ corresponds to

$$\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}\big|_{(0,0,0)} = \frac{1}{1024\pi^3 s_0^2 \tau_0^{3/2}}. \quad (81)$$

If we additionally renormalize the original Gaussian blink to having maximum value equal to C according to (54), then the magnitude value at the extremum over spatio-temporal scales will instead be given by

$$\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}\big|_{(0,0,0)} = \frac{C^2 \left(q^2 \tau_0\right)^{\frac{3q^2}{q^2+1}}}{16 \left(q^2+1\right)^3 \tau_0^2}, \quad (82)$$

where specifically the choice $q = 1$ corresponds to

$$\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}\big|_{(0,0,0)} = \frac{C^2}{128\sqrt{\tau_0}} \quad (83)$$

and implying that if we want to compare responses between different spatio-temporal scale levels, we should consider the following post-normalized magnitude measure to achieve scale invariance over both spatial and temporal scales

$$\begin{aligned} &\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}} \\ &= \tau^{\frac{2(2-q^2)}{q^2+1}} \det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}\big|_{\gamma_s=1, \gamma_\tau=\frac{3q^2}{2(q^2+1)}} \\ &= s^2\, \tau^2 \left(L_{xxtt}\, L_{yytt} - L_{xytt}^2\right). \end{aligned} \quad (84)$$

4.4 The Determinant of the Spatial Hessian Matrix Applied to the First-Order Temporal Derivative

Analogously to the determinant of the spatial Hessian applied to the second-order temporal derivative, we can also apply the determinant of the spatial Hessian to the first-order temporal derivative

$$\begin{aligned} \det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}} &= s^{2\gamma_s}\, \tau^{\gamma_\tau} \det \mathcal{H}_{(x,y)} L_t \\ &= s^{2\gamma_s}\, \tau^{\gamma_\tau} \left(L_{xxt} L_{yyt} - L_{xyt}^2\right). \end{aligned} \quad (85)$$

This operator can be expected to give a strong response when both the first-order temporal derivative and the determinant of the spatial Hessian give strong responses, e.g., when there are strong first-order temporal variations in combination with simultaneously strong spatial variations in two orthogonal spatial directions, such as for onset or offsets blobs or corners.

When applied to an onset Gaussian blob of the form (59) having a spatio-temporal scale-space representation of the form (60), the first-order temporal derivative of the determinant of the spatial Hessian at the origin then assumes the form

$$\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}\big|_{(0,0,0)} = \frac{s^{2\gamma_s} \tau^{\gamma_\tau}}{8\pi^3 (s+s_0)^4 (\tau+\tau_0)} \quad (86)$$

and assumes its extremum over spatial and temporal scales at

$$\hat{s} = \frac{\gamma_s s_0}{2-\gamma_s}, \quad (87)$$

$$\hat{\tau} = \frac{\gamma_\tau \tau_0}{1 - \gamma_\tau}. \tag{88}$$

If we require the spatial and temporal scale estimates to reflect the spatial and temporal extent of the Gaussian onset blob according to $\hat{s} = s_0$ and $\hat{\tau} = q^2 \tau_0$, then this implies that we should calibrate the scale normalization parameters γ_s and γ_τ according to

$$\gamma_s = 1, \tag{89}$$

$$\gamma_\tau = \frac{q^2}{q^2 + 1}, \tag{90}$$

where specifically the choice $q = 1$ corresponds to $\gamma_\tau = 1/2$. For these values of γ_s and γ_τ, the scale-normalized magnitude expression at the extremum over spatial and temporal scales will be given by

$$\det \mathcal{H}_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}}\big|_{(0,0,0)} = \frac{\left(q^2 \tau_0\right)^{\frac{q^2}{q^2+1}}}{128\pi^3 \left(q^2+1\right) s_0^2 \tau_0}, \tag{91}$$

where specifically the choice $q = 1$ corresponds to

$$\det \mathcal{H}_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}}\big|_{(0,0,0)} = \frac{1}{256\pi^3 s_0^2 \sqrt{\tau_0}}. \tag{92}$$

If we additionally renormalize the original Gaussian onset blob to having maximum value equal to C according to (70), then the magnitude value at the extremum over spatio-temporal scales will instead be given by

$$\det \mathcal{H}_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}}\big|_{(0,0,0)} = \frac{C^2 \left(q^2 \tau_0\right)^{\frac{q^2}{q^2+1}}}{32\pi q^2 \tau_0 + 32\pi \tau_0}, \tag{93}$$

where specifically the choice $q = 1$ corresponds to

$$\det \mathcal{H}_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}}\big|_{(0,0,0)} = \frac{C^2}{64\pi \sqrt{\tau_0}} \tag{94}$$

and implying that if we want to compare responses between different spatio-temporal scale levels, we should consider the following post-normalized magnitude measure to achieve scale invariance over both spatial and temporal scales

$$\begin{aligned} &\det \mathcal{H}_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}} \\ &= \tau^{\frac{q^2+2}{2(q^2+1)}} \det \mathcal{H}_{(x,y),\mathrm{norm}} L_{t,\mathrm{norm}}\big|_{\gamma_s=1,\gamma_\tau=\frac{q^2}{q^2+1}} \\ &= s^2 \tau \left(L_{xxt} L_{yyt} - 2L_{xyt}^2\right). \end{aligned} \tag{95}$$

4.5 The Determinant of the Spatio-Temporal Hessian Matrix

For general values of the spatial and temporal scale normalization parameters γ_s and γ_τ, the scale-normalized determinant of the spatio-temporal Hessian is given by

$$\begin{aligned} &\det \mathcal{H}_{(x,y,t),\mathrm{norm}} L \\ &= s^{2\gamma_s} \tau^{\gamma_\tau} \left(L_{xx} L_{yy} L_{tt} + 2L_{xy} L_{xt} L_{yt}\right. \\ &\qquad \left. - L_{xx} L_{yt}^2 - L_{yy} L_{xt}^2 - L_{tt} L_{xy}^2\right). \end{aligned} \tag{96}$$

This operator can be expected to give strong responses when there are simultaneously strong second-order variations in three strongly different directions in joint space–time.

When applied to a Gaussian blink of the form (43) having a spatio-temporal scale-space representation of the form (44), the scale-normalized determinant of the spatio-temporal Hessian at the origin then assumes the form

$$\begin{aligned} &\det(\mathcal{H}_{(x,y,t),\mathrm{norm}} L)\big|_{(0,0,0)} \\ &= -\frac{s^{2\gamma_s} \tau^{\gamma_\tau}}{16\sqrt{2}\pi^{9/2} (s + s_0)^5 (\tau + \tau_0)^{5/2}} \end{aligned} \tag{97}$$

and assumes its extremum over spatial and temporal scales at

$$\hat{s} = \frac{2\gamma_s s_0}{5 - 2\gamma_s}, \tag{98}$$

$$\hat{\tau} = \frac{2\gamma_\tau \tau_0}{5 - 2\gamma_\tau}. \tag{99}$$

Requiring the spatial and temporal scale estimates to reflect the spatial and temporal extent of the Gaussian blink according to $\hat{s} = s_0$ and $\hat{\tau} = q^2 \tau_0$ implies that we should calibrate the scale normalization parameters γ_s and γ_τ according to

$$\gamma_s = \frac{5}{4}, \tag{100}$$

$$\gamma_\tau = \frac{5q^2}{2(q^2 + 1)}, \tag{101}$$

where specifically the choice $q = 1$ corresponds to $\gamma_\tau = 5/4$. For these values of γ_s and γ_τ, the scale-normalized magnitude expression at the extremum over spatial and temporal scales is given by

$$\begin{aligned} &\det(\mathcal{H}_{(x,y,t),\mathrm{norm}} L)\big|_{(0,0,0)} \\ &= -\frac{\left(q^2 \tau_0\right)^{\frac{5q^2}{2(q^2+1)}}}{512\sqrt{2}\pi^{9/2} s_0^{5/2} \left(\left(q^2+1\right) \tau_0\right)^{5/2}}, \end{aligned} \tag{102}$$

where specifically the choice $q = 1$ corresponds to

$$\det(\mathcal{H}_{(x,y,t),\text{norm}} L)\big|_{(0,0,0)} = -\frac{1}{4096\pi^{9/2} s_0^{5/2} \tau_0^{5/4}}. \tag{103}$$

If we additionally renormalize the original Gaussian blink to having maximum value equal to C according to (54), then the magnitude value at the extremum over spatio-temporal scales will instead be given by

$$\det(\mathcal{H}_{(x,y,t),\text{norm}} L)\big|_{(0,0,0)} = -\frac{C^3 \sqrt{s_0}\, \tau_0^{3/2} \left(q^2 \tau_0\right)^{\frac{5q^2}{2(q^2+1)}}}{32 \left(\left(q^2+1\right) \tau_0\right)^{5/2}}, \tag{104}$$

where specifically the choice $q = 1$ corresponds to

$$\det(\mathcal{H}_{(x,y,t),\text{norm}} L)\big|_{(0,0,0)} = -\frac{C^3 \sqrt{s_0} \sqrt[4]{\tau_0}}{128\sqrt{2}} \tag{105}$$

and implying that if we want to compare responses between different spatio-temporal scale levels, we should consider the following post-normalized magnitude measure to achieve scale invariance over both spatial and temporal scales

$$\begin{aligned}
&\det(\mathcal{H}_{(x,y,t),\text{postnorm}} L) \\
&= \frac{\tau^{\frac{2-3q^2}{2(q^2+1)}}}{\sqrt{s}} \det(\mathcal{H}_{(x,y,t),\text{norm}} L)\big|_{\gamma_s = \frac{5}{4}, \gamma_\tau = \frac{5q^2}{2(q^2+1)}} \\
&= s^2 \tau \Big(L_{xx} L_{yy} L_{tt} + 2 L_{xy} L_{xt} L_{yt} \\
&\qquad - L_{xx} L_{yt}^2 - L_{yy} L_{xt}^2 - L_{tt} L_{xy}^2\Big).
\end{aligned} \tag{106}$$

In view of these results, it is illuminating to compare to the analysis by Willems et al. [122], who defined a scale-normalized determinant of the Hessian corresponding to (96) based on $\gamma_s = 1$ and $\gamma_\tau = 1$, which in turn implies that the spatial and temporal scale estimates were instead given by

$$\hat{s} = \frac{2}{3} s_0, \tag{107}$$

$$\hat{\tau} = \frac{2}{3} \tau_0. \tag{108}$$

If we would like the features to be detected at the scales at which they occur, such that $\hat{s} = s_0$ and $\hat{\tau} = \tau_0$, we should, however, instead choose the scale normalization powers γ_s and γ_τ according to (100) and (101) for $q = 1$, so that we achieve maximum similarity between the response property of the spatio-temporal feature detector in relation to the spatio-temporal features we would like to detect. If using a lower value of the parameter $q < 1$, then this property is sacrificed for the possible gain of obtaining faster temporal responses in a time-causal implementation, where otherwise the detection of image features at coarser temporal scales implies longer temporal delays (compare with Sect. 2.2). Over the spatial domain or over a non-causal temporal domain as used in the original work by Willems et al. [122], it should, however, from signal detection theory be better to calibrate the method such that $\hat{s} = s_0$ and $\hat{\tau} = \tau_0$. Notwithstanding the potential gain of achieving a shorter temporal delay by using a lower value of $q < 1$, from a signal detection theory background there should be no motivation to calibrate the feature detector to choosing finer spatial scale levels than s_0.

4.6 The Second-Order Temporal Derivative of the Determinant of the Spatial Hessian Matrix

When using the spatial Laplacian operator over the spatial domain as a basis for defining spatio-temporal interest operators, the spatial Laplacian does because of its linearity commute with the first- and second-order temporal derivatives. Thereby, the spatial Laplacian of the second-order temporal derivative is equal to the second-order temporal derivative of the spatial Laplacian. When replacing the Laplacian interest operator in the spatio-temporal interest operator $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$ by the determinant of the spatial Hessian, an alternative possibility to considering the determinant of the second-order temporal derivative $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$ is therefore to consider the second-order temporal derivative of the determinant of the spatial Hessian

$$\begin{aligned}
&\partial_{tt,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L) \\
&= s^{2\gamma_s} \tau^{\gamma_\tau}\, \partial_{tt}(\det \mathcal{H}_{(x,y)} L) \\
&= s^{2\gamma_s} \tau^{\gamma_\tau} \Big(L_{xxtt} L_{yy} + 2 L_{xxt} L_{yyt} + L_{xx} L_{yytt} \\
&\qquad - 2 L_{xyt}^2 - 2 L_{xy} L_{xytt}\Big).
\end{aligned} \tag{109}$$

This operator can be expected to give strong responses when the spatial slice of joint space–time contains strong second-order variations on two orthogonal spatial directions, and this structure in turn also leads to strong second-order temporal variations as time evolves.

When applied to a Gaussian blink of the form (43) having a spatio-temporal scale-space representation of the form (44), the scale-normalized determinant of the spatio-temporal Hessian at the origin then assumes the form

$$\begin{aligned}
&\partial_{tt,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{(0,0,0)} \\
&= -\frac{s^{2\gamma_s} \tau^{\gamma_\tau}}{4\pi^3 (s+s_0)^4 (\tau+\tau_0)^2}
\end{aligned} \tag{110}$$

and assumes its extremum over spatial and temporal scales at

$$\hat{s} = \frac{\gamma_s s_0}{2 - \gamma_s}, \tag{111}$$

$$\hat{\tau} = \frac{\gamma_\tau \tau_0}{2 - \gamma_\tau}. \tag{112}$$

If we require the spatial and temporal scale estimates to reflect the spatial and temporal extent of the Gaussian blink according to $\hat{s} = s_0$ and $\hat{\tau} = q^2 \tau_0$, then this implies that we should calibrate the scale normalization parameters γ_s and γ_τ according to

$$\gamma_s = 1, \tag{113}$$

$$\gamma_\tau = \frac{2q^2}{q^2 + 1}, \tag{114}$$

where specifically the choice $q = 1$ corresponds to $\gamma_\tau = 1$. For these values of γ_s and γ_τ, the scale-normalized magnitude expression at the extremum over spatial and temporal scales will be given by

$$\partial_{tt,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{(0,0,0)} = -\frac{\left(q^2 \tau_0\right)^{\frac{2q^2}{q^2+1}}}{64\pi^3 \left(q^2+1\right)^2 s_0^2 \tau_0^2}, \tag{115}$$

where specifically the choice $q = 1$ corresponds to

$$\partial_{tt,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{(0,0,0)} = -\frac{1}{256\pi^3 s_0^2 \tau_0}. \tag{116}$$

If we additionally renormalize the original Gaussian blink to having maximum value equal to C according to (54), then the magnitude value at the extremum over spatio-temporal scales will instead be given by

$$\partial_{tt,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{(0,0,0)} = -\frac{C^2 \left(q^2 \tau_0\right)^{\frac{2q^2}{q^2+1}}}{8\left(q^2+1\right)^2 \tau_0}, \tag{117}$$

where specifically the choice $q = 1$ corresponds to

$$\partial_{tt,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{(0,0,0)} = -\frac{C^2}{32} \tag{118}$$

and implying that if we want to compare responses between different spatio-temporal scale levels, we should consider the following post-normalized magnitude measure to achieve scale invariance over both spatial and temporal scales

$$\begin{aligned} &\partial_{tt,\text{postnorm}}(\det \mathcal{H}_{(x,y),\text{postnorm}} L) \\ &= \tau^{\frac{1-q^2}{1+q^2}} \, \partial_{tt,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{\gamma_s=1,\gamma_\tau=\frac{2q^2}{1+q^2}} \\ &= s^2 \tau \left(L_{xxtt} L_{yy} + 2 L_{xxt} L_{yyt} + L_{xx} L_{yytt} \right. \\ &\qquad \left. - 2 L_{xyt}^2 - 2 L_{xy} L_{xytt} \right). \end{aligned} \tag{119}$$

4.7 The First-Order Temporal Derivative of the Determinant of the Spatial Hessian Matrix

Analogously to the second-order temporal derivative of the determinant of the spatial Hessian, we can also define the first-order temporal derivative of the determinant of the spatial Hessian

$$\begin{aligned} &\partial_{t,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L) \\ &= s^{2\gamma_s} \, \tau^{\gamma_\tau/2} \, \partial_t(\det \mathcal{H}_{(x,y)} L) \\ &= s^{2\gamma_s} \, \tau^{\gamma_\tau/2} \, \left(L_{xxt} L_{yy} + L_{xx} L_{yyt} - 2 L_{xy} L_{xyt} \right). \end{aligned} \tag{120}$$

This operator can be expected to give strong responses when the spatial slice of joint space–time contains strong second-order variations on two orthogonal spatial directions, and this structure in turn also leads to strong first-order temporal variations as time evolves.

When applied to an onset Gaussian blob of the form (59) having a spatio-temporal scale-space representation of the form (60), the first-order temporal derivative of the determinant of the spatial Hessian at the origin then assumes the form

$$\begin{aligned} &\partial_{t,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{(0,0,0)} \\ &= \frac{s^{2\gamma_s} \tau^{\gamma_\tau/2}}{4\sqrt{2}\pi^{5/2}(s+s_0)^4 \sqrt{\tau+\tau_0}} \end{aligned} \tag{121}$$

and assumes its extremum over spatial and temporal scales at

$$\hat{s} = \frac{\gamma_s s_0}{2 - \gamma_s}, \tag{122}$$

$$\hat{\tau} = \frac{\gamma_\tau \tau_0}{1 - \gamma_\tau}. \tag{123}$$

If we require the spatial and temporal scale estimates to reflect the spatial and temporal extent of the Gaussian onset blob according to $\hat{s} = s_0$ and $\hat{\tau} = q^2 \tau_0$, then this implies that we should calibrate the scale normalization parameters γ_s and γ_τ according to

$$\gamma_s = 1, \tag{124}$$

$$\gamma_\tau = \frac{q^2}{q^2 + 1}, \tag{125}$$

where specifically the choice $q = 1$ corresponds to $\gamma_\tau = 1/2$. For these values of γ_s and γ_τ, the scale-normalized magnitude expression at the extremum over spatial and temporal scales will be given by

$$\partial_{t,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{(0,0,0)}$$
$$= \frac{\left(q^2 \tau_0\right)^{\frac{q^2}{2q^2+2}}}{64\sqrt{2}\pi^{5/2} s_0^2 \sqrt{\left(q^2+1\right)\tau_0}}, \tag{126}$$

where specifically the choice $q = 1$ corresponds to

$$\partial_{t,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{(0,0,0)} = \frac{1}{128\pi^{5/2} s_0^2 \sqrt[4]{\tau_0}}. \tag{127}$$

If we additionally renormalize the original Gaussian onset blob to having maximum value equal to C according to (70), then the magnitude value at the extremum over spatio-temporal scales will instead be given by

$$\partial_{t,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{(0,0,0)} = \frac{C^2 \left(q^2 \tau_0\right)^{\frac{q^2}{2q^2+2}}}{16\sqrt{2\pi}\sqrt{\left(1+q^2\right)\tau_0}}, \tag{128}$$

where specifically the choice $q = 1$ corresponds to

$$\partial_{t,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{(0,0,0)} = \frac{C^2}{32\sqrt{\pi}\sqrt[4]{\tau_0}} \tag{129}$$

and implying that if we want to compare responses between different spatio-temporal scale levels, we should consider the following post-normalized magnitude measure to achieve scale invariance over both spatial and temporal scales

$$\partial_{t,\text{postnorm}}(\det \mathcal{H}_{(x,y),\text{postnorm}} L)$$
$$= \tau^{\frac{1}{2(q^2+1)}} \; \partial_{t,\text{norm}}(\det \mathcal{H}_{(x,y),\text{norm}} L)\big|_{\gamma_s=1,\gamma_\tau=\frac{q^2}{q^2+1}}$$
$$= s^2\sqrt{\tau}\left(L_{xxt}L_{yy} + L_{xx}L_{yyt} - 2L_{xy}L_{xyt}\right). \tag{130}$$

4.8 The Spatio-Temporal Laplacian

If aiming at defining a spatio-temporal analogue of the Laplacian operator, one does, however, need to consider that the most straightforward way of defining such an operator

$$\nabla^2_{(x,y,t)} L = L_{xx} + L_{yy} + L_{tt} \tag{131}$$

is not covariant under independent scaling transformations of the spatial and temporal domains as occurs if observing the same scene with cameras having independently different spatial and temporal sampling rates. Therefore, if attempting to define a spatio-temporal analogue of the Laplacian of the Gaussian operator, one could in principle consider introducing an arbitrary scaling factor $\varkappa^2$ between the temporal versus the spatial derivatives

$$\nabla^2_{(x,y,t)} L = L_{xx} + L_{yy} + \varkappa^2 L_{tt}. \tag{132}$$

This operator can be expected to give strong response when there is strong second-order variation in at least one spatial dimension or in the temporal dimension. It is, however, not necessary that that there are simultaneous strong variations over both space and time, implying that this operator cannot be expected to be as selective as the other seven spatio-temporal interest point detectors studied above.

With the previously introduced recipe of replacing spatial and temporal derivatives by corresponding scale-normalized derivatives, the corresponding scale-normalized expression then becomes

$$\nabla^2_{(x,y,t),\text{norm}} L = s^{\gamma_s}(L_{xx} + L_{yy}) + \varkappa^2 \tau^{\gamma_\tau} L_{tt}, \tag{133}$$

which, however, is not within the family of spatio-temporal differential invariants (18) guaranteed to lead to scale-covariant spatio-temporal scale selection.

When applied to a Gaussian blink of the form (43) having a spatio-temporal scale-space representation of the form (44), the scale-normalized spatio-temporal Laplacian at the origin then assumes the form

$$\nabla^2_{(x,y,t),\text{norm}} L\Big|_{(0,0,0)} = \frac{-2s^{\gamma_s}(\tau+\tau_0) - \varkappa^2\tau^{\gamma_\tau}(s+s_0)}{2\sqrt{2}\pi^{3/2}(s+s_0)^2(\tau+\tau_0)^{3/2}}. \tag{134}$$

Unfortunately, the algebraic equations that determine the spatial and temporal scale estimates as function of s_0 and τ_0

$$\begin{aligned} &-2(\gamma_s-2)s^{\gamma_s+1}(\tau+\tau_0) - 2\gamma_s s_0 s^{\gamma_s}(\tau+\tau_0) \\ &\quad + s^2\varkappa^2\tau^{\gamma_\tau} + s s_0 \varkappa^2 \tau^{\gamma_\tau} = 0, \end{aligned} \tag{135}$$

$$\begin{aligned} &2\tau s^{\gamma_s}(\tau+\tau_0) - s\varkappa^2\tau^{\gamma_\tau}((2\gamma_\tau-3)\tau + 2\gamma_\tau\tau_0) \\ &\quad - s_0\varkappa^2\tau^{\gamma_\tau}((2\gamma_\tau-3)\tau + 2\gamma_\tau\tau_0) = 0, \end{aligned} \tag{136}$$

are hard to solve for general values of the scale normalization parameters γ_s and γ_τ. By solving these equations in the specific case of $\gamma_s = 1$ and $\gamma_\tau = 1$, we can, however, note that the resulting scale estimates

$$\hat{s} = s_0\left(\frac{5}{2+\varkappa^2} - 1\right), \tag{137}$$

$$\hat{\tau} = 2\tau_0\left(2 - \frac{5}{2+\varkappa^2}\right), \tag{138}$$

will be explicitly dependent on the relative scaling factor $\varkappa^2$ between the derivatives with respect to the temporal versus the spatial domains. This situation is in clear contrast to the previously considered spatio-temporal differential invariants for spatio-temporal scale selection: (i)–(ii) the spatial Laplacian of the first- and second-order temporal derivatives (58), (iii)–(iv) the determinant of the Hessian applied to the first- and second-order temporal derivatives (58) and (42), (v) the determinant of the spatio-temporal Hessian (96) or (vi)–(vii) the first- and second-order temporal derivatives of the determinant of the spatial Hessian (109) and (120), for which a corresponding multiplication of the temporal derivative operator by a temporal rescaling factor $\varkappa$ does not affect the spatio-temporal scale estimates.

The underlying theoretical reason for this lack of spatial and temporal scale invariance is that the attempt to define a spatio-temporal Laplacian operator according to (132) is not covariant under independent rescaling transformations of the spatial and temporal domains. The spatial Laplacian of the first- and second-order temporal derivatives, the determinant of the Hessian of the first- and second-order temporal derivatives and the determinant of the spatio-temporal Hessian are on the other hand truly covariant under such independent relative scaling transformations of the spatial and temporal domains.

The corresponding magnitude estimate at the extremum over spatio-temporal scales is for $\gamma_s = 1$ and $\gamma_\tau = 1$ given by

$$\nabla^2_{(x,y,t),\text{norm}} L\Big|_{(0,0,0)} = \frac{3\sqrt{\frac{3}{10}}\,(2+\varkappa^2)}{25\pi^{3/2} s_0 \sqrt{\tau_0}}. \tag{139}$$

If we additionally renormalize the original Gaussian blink to having maximum value equal to C according to (54), then the magnitude value at the extremum over spatio-temporal scales will instead be given by

$$\nabla^2_{(x,y,t),\text{norm}} L\Big|_{(0,0,0)} = -\frac{6}{25}\sqrt{\frac{3}{5}}\left(2+\varkappa^2\right) C \tag{140}$$

and also dependent on the in principle arbitrary relative weighting factor $\varkappa^2$ between the temporal and spatial derivatives.

To illustrate the practical consequence of the lack of spatio-temporal scale covariance for a differential entity used for spatio-temporal scale selection, let us consider two different video cameras that are observing the same scene. Let us for simplicity assume that the sensors in the two video cameras have the same spatial resolution, whereas the temporal resolutions differ by say a factor of two. If we define a spatio-temporal Laplacian operator for each video domain based on the native coordinate system of each respective individual camera, then the spatio-temporal Laplacian operator in the first video domain will correspond to a spatio-temporal Laplacian operator in the second video domain that differs by a factor of two in the value of $\varkappa$. Thus, if we perform spatio-temporal scale selection by detecting local extrema over spatio-temporal scales of the spatio-temporal Laplacian, we will detect extrema in effective spatio-temporal differential expressions that differ between the two video domains. Specifically, this implies that we cannot exactly interrelate the spatio-temporal Laplacian responses between the two domains in the way necessary to carry out a proof of scale invariance for general classes of spatio-temporal image structures. Although the scale estimates could for another form of scale normalization be computed for the specific spatio-temporal image model of a Gaussian blink [49], corresponding scale selection properties are then not guaranteed to generalize to more general spatio-temporal image structures beyond the specific subfamily of image structures for which the scale calibration was performed. Because of the covariance properties of the spatio-temporal differential invariants $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y,t),\text{norm}} L$, $\partial_{t,\text{norm}} (\mathcal{H}_{(x,y),\text{norm}} L)$ and $\partial_{tt,\text{norm}} (\mathcal{H}_{(x,y),\text{norm}} L)$, such interrelations can, however, be carried out for those differential operators between two video domains with undetermined relative scaling factors between the spatial and temporal domains. Consequently, these differential entities are therefore much better for spatio-temporal scale selection than the attempt to define a spatio-temporal Laplacian operator.

Additionally, if one would attempt to rank image features based on the corresponding scale-normalized magnitude measure $\nabla^2_{(x,y,t),\text{norm}} L$, then the relative ranking of the image features could therefore also be different between the two domains of the two video cameras, whereas the corresponding relative ranking of image features is preserved for spatio-temporal scale selection based on the differential invariants $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y,t),\text{norm}} L$, $\partial_{t,\text{norm}} (\mathcal{H}_{(x,y),\text{norm}} L)$ and $\partial_{tt,\text{norm}} (\mathcal{H}_{(x,y),\text{norm}} L)$.

In the spatio-temporal interest point detector proposed in [49], a scale-normalized spatio-temporal Laplacian operator corresponding to the specific choice of $\varkappa = 1$ was indeed used for spatio-temporal scale selection, although with a different form of scale normalization of the form

$$\nabla^2_{(x,y,t),\text{norm}} L = s^a \tau^b (L_{xx} + L_{yy}) + s^c \tau^d L_{tt} \tag{141}$$

for the specific choices of $a = 1, b = 1, c = 1/2$ and $d = 3/4$. In addition to the above-mentioned fundamental limitation of using a spatio-temporal Laplacian operator for spatio-temporal scale selection, by the mixed scale normalization in (141) with the temporal scale parameter τ affecting the spatial derivate expressions L_{xx} and L_{yy} and the spatial scale parameter s affecting the temporal derivative expression L_{tt}, it will, however, not be possible to establish a relation between such spatio-temporal Laplacian operators between different spatio-temporal domains that are affected by independent relative rescalings of the spatial and temporal domains. Specifically, it will therefore not be possible to establish a covariance relation between two such independently rescaled spatio-temporal image domains as would be needed to prove scale covariance of the spatial and temporal scale estimates for general spatio-temporal image structures according to the spatio-temporal scale selection theory in Sect. 3. By these theoretical arguments, we can therefore explain why the scale estimates from the spatial and temporal selection mechanisms in [49] were later empirically found to not be sufficiently robust.

If a scale-normalized spatio-temporal Laplacian operator is to be used for spatio-temporal feature detection anyway, the scale normalization according to (133) should, however, lead to better experimental results than the scale normalization according to (141), since the partial derivates with respects to the different dimensions of space and time in the scale-normalized differential expression (141) are not added in terms of dimensionless scale-normalized differential entities for the given values of a, b, c and d, whereas the partial derivatives with respect to space versus time are added in a dimensionless manner in the scale-normalized differential expression (133) if $\gamma_s = 1$ and $\gamma_\tau = 1$ (and corresponding to $a = 1, b = 0, c = 0$ and $d = 1$ in (141) for the specific choice of $\varkappa = 1$).

4.9 Scale Normalization Powers of Spatio-Temporal Interest Point Detectors

To summarize, the analysis of scale calibration in Sects. 4.1–4.7 shows that the scale normalization powers γ_s and γ_τ for the different spatio-temporal interest point detectors should be determined according to Table 1.

4.10 Relating Magnitude Thresholds Between Different Spatio-Temporal Feature Detectors

By considering the scale-normalized magnitude responses (55), (71), (82), (93), (104) (117) and (128) of the above scale-covariant spatio-temporal feature detectors and applying post-normalization of these entities to make the feature responses fully scale-invariant, we can express relations between their magnitude responses in terms of the contrast C of the spatio-temporal image pattern that gave rise to the feature response according to Table 2. These relations can in turn be used for expressing coarse relations between magnitude thresholds for the different types of spatio-temporal interest operators.

Table 1 Scale normalization powers γ_s and γ_τ as determined from scale calibration of the seven spatio-temporal interest point detectors $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y,t),\text{norm}} L$, $\partial_{t,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ and $\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ that are guaranteed to lead to scale-covariant spatio-temporal scale estimates

$\mathcal{D}L$	γ_s	γ_τ
$\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$	1	$\frac{q^2}{q^2+1}$
$\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$	1	$\frac{3q^2}{2(q^2+1)}$
$\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$	1	$\frac{q^2}{q^2+1}$
$\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$	1	$\frac{3q^2}{2(q^2+1)}$
$\det \mathcal{H}_{(x,y,t),\text{norm}} L$	$\frac{5}{4}$	$\frac{5q^2}{2(q^2+1)}$
$\partial_{t,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$	1	$\frac{q^2}{q^2+1}$
$\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$	1	$\frac{2q^2}{q^2+1}$

Table 2 Relations between magnitude thresholds for seven of the spatio-temporal interest point detectors studied in this paper in terms of a common local contrast parameter C

Magnitude thresholds for spatio-temporal interest operators		
$\mathcal{D}L$	For $q = 1$	For general q
$\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$	$\frac{C}{4\sqrt{\pi}}$	$\frac{C\, q^{\frac{q^2}{q^2+1}}}{2\sqrt{2\pi}\sqrt{q^2+1}}$
$\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$	$\frac{C}{4\sqrt{2}}$	$\frac{C q^{\frac{3q^2}{q^2+1}}}{2(q^2+1)^{3/2}}$
$\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$	$\frac{C^2}{64\pi}$	$\frac{C^2 q^{\frac{2q^2}{q^2+1}}}{32\pi q^2+32\pi}$
$\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$	$\frac{C^2}{128}$	$\frac{C^2 q^{\frac{6q^2}{q^2+1}}}{16(q^2+1)^3}$
$\det \mathcal{H}_{(x,y,t),\text{norm}} L$	$\frac{C^3}{128\sqrt{2}}$	$\frac{C^3\, q^{\frac{5q^2}{q^2+1}}}{32(q^2+1)^{5/2}}$
$\partial_{t,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$	$\frac{C^2}{32\sqrt{\pi}}$	$\frac{C^2\, q^{\frac{q^2}{q^2+1}}}{16\sqrt{2\pi}\sqrt{1+q^2}}$
$\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$	$\frac{C^2}{32}$	$\frac{C^2\, q^{\frac{4q^2}{q^2+1}}}{8(q^2+1)^2}$

5 Spatio-Temporal Interest Points Detected as Spatio-Temporal Scale-Space Extrema Over Space–Time

In this section, we shall use the scale-normalized differential entities $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y,t),\text{norm}} L$, $\partial_{t,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$, $\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ and $\nabla^2_{(x,y,t),\text{norm}} L$ according to (42), (58), (74), (85), (96), (109), (120) and (133) for detecting spatio-temporal interest points. The overall idea of the most basic form of such an algorithm is to simultaneously detect both spatio-temporal points $(\hat{x}, \hat{y}, \hat{t})$ and spatio-temporal scales $(\hat{s}, \hat{\tau})$ at which the scale-normalized differential entity $(\mathcal{D}_{\text{norm}} L)(x, y, t;\ s, \tau)$ simultaneously assumes local extrema with respect to both space–time (x, y, t) and spatio-temporal scales (s, τ).

For the use case of offline analysis of pre-recorded video using a non-causal spatio-temporal scale-space representation, such a spatio-temporal scale-space extrema algorithm could be expressed as a straightforward generalization of the corresponding spatial scale-space extrema algorithm proposed in Lindeberg [65] and summarized on more compact form in "Appendix A". The only major conceptual differences are that: (i) the image data should be expanded over both a spatial and a temporal scale parameter instead of just a spatial scale parameter and (ii) the local comparisons for detecting local extrema should be performed over a $3 \times 3 \times 3 \times 3 \times 3$-neighbourhood over $(x, y, t;\ s, \tau)$ instead of over a $3 \times 3 \times 3$-neighbourhood over $(x, y;\ s)$.

A computational problem when expanding a video sequence over both spatial and temporal scales, however, is that the amount of data may become very large, if expanding the video data into the 5-D spatio-temporal scale-space representation over the spatial domain (x, y), the temporal domain t and the spatio-temporal scale parameters (s, τ). For this reason, we shall instead consider a time-recursive implementation that steps forward over time t and only maintains a much more compact time-recursive memory of past information, as a 4-D representation over the spatial image coordinates (x, y) and the spatio-temporal scale levels (s, τ) at each time moment t. Therefore, the time-recursive implementation avoids expanding the internal memory over the temporal dimension and does also directly apply to a time-causal situation in which the future cannot be accessed.

5.1 Time-Causal and Time-Recursive Algorithm for Spatio-Temporal Scale-Space Extrema Detection

Let us approximate the spatial smoothing operation in the continuous spatio-temporal scale-space representation according to (9) by smoothing with the discrete analogue of the Gaussian kernel over the spatial domain [56]

$$T(n_1, n_2;\ s) = \mathrm{e}^{-2s} I_{n_1}(s)\, I_{n_2}(s), \tag{142}$$

which obeys the semi-group property over spatial scales

$$T(\cdot, \cdot;\ s_1) * T(\cdot, \cdot;\ s_2) = T(\cdot, \cdot;\ s_1 + s_2) \tag{143}$$

and where I_n denotes the modified Bessel functions of integer order [2].

Let us additionally approximate the time-causal limit kernel, which can be described by a cascade of first-order integrators [75, Equation (15)]

$$\partial_t L(t;\ \tau_k) = \frac{1}{\mu_k} \left(L(t;\ \tau_{k-1}) - L(t;\ \tau_k) \right), \tag{144}$$

by a cascade of first-order recursive filters of the form [75, Equation (56)]

$$f_{out}(t) - f_{out}(t-1) = \frac{1}{1+\mu_k} \left(f_{in}(t) - f_{out}(t-1) \right). \tag{145}$$

Assuming that the input video data $f(x, y, t)$ is acquired at spatial scale s_0 and temporal scale τ_0, we can then state a basic algorithm for computing the time-causal and time-recursive spatio-temporal scale-space representation and for detecting spatio-temporal scale-space extrema of scale-normalized differential invariants from it as follows:

1. Determine a set of logarithmically distributed temporal scale levels τ_k and spatial scale levels s_l at which the algorithm is to operate by computing spatio-temporal scale-space representations at these spatio-temporal scales.
2. Compute time constants $\mu_k = (\sqrt{1 + 4 r^2 (\tau_k - \tau_{k-1})} - 1)/2$ according to Lindeberg [75, Equations (58) and (55)] for approximating the time-causal limit kernel by a finite number of recursive filters, where r denotes the frame rate and the temporal scale levels τ_k are given in units of $[\text{seconds}]^2$.
3. Expand the first image frame $f(x, y, 0)$ into its purely spatial scale-space representation $L(x, y, 0;\ s, \tau_0)$ over the spatial scale levels s_l at the finest temporal scale τ_0 using the semi-group property of the discrete analogue of the Gaussian kernel
$$L(\cdot, \cdot, 0;\ s_l, \tau_0) = T(\cdot, \cdot;\ s_l - s_{l-1}) * L(\cdot, \cdot, 0;\ s_{l-1}, \tau_0) \tag{146}$$
with initial condition $L(x, y, 0;\ s_0, \tau_0) = f(x, y, 0)$ at the finest spatial scale s_0.
4. For each temporal scale level τ_k, initiate a temporal buffer for temporal scale-space smoothing at this temporal scale using the purely spatial scale-space representation of the first frame as initial condition $B(x, y, k, l) = L(x, y, 0;\ s_l, \tau_0)$.
5. For each spatial and temporal scale level, initiate a small number of temporal buffers for the nearest past frames. (This number should be equal to the maximum order of temporal differentiation.)
6. Loop forwards over time t (in units of time steps):
   (a) Given a new image frame $f(x, y, t)$, expand this frame into its purely spatial scale-space representation $L(x, y, t;\ s, \tau_0)$ at the finest temporal scale τ_0
$$L(\cdot, \cdot, t;\ s_l, \tau_0) = T(\cdot, \cdot;\ s_l - s_{l-1}) * L(\cdot, \cdot, t;\ s_{l-1}, \tau_0). \tag{147}$$

with initial condition $L(x, y, t;\; s_0, \tau_0) = f(x, y, t)$ at the finest spatial scale s_0.

(b) Loop over the temporal scale levels k in ascending order:

  i. For each spatio-temporal scale level (k, l), perform temporal smoothing according to (with $B(x, y, 0, l) = L(x, y, t;\; s_l, \tau_0)$)

$$B(x, y, k, l) := B(x, y, k, l) + \frac{1}{1 + \mu_k}(B(x, y, k-1, l) - B(x, y, k, l)). \qquad (148)$$

(c) For all temporal and spatial scales, compute temporal derivatives using backward differences over the buffers from past frames.

(c) For all temporal and spatial scales, compute the scale-normalized differential entity $(\mathcal{D}_{\text{norm}} L)(x, y, t;\; s_l, \tau_k)$ at that spatio-temporal scale.

(e) For all points and spatio-temporal scales $(x, y;\; s_l, \tau_k)$ for which the magnitude of the post-normalized differential entity is above a pre-defined threshold

$$|(\mathcal{D}_{\text{postnorm}} L)(x, y, t;\; s_l, \tau_k)| \geq \theta, \qquad (149)$$

and optionally, if using complementary thresholding [74], the sign of a complementary differential expression[3] $\bar{\mathcal{D}} L$ is additionally positive

$$|(\bar{\mathcal{D}} L)(x, y, t;\; s_l, \tau_k)| \geq 0, \qquad (150)$$

determine if the point is either a positive maximum or a negative minimum in comparison with its nearest neighbours over space (x, y), time t, spatial scales s_l and temporal scales τ_k.Because the detection of local extrema over time requires a future reference in the temporal direction, this comparison is not done at the most recent frame but at the nearest past frame.

  i. For each detected scale-space extremum, compute more accurate estimates of its spatio-temporal position $(\hat{x}, \hat{y}, \hat{t})$ and spatio-temporal scale $(\hat{s}, \hat{\tau})$ using parabolic interpolation along each dimension according to Lindeberg [77, Equation (115)]. Do also compensate the magnitude estimates by a magnitude correction factor computed for each dimension.

When detecting local extrema with respect to the spatial, temporal and scale dimensions, we stop performing the comparisons at any point in spatio-temporal scale-space as soon as it can be stated that a spatio-temporal point $(x, y, t;\; s, \tau)$ is neither a local maximum nor a local minimum.

Note specifically that by performing the spatial smoothing in the outer loop over spatio-temporal scales, the computationally more demanding spatial smoothing is performed only once for each spatial scale level, whereas the computationally more efficient temporal smoothing is performed in the inner loop over all combinations of spatial and temporal scales. The algorithm is also inherently parallel over spatio-temporal scale levels and lends itself to parallel implementation over a multi-core architecture.

5.2 Post-filtering of Spatio-Temporal Scale-Space Extrema

Additionally, to handle the different amounts of temporal delay at adjacent temporal scales, which may strongly affect the detection of local extrema over temporal scales by nearest neighbour processing of temporal scales in a time-causal context, we perform a post-filtering step of the spatio-temporal scale-space extrema as an extension of the post-filtering method proposed for temporal scale-space extrema of a purely temporal time-causal scale-space representation as detailed in Lindeberg [77, Section 7.1]:

- To post-filter spatio-temporal scale-space extrema with respect to the nearest finer temporal scale, we introduce buffers for keeping a short-term memory of purely temporal extrema of the scale-normalized differential expression $(\mathcal{D}_{\text{norm}} L)(x, y, t;\; s \tau)$. If a point $(x, y, t;\; s \tau)$ is a local maximum (minimum) over time t, then keep this point in a the buffer of local maxima (minima) as long as the values monotonically decrease (increase) with time to later time moments. When a point has been detected as a candidate for a spatio-temporal scale-space maximum (minimum), check if there are active buffers of local maxima (minima) in a local spatial 3×3-neighbourhood over space at the nearest finer temporal scale. If there is such a maximum (minimum) having a higher (lower) value than the current spatio-temporal scale-space maximum (minimum), then the current point is not allowed to become a scale-space extremum.
- To post-filter spatio-temporal scale-space extrema with respect to the nearest coarser temporal scale, we put a record of the spatio-temporal scale-space extremum in

[3] For example, if performing spatio-temporal interest point detection using the spatial Laplacian operator $\nabla^2 L$ applied to either of the first- or the second-order temporal derivatives L_t or L_{tt}, complementary thresholding can be performed by applying the unsigned Hessian feature strength measure $\mathcal{D}_1 L = L_{xx} L_{yy} - L_{xy}^2 - k\,(L_{xx} + L_{yy})^2$ [74] to either the first- or the second-order temporal derivatives, respectively, for $k \in [0, 1/4[$ with preferred choice of $k \in [0.04, 0.10]$, to suppress multiple responses along elongated image structures over the spatial domain. This implies that complementary thresholding for these Laplacian-based spatio-temporal interest operators should be performed based on $\mathcal{D}_1 L_t = L_{xxt} L_{yyt} - L_{xyt}^2 - k\,(L_{xxt} + L_{yyt})^2 > 0$ or $\mathcal{D}_1 L_{tt} = L_{xxtt} L_{yytt} - L_{xytt}^2 - k\,(L_{xxtt} + L_{yytt})^2 > 0$.

a spatial 3×3-neighbourhood over space at the nearest coarser temporal scale. If the original point was a scale-space maximum (minimum), then the short-term memory is kept active as long as the scale-normalized differential expression $(\mathcal{D}_{\text{norm}} L)(x, y, t;\ s\tau)$ continues to increase (decrease) over time. If the scale-normalized magnitude additionally would increase above the scale-normalized magnitude of the original candidate scale-space extremum, then the original candidate to a scale-space extremum is disregarded.

With these two mechanisms running in parallel to the above time-causal and time-recursive spatio-temporal scale-space extrema detection algorithm, we can compensate for the different temporal delays at adjacent temporal scale levels, which implies that a spatio-temporal event in the world will appear as an extremum earlier in the time-causal scale-space representation at finer temporal scales in relation to the time-causal scale-space representation at coarser temporal scales.

5.3 Experimental Results

Figures 7, 8, 9, 10 and 11 show the result of detecting spatio-temporal scale-space extrema in this way for three video sequences from the UCF-101 dataset [104] and one video sequence from the KITTI dataset [26]. For these experiments, we used 21 spatial scale levels between $\sigma_s = 2$ and 21 pixels and 7 temporal scale levels between $\sigma_\tau = 40$ ms and 2.56 s with seven additional pre-scales and distribution parameter $c = 2$ for the time-causal limit kernel. To obtain comparable numbers of features from the different types of feature detectors, we adapted the thresholds on the scale-normalized differential invariants such that the average number of features from each feature detector was 50 features per frame for the kayaking video, a lower number of 30 features per frame for the videos of the table tennis player and the archer where the background is static, and a larger number of 200 features per frame for the driving scene, where the camera is moving relative to a cluttered environment.

Figure 7 and the first row of Fig. 10 show results computed from the same video of a kayaker as used for the illustrations in Figs. 1, 2, 3 and 4. As can be seen from the results, all eight feature detectors respond to regions in the video sequence where there are strong variations in image intensity over space and time. There are, however, also some qualitative differences between the results from the different spatio-temporal interest point detectors. The LGN-inspired feature detectors $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$ and $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$ respond both to the motion patterns of the paddler and to the spatio-temporal texture corresponding to the waves on the water surface that lead to temporal flickering effects and so do the operators $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$ and $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$. The more corner detector inspired feature detectors $\det \mathcal{H}_{(x,y,t),\text{norm}} L$, $\partial_{t,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ and $\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ respond more to image features where there are simultaneously rapid variations over both of the spatial dimensions and the temporal dimension.

Figure 8 and the second row of Fig. 10 show corresponding results of detecting spatio-temporal scale-space extrema from a video sequence with a table tennis player. Here, we can note that the seven spatio-temporal interest point detectors $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y,t),\text{norm}} L$, $\partial_{t,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ and $\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ do all give rise to rather rich distributions of feature responses corresponding to the motion pattern of the tennis player. (The responses on the left part of the table tennis table are caused by cast shadows of the tennis player from the lamp in the ceiling). The LGN-inspired feature detectors $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$ and $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$ do both specifically generate responses when the ball flies over the net and so do the determinant of the spatio-temporal Hessian as well as the first- and second-order temporal derivatives of the spatial Laplacian. The responses due to the spatio-temporal Laplacian are, however, less specific to specific motion events, and with numerous responses from the almost static background. Incorporating also the theoretical limitations of the spatio-temporal Laplacian described in Sect. 4.8 as well as other limitations that will be described below, we conclude that this operator should therefore not be considered as a suitable feature detector for processing video data.

Figure 9 and the third row of Fig. 10 show the results of detecting corresponding spatio-temporal scale-space extrema from a video sequence with an archer. Here, we can note that the five spatio-temporal interest point detectors $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$ and $\det \mathcal{H}_{(x,y,t),\text{norm}} L$ do in a corresponding manner give rise to rather rich distributions of feature responses corresponding to the motion pattern of the archer. For the determinant of the spatio-temporal Hessian, which operates like a three-dimensional corner detector, there are, however, many more responses along the edges of the archer than for the other four feature detectors. The four feature detectors $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$ based on first- or second-order temporal derivatives do all generate multiple responses when the arrow hits the cloth on the wall. The response of the determinant of the spatio-temporal Hessian is, however, delayed and not as strong as for the other competing spatio-temporal events in the scene.

Figure 11 shows results of applying these spatio-temporal scale extrema detection algorithms to a scene with a car driving along a road. Because image feature detection based on space–time separable spatio-temporal receptive fields is here applied to a scene where the camera is moving relative to the

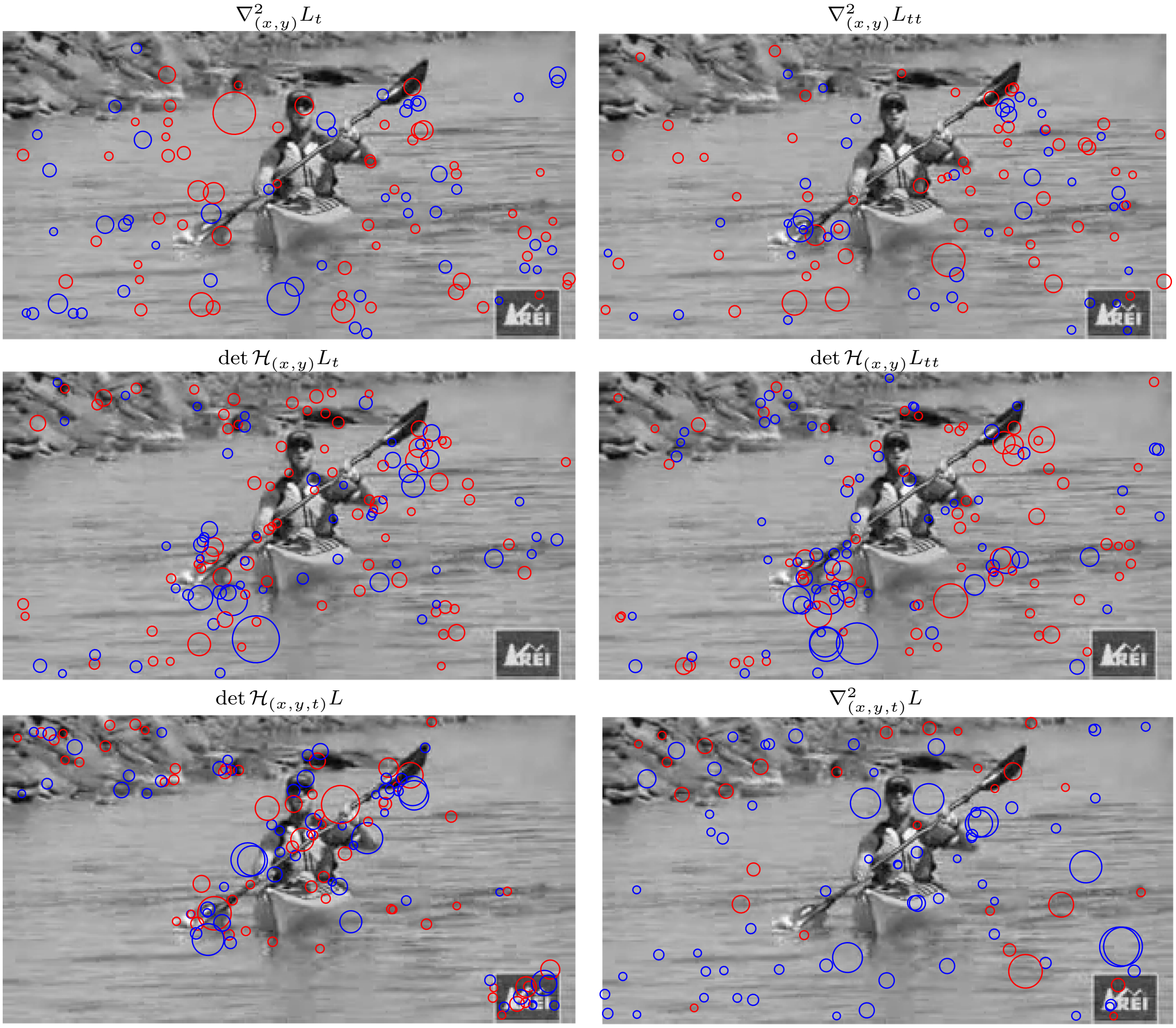
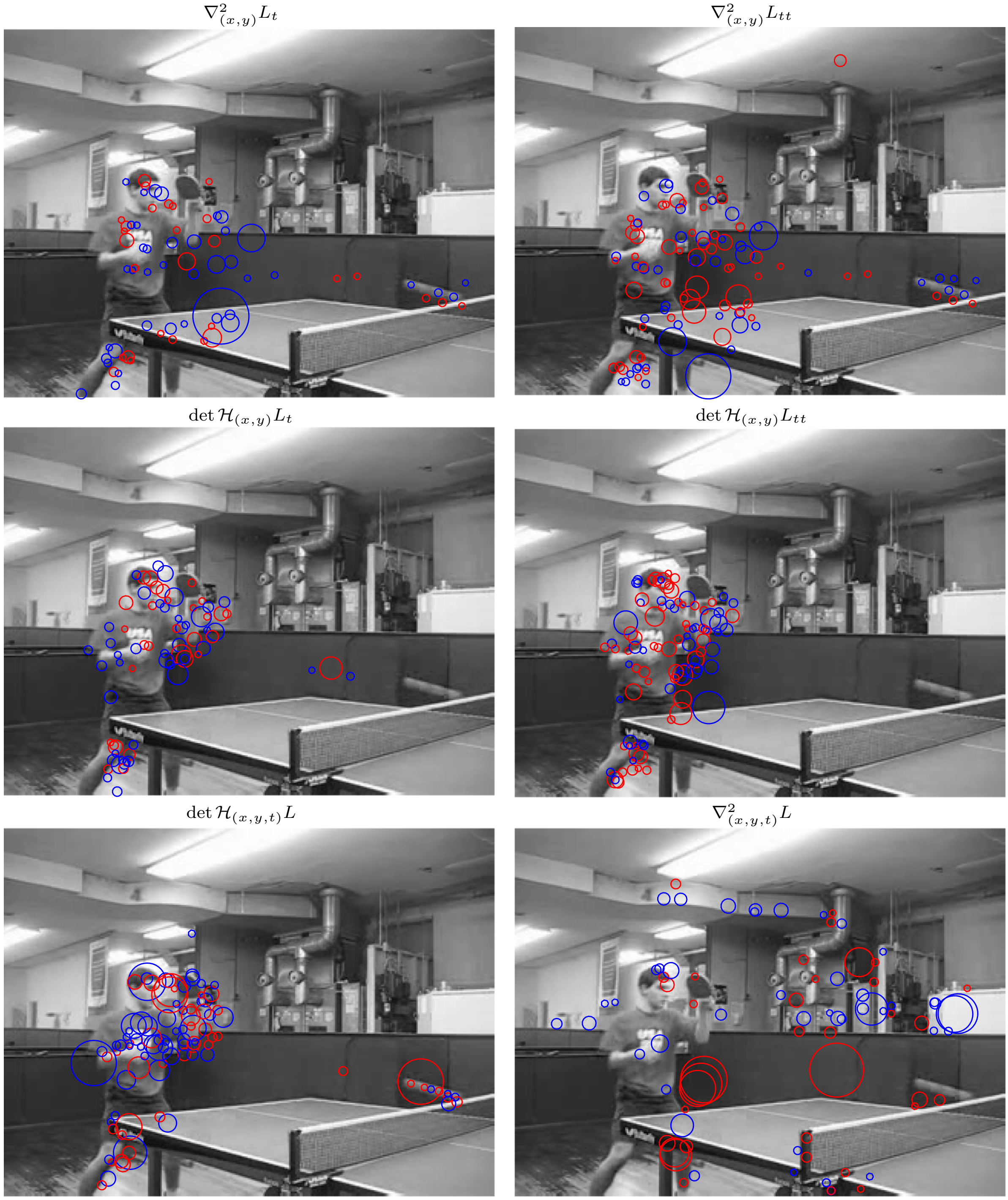

$\nabla^2_{(x,y)} L_t$
$\nabla^2_{(x,y)} L_{tt}$
$\det \mathcal{H}_{(x,y)} L_t$
$\det \mathcal{H}_{(x,y)} L_{tt}$
$\det \mathcal{H}_{(x,y,t)} L$
$\nabla^2_{(x,y,t)} L$

Fig. 7 Spatio-temporal interest points computed from a video sequence in the UCF-101 dataset (Kayaking_g01_c01.avi, cropped) for different scale-normalized spatio-temporal entities and using the presented time-causal and time-recursive spatio-temporal scale-space extrema detection algorithm with the temporal scale-space smoothing performed by a time-discrete approximation of the time-causal limit kernel for $c = 2$ and temporal scale calibration based on $q = 1$: (top left) The spatial Laplacian of the first-order temporal derivative $\nabla^2_{(x,y)} L_t$. (top right) The spatial Laplacian of the second-order temporal derivative $\nabla^2_{(x,y)} L_{tt}$. (middle row left) The determinant of the spatial Hessian of the first-order temporal derivative $\det \mathcal{H}_{(x,y)} L_t$. (middle row right) The determinant of the spatial Hessian of the second-order temporal derivative $\det \mathcal{H}_{(x,y)} L_{tt}$. (bottom row left) The determinant of the spatio-temporal Hessian $\det \mathcal{H}_{(x,y,t)} L$. (bottom row right) The spatio-temporal Laplacian $\nabla^2_{(x,y,t)} L$. Each figure shows a snapshot at frame 90 with a threshold on the magnitude of the scale-normalized differential expression determined such that the average number of features is 50 features per frame. The radius of each circle reflects the spatial scale of the spatio-temporal scale-space extremum (image size: 320 × 172 pixels of original 320 × 240 pixels; frame 90 of 226 frames at 25 frames/s)

environment, static spatial image features in the world that move relative to the motion direction will here lead to spatio-temporal receptive field responses.

For the six basic spatio-temporal interest point detectors that constitute combinations of differential entities used for spatial interest point detection with temporal derivates: (i)–(ii) the spatial Laplacian applied to the first- and second-order temporal derivatives, (iii)–(iv) the determinant of the spatial Hessian applied to the first- and second-order temporal derivatives and (v)–(vi) the first- and second-order temporal derivatives of the determinant of the spatial Hessian matrix, we can note that all these spatio-temporal interest point detectors lead to feature responses for the parked cars that have qualitatively similarities to the responses from applying spatial interest point detectors to a static scene, with the additional constraint that there should also be relative motions



$\nabla^2_{(x,y)} L_t$

$\nabla^2_{(x,y)} L_{tt}$

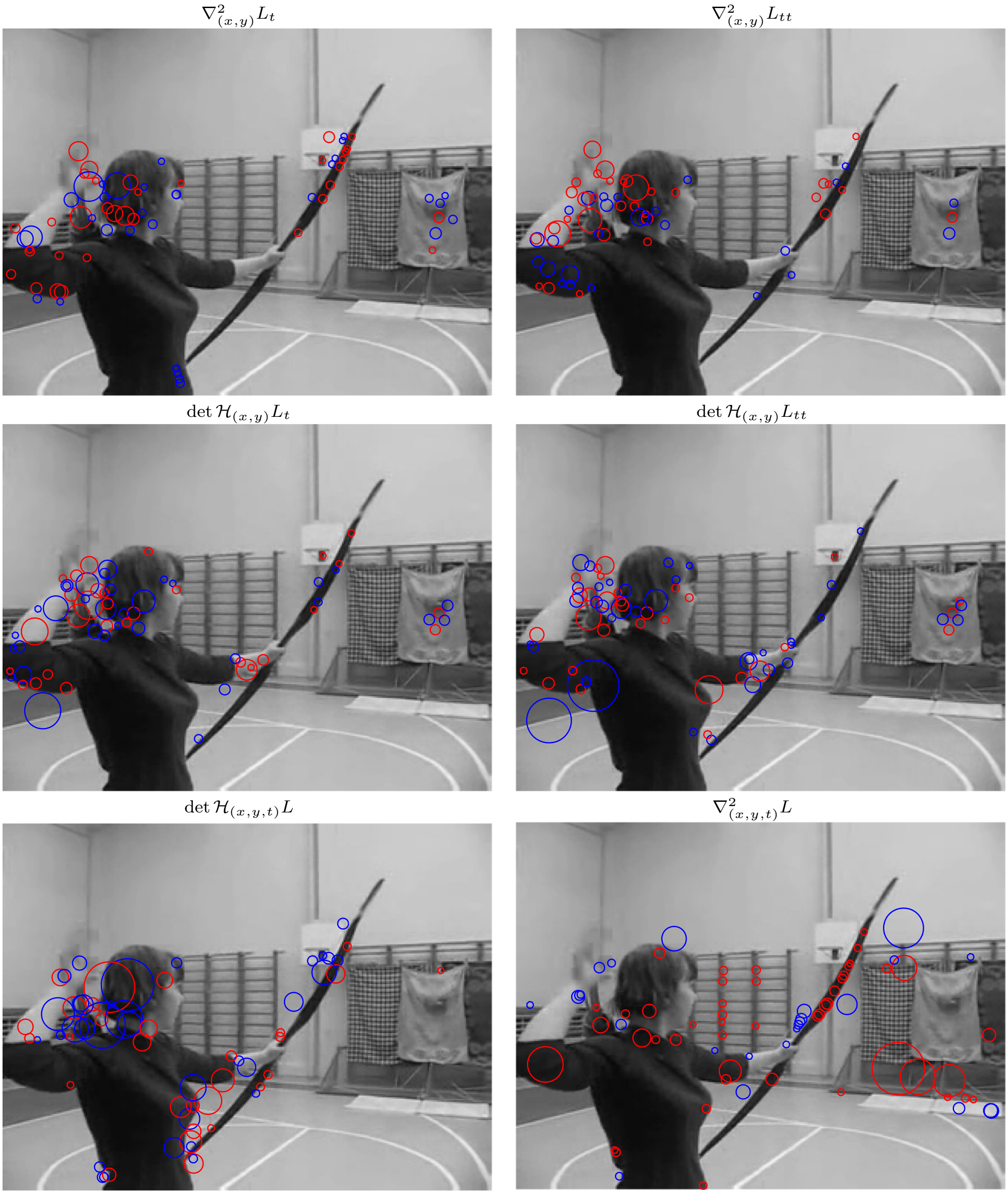
$\det \mathcal{H}_{(x,y)} L_t$

$\det \mathcal{H}_{(x,y)} L_{tt}$

$\det \mathcal{H}_{(x,y,t)} L$

$\nabla^2_{(x,y,t)} L$

Fig. 8 Spatio-temporal interest points computed from a video sequence in the UCF-101 dataset (TableTennisShot_g10_c01.avi) for different scale-normalized spatio-temporal entities and using the presented time-causal and time-recursive spatio-temporal scale-space extrema detection algorithm with the temporal scale-space smoothing performed by a time-discrete approximation of the time-causal limit kernel for $c = 2$ and temporal scale calibration based on $q = 1$: (Top left) The spatial Laplacian of the first-order temporal derivative $\nabla^2_{(x,y)} L_t$. (Top right) The spatial Laplacian of the second-order temporal derivative $\nabla^2_{(x,y)} L_{tt}$. (Middle row left) The determinant of the spatial Hessian of the first-order temporal derivative $\det \mathcal{H}_{(x,y)} L_t$. (Middle row right) The determinant of the spatial Hessian of the second-order temporal derivative $\det \mathcal{H}_{(x,y)} L_{tt}$. (Bottom row left) The determinant of the spatio-temporal Hessian $\det \mathcal{H}_{(x,y,t)} L$. (Bottom row right) The spatio-temporal Laplacian $\nabla^2_{(x,y,t)} L$. Each figure shows a snapshot at frame 37 with a threshold on the magnitude of the scale-normalized differential expression determined such that the average number of features is 30 features per frame. The radius of each circle reflects the spatial scale of the spatio-temporal scale-space extremum (image size: 320 × 240 pixels; frame 37 of 178 frames at 25 frames/s)



$\nabla^2_{(x,y)} L_t$

$\nabla^2_{(x,y)} L_{tt}$

$\det \mathcal{H}_{(x,y)} L_t$

$\det \mathcal{H}_{(x,y)} L_{tt}$

$\det \mathcal{H}_{(x,y,t)} L$

$\nabla^2_{(x,y,t)} L$

Fig. 9 Spatio-temporal interest points computed from a video sequence in the UCF-101 dataset (Archery_g01_c07.avi) for different scale-normalized spatio-temporal entities and using the presented time-causal and time-recursive spatio-temporal scale-space extrema detection algorithm with the temporal scale-space smoothing performed by a time-discrete approximation of the time-causal limit kernel for $c = 2$ and temporal scale calibration based on $q = 1$: (Top left) The spatial Laplacian of the first-order temporal derivative $\nabla^2_{(x,y)} L_t$. (Top right) The spatial Laplacian of the second-order temporal derivative $\nabla^2_{(x,y)} L_{tt}$. (Middle row left) The determinant of the spatial Hessian of the first-order temporal derivative $\det \mathcal{H}_{(x,y)} L_t$. (Middle row right) The determinant of the spatial Hessian of the second-order temporal derivative $\det \mathcal{H}_{(x,y)} L_{tt}$. (Bottom row left) The determinant of the spatio-temporal Hessian $\det \mathcal{H}_{(x,y,t)} L$. (Bottom row right) The spatio-temporal Laplacian $\nabla^2_{(x,y,t)} L$. Each figure shows a snapshot at frame 71 with a threshold on the magnitude of the scale-normalized differential expression determined such that the average number of features is 30 features per frame. The radius of each circle reflects the spatial scale of the spatio-temporal scale-space extremum (image size: 320 × 240 pixels; frame 71 of 143 frames at 25 frames/s)

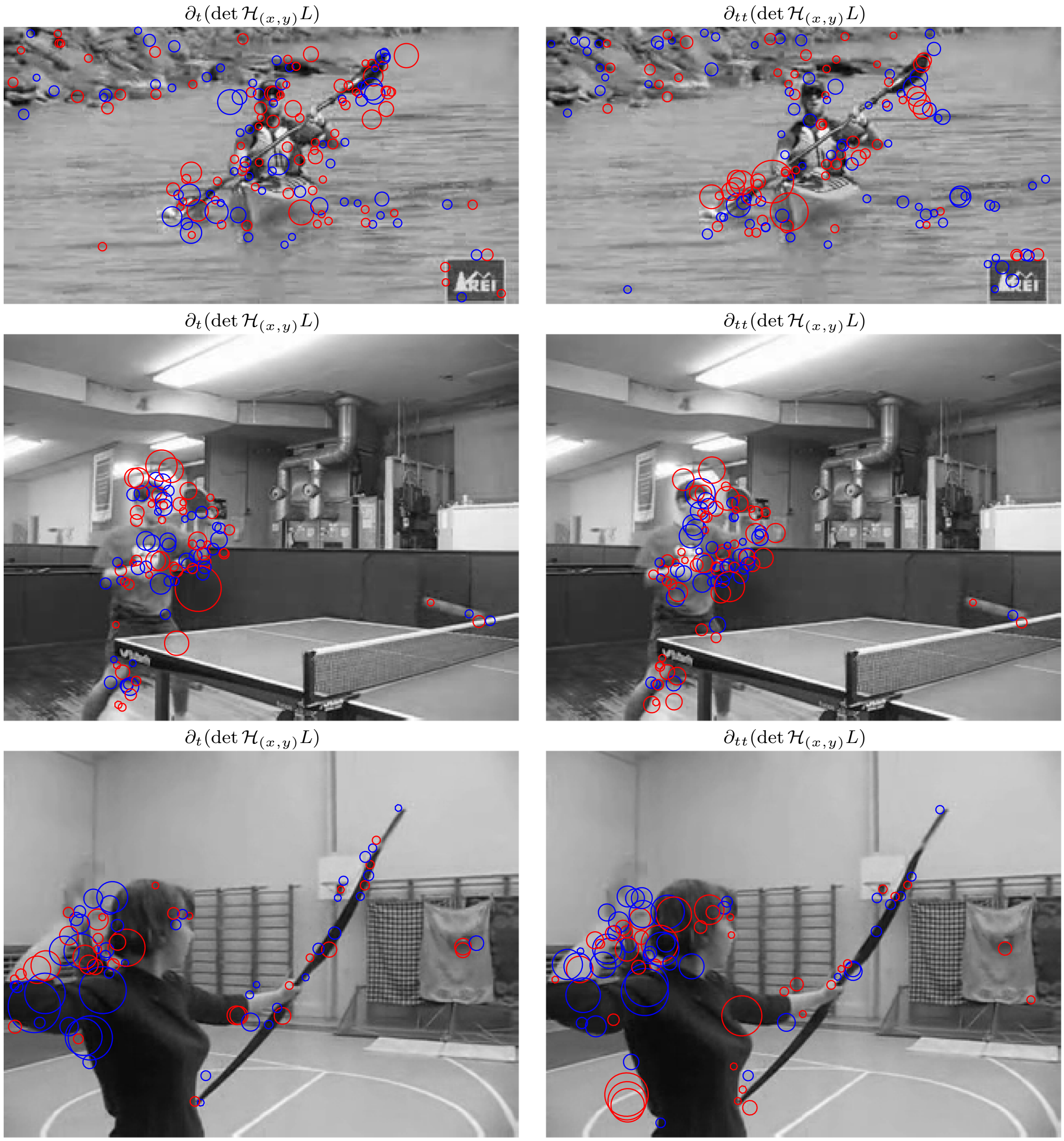
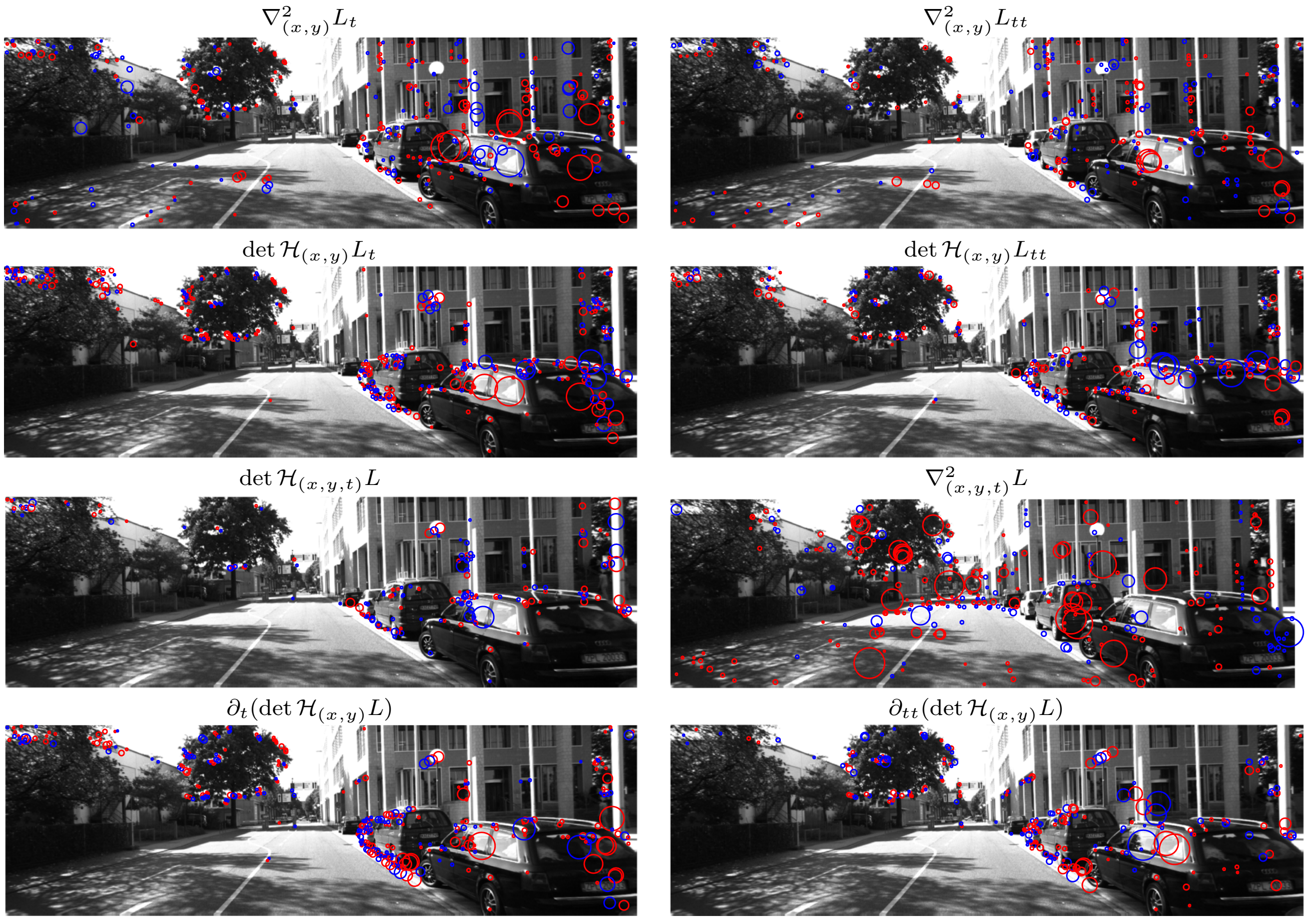

$\partial_t(\det \mathcal{H}_{(x,y)} L)$
$\partial_{tt}(\det \mathcal{H}_{(x,y)} L)$
$\partial_t(\det \mathcal{H}_{(x,y)} L)$
$\partial_{tt}(\det \mathcal{H}_{(x,y)} L)$
$\partial_t(\det \mathcal{H}_{(x,y)} L)$
$\partial_{tt}(\det \mathcal{H}_{(x,y)} L)$

Fig. 10 Spatio-temporal interest points computed from three video sequences in the UCF-101 dataset (Kayaking_g01_c01.avi, cropped, TableTennisShot_g10_c01.avi and Archery_g01_c07.avi) for different scale-normalized spatio-temporal entities and using the presented time-causal and time-recursive spatio-temporal scale-space extrema detection algorithm with the temporal scale-space smoothing performed by a time-discrete approximation of the time-causal limit kernel for $c = 2$ and temporal scale calibration based on $q = 1$: (Left column) The first-order temporal derivative of the determinant of the spatial Hessian $\partial_t(\det \mathcal{H}_{(x,y)} L)$. (Right column) The second-order temporal derivative of the determinant of the spatial Hessian $\partial_{tt}(\det \mathcal{H}_{(x,y)} L)$. Each figure shows a snapshot at a given frame with a threshold on the magnitude of the scale-normalized differential expression determined such that the average number of features is 50 features per frame for the kayak video and 30 features per frame for the table tennis and archery videos. The radius of each circle reflects the spatial scale of the spatio-temporal scale-space extremum (Image size: 320 × 172 pixels of original 320 × 240 pixels. Top row: frame 90 of 226 frames. Middle row: frame 37 of 178 frames. Bottom row: frame 71 of 143 frames. All videos with 25 frames/s)



$\nabla^2_{(x,y)} L_t$

$\nabla^2_{(x,y)} L_{tt}$

$\det \mathcal{H}_{(x,y)} L_t$

$\det \mathcal{H}_{(x,y)} L_{tt}$

$\det \mathcal{H}_{(x,y,t)} L$

$\nabla^2_{(x,y,t)} L$

$\partial_t (\det \mathcal{H}_{(x,y)} L)$

$\partial_{tt} (\det \mathcal{H}_{(x,y)} L)$

Fig. 11 Spatio-temporal interest points computed from a video sequence in the KITTI dataset (tracking test video nr 16) for different scale-normalized spatio-temporal entities and using the presented time-causal and time-recursive spatio-temporal scale-space extrema detection algorithm with the temporal scale-space smoothing performed by a time-discrete approximation of the time-causal limit kernel for $c = 2$ and temporal scale calibration based on $q = 1$: (Top left) The spatial Laplacian of the first-order temporal derivative $\nabla^2_{(x,y)} L_t$. (Top right) The spatial Laplacian of the second-order temporal derivative $\nabla^2_{(x,y)} L_{tt}$. (Middle row left) The determinant of the spatial Hessian of the first-order temporal derivative $\det \mathcal{H}_{(x,y)} L_t$. (Middle row right) The determinant of the spatial Hessian of the second-order temporal derivative $\det \mathcal{H}_{(x,y)} L_{tt}$. (bottom row left) The determinant of the spatio-temporal Hessian $\det \mathcal{H}_{(x,y,t)} L$. (Bottom row right) The spatio-temporal Laplacian $\nabla^2_{(x,y,t)} L$. Each figure shows a snapshot at frame 77 with a threshold on the magnitude of the scale-normalized differential expression determined such that the average number of features is 200 features per frame. The radius of each circle reflects the spatial scale of the spatio-temporal scale-space extremum (image size: 1242 × 375 pixels; frame 77 of 509 frames at 10 frames/s). (A temporal sampling rate of 10 frames/s is, however, too sparse for this type of local differential analysis of such fast changes in the image structures over time)

between the camera and the environment. For (vii) the genuine 3-D determinant of the spatio-temporal Hessian, the responses are on the other hand more selective, while for (viii) the spatio-temporal Laplacian, the responses are far less selective and less informative.

An alternative way of handling spatio-temporal scenes with dominant relative motions between the camera and the environment, in contrast to this use of space–time separable receptive fields for only image velocity $v = 0$, is by exploiting the full structure of the spatio-temporal receptive field model (1), by considering spatio-temporal receptive fields with nonzero image velocities $v \neq 0$, which can be locally adapted to the local motion direction corresponding to velocity adaptation [50,51,61] or alternatively performing local, regional or global image stabilization. Then, the image operations can be made truly covariant under local, regional or global Galilean image transformations [67,71] and allow for a more explicit separation of spatio-temporal receptive field responses that correspond to more complex spatio-temporal image structures than local Galilean motions.

5.4 Covariance and Invariance Properties

From the theoretical scale selection properties of the spatial scale-normalized derivative operators according to the spatial scale selection theory in Lindeberg [65] in combination with the temporal scale selection properties of the temporal scale selection theory in Lindeberg [77] with the scale covariance of the underlying spatio-temporal derivative expressions $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$, $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$, $\det \mathcal{H}_{(x,y,t),\text{norm}} L$, $\partial_{t,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ and $\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ described in Lindeberg [75], it follows that these spatio-temporal interest point detectors are truly scale covariant under independent scaling transformations of the spatial and the temporal domains if the temporal smoothing is performed by either a non-causal Gaussian kernel $g(t;\; \tau)$ over the temporal domain or the time-causal limit kernel $\Psi(t;\; \tau, c)$. From the general proof in Sect. 3, it follows that the selected spatio-temporal scale levels transform in a scale-covariant way under independent scaling transformations of the spatial and the temporal domains. Additionally, the post-normalized magnitude estimates from these seven spatio-temporal differential invariants are truly scale invariant.

6 Quantifying the Accuracy of the Scale Estimates and the Amounts of Temporal Delays

The theoretical analysis of the scale selection properties of the different types of spatio-temporal interest point detectors presented in Sect. 4 was performed for a non-causal Gaussian spatio-temporal concept and using model signals based on Gaussian or integrated Gaussian intensity profiles over time. While it was conceptually shown in Lindeberg [77] that important scale selection properties in terms of temporal scale-invariance transfer from a non-causal Gaussian temporal scale-space concept to the time-causal temporal scale-space concept based on the time-causal limit kernel, it is of interest to also quantify the numerical properties in terms of the spatio-temporal scale estimates and the temporal delays obtained from a truly time-causal scale-space concept and a time-causal implementation.

In this section, we will experimentally quantify: (i) how well the spatio-temporal scale selection properties transfer to a discrete implementation, specifically how accurate the spatial and temporal scale estimates are for idealized model patterns with ground truth, as well as (ii) how the different spatio-temporal feature detectors differ in their ability to respond fast with regard to time-critical applications.

6.1 Time-Causal Gaussian Blink

To quantify the transfer of the spatio-temporal scale selection properties to a time-causal spatio-temporal domain, we first generated a set of videos with time-causal Gaussian blinks obtained by filtering a discrete delta function with a discrete Gaussian kernel over the spatial domain and a discrete approximation of the time-causal limit kernel over the temporal domain. Such videos sequences were generated with spatial extent $\sigma_{s,0} = 8$ pixels and temporal durations of $\sigma_{\tau,0} = 40$, 80, 160, 320 and 640 ms at a frame rate of 50 frames/s and for distribution parameter $c = 2$ of the time-causal limit kernel. The reason for not varying the spatial scale parameter in this experiment is that the properties of the spatial scale selection mechanism have already been sufficiently well established and tested.

Then, we detected spatio-temporal scale-space extrema of: (i) the spatial Laplacian of the second-order temporal derivative of $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$, (ii) the determinant of the spatial Hessian of the second-order temporal derivative $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$, (iii) the determinant of the spatio-temporal Hessian $\det \mathcal{H}_{(x,y,t),\text{norm}} L$ and (iv) the second-order temporal derivative of the determinant of the spatial Hessian $\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ for each one of these videos, and recorded (i) the selected spatial scale $\hat{\sigma}_s$ in units of pixels, (ii) the selected temporal scale $\hat{\sigma}_\tau$ in units of milliseconds and (iii) a measure of the effective temporal delay $\delta = \hat{t} - t_{\max}$ defined as the time difference between the time moment $\hat{t}$ at which the spatio-temporal scale-space extremum is detected and the time moment $t_{\max}$ at which the spatio-temporal maximum in the input function occurred. The motivation for the latter choice is that because of the time-causal model, each spatio-temporal pattern is associated with an inherent temporal delay. By compensating for this delay, the intention is that the compensated delay score should more reflect the additional amount of temporal delay caused by the time-causal feature detection method.

The results of these experiments are given in Table 3 for two different settings of the temporal scale calibration parameter q. Note that (i) the spatial scale estimates $\hat{\sigma}_s$ are highly accurate and that (ii) when using $q = 1$ the temporal scale estimates $\hat{\sigma}_\tau$ do also give good estimates of the temporal duration of the underlying spatio-temporal image structures considering the coarse sampling of the temporal scale levels induced by a distribution parameter of $c = 2$, which means that the ratio between adjacent temporal scale levels is equal to two in units of dimension [time] and which in turn limits the effective resolution of the temporal scale estimates. Additionally, the implementation differs from the presented scale selection theory in the respects that: (i) the theoretical analysis has been performed based on the non-causal Gaussian temporal scale-space model, whereas the experiments are performed using the time-causal scale-space model, (ii) the spatio-temporal scale selection theory is continuous, whereas the discrete implementation is based on the discrete analogue of the Gaussian kernel [56] over space and recursive filters over time and (iii) for shorter temporal scales, the temporal scales of the model signals are close to the inner temporal scale in the video, determined by the frame rate of 50 fps corresponding to 20 ms between adjacent frames, implying

Table 3 Numerical quantification of the spatio-temporal scale selection properties of four spatio-temporal interest point detectors when applied to model signals defined as time-causal Gaussian blinks of spatial extent $\sigma_{s,0} = 8$ pixels and different temporal durations $\sigma_{\tau,0} = 40$, 80, 160, 320 and 640 ms

$\sigma_{s,0}$	$\sigma_{\tau,0}$	$\nabla^2_{(x,y)} L_{tt}$			$\det \mathcal{H}_{(x,y)} L_{tt}$			$\det \mathcal{H}_{(x,y,t)} L$		
		$\hat{\sigma}_s$	$\hat{\sigma}_\tau$	δ	$\hat{\sigma}_s$	$\hat{\sigma}_\tau$	δ	$\hat{\sigma}_s$	$\hat{\sigma}_\tau$	δ
Scale selection for a time-causal Gaussian blink using $q = 1$										
8	40	7.99	37	6	7.99	37	6	7.99	42	60
8	80	7.99	71	−5	7.99	73	−5	7.99	79	107
8	160	7.99	179	−18	7.99	173	−18	7.99	157	210
8	320	7.99	334	−36	7.99	330	−36	7.99	313	426
8	640	7.99	676	−64	7.99	663	−64	7.99	626	869
Scale selection for a time-causal Gaussian blink using $q = 3/4$										
8	40	7.99	36	3	7.99	34	6	7.99	33	42
8	80	7.99	36	−27	7.99	48	58	7.99	48	60
8	160	7.99	117	−57	7.99	114	−56	7.99	105	109
8	320	7.99	223	−123	7.99	220	−123	7.99	204	213
8	640	7.99	439	−246	7.99	436	−246	7.99	418	433

$\sigma_{s,0}$	$\sigma_{\tau,0}$	$\partial_{tt} (\det \mathcal{H}_{(x,y)} L)$			$\partial_{tt} (\det \mathcal{H}_{(x,y)} L)$		
		$\hat{\sigma}_s$	$\hat{\sigma}_\tau$	δ	$\hat{\sigma}_s$	$\hat{\sigma}_\tau$	δ
Scale selection for a time-causal Gaussian blink using $q = 1$ or $q = 3/4$							
8	40	7.99	37	67	7.99	29	48
8	80	7.99	73	116	7.99	51	69
8	160	7.99	152	222	7.99	95	119
8	320	7.99	298	445	7.99	194	229
8	640	7.99	596	901	7.99	392	460

For each one of the differential entities (i) the spatial Laplacian of the second-order temporal derivative $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$, (ii) the determinant of the spatial Hessian of the second-order temporal derivative $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$, (iii) the determinant of the spatio-temporal Hessian $\det \mathcal{H}_{(x,y,t),\text{norm}} L$ and (iv) the second-order temporal derivative of the determinant of the spatial Hessian $\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$, the results show (i) the selected spatial scale $\hat{\sigma}_s$ in units of pixels, (ii) the selected temporal scale $\hat{\sigma}_\tau$ in units of milliseconds and (iii) the effective temporal delay $\delta = \hat{t} - t_{\max}$ defined as the time difference between the time moment $\hat{t}$ at which the spatio-temporal scale-space extremum is detected and the time moment $t_{\max}$ at which the spatio-temporal maximum of the input function occurs. (The experiments have been performed at a frame rate of 50 fps corresponding to 20 ms between adjacent frames and for distribution parameter $c = 2$ of the time-causal limit kernel corresponding to a sampling of the temporal scale parameter by a factor of two between adjacent temporal scale levels in units of dimension [time])

that the temporal discretization effects at shorter temporal scales become stronger.

For this family of model signals, the spatial Laplacian of the second-order temporal derivative $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$ and the determinant of the Hessian of the second-order temporal derivative $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$ respond very fast to the onset of a spatio-temporal Gaussian blob when using $q = 1$. For the determinant of the spatio-temporal Hessian $\det \mathcal{H}_{(x,y,t),\text{norm}} L$ and the second-order temporal derivative of the determinant of the spatial Hessian $\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$, the temporal delays are, however, substantial when using $q = 1$. By instead choosing the temporal scale calibration parameter q to a lower value of $q = 3/4$, the effective temporal delays can be substantially reduced in many cases up to a reduction near 50% for the determinant of the spatio-temporal Hessian $\det \mathcal{H}_{(x,y,t),\text{norm}} L$ and the second-order temporal derivative of the determinant of the spatial Hessian $\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ at the cost of less accurate but still not completely unreasonable estimates of the temporal duration of the underlying spatio-temporal image structures.

A general conclusion that we can draw from this experiments is that the operators $\nabla^2_{(x,y),\text{norm}} L_{tt,\text{norm}}$ and $\det \mathcal{H}_{(x,y),\text{norm}} L_{tt,\text{norm}}$ that operate directly on temporal derivatives respond significantly faster compared to the operator $\det \mathcal{H}_{(x,y,t),\text{norm}} L$ that operates on the joint space–time structure and the operator $\partial_{tt,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ that operates on temporal derivatives of a nonlinear spatial differential invariant.

6.2 Time-Causal Gaussian Onset Blob

To quantify the transfer of the spatio-temporal scale selection properties for another class of model signals, we then

Table 4 Numerical quantification of the spatio-temporal scale selection properties of three spatio-temporal interest point detectors when applied to model signals defined as time-causal Gaussian onset blobs of spatial extent $\sigma_{s,0} = 8$ pixels and different temporal durations $\sigma_{\tau,0} = 40, 80, 160, 320$ and 640 ms

$\sigma_{s,0}$	$\sigma_{\tau,0}$	$\nabla^2_{(x,y)} L_t$			$\det \mathcal{H}_{(x,y)} L_t$			$\partial_t (\det \mathcal{H}_{(x,y)} L)$		
		$\hat{\sigma}_s$	$\hat{\sigma}_\tau$	δ	$\hat{\sigma}_s$	$\hat{\sigma}_\tau$	δ	$\hat{\sigma}_s$	$\hat{\sigma}_\tau$	δ
Scale selection for a time-causal Gaussian onset blob using $q = 1$										
8	40	7.99	43	37	7.99	43	57	7.99	36	87
8	80	7.99	74	116	7.99	75	116	7.99	72	179
8	160	7.99	150	240	7.99	152	240	7.99	151	370
8	320	7.99	311	498	7.99	313	498	7.99	302	762
8	640	7.99	616	1023	7.99	620	1023	7.99	605	1557
Scale selection for a time-causal Gaussian onset blob using $q = 3/4$										
8	40	7.99	32	34	7.99	30	34	7.99	35	58
8	80	7.99	56	65	7.99	56	65	7.99	50	113
8	160	7.99	106	130	7.99	106	130	7.99	103	228
8	320	7.99	207	267	7.99	208	267	7.99	201	469
8	640	7.99	421	552	7.99	422	552	7.99	406	961

For each one of the differential entities (i) the spatial Laplacian of the first-order temporal derivative $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$, (ii) the determinant of the spatial Hessian of the first-order temporal derivative $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$ and (iii) the first-order temporal derivative of the determinant of the spatial Hessian $\partial_{t,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$, the results show (i) the selected spatial scale $\hat{\sigma}_s$ in units of pixels, (ii) the selected temporal scale $\hat{\sigma}_\tau$ in units of milliseconds and (iii) the effective temporal delay $\delta = \hat{t} - t_{\max}$ defined as the time difference between the time moment $\hat{t}$ at which the spatio-temporal scale-space extremum is detected and the time moment $t_{\max}$ at which the spatio-temporal maximum of the spatio-temporal smoothing kernel at the same spatial and temporal scales occurs. (The experiments have been performed at a frame rate of 50 fps corresponding to 20 ms between adjacent frames and for distribution parameter $c = 2$ of the time-causal limit kernel corresponding to a sampling of the temporal scale parameter by a factor of two between adjacent temporal scale levels in units of dimension [time])

generated a set of videos with time-causal Gaussian onset blobs obtained by filtering a the tensor product between a discrete delta function over the spatial domain and discrete Heaviside function over the temporal domain function with a discrete Gaussian kernel over the spatial domain and a discrete approximation of the time-causal limit kernel over the temporal domain. Such videos sequences were generated with spatial extent $\sigma_{s,0} = 8$ pixels and temporal durations of $\sigma_{\tau,0} = 40, 80, 160, 320$ and 640 ms at a frame rate of 50 frames/s and for distribution parameter $c = 2$ of the time-causal limit kernel.

Then, we detected spatio-temporal scale-space extrema of: (i) the spatial Laplacian of the first-order temporal derivative $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$, (ii) the determinant of the spatial Hessian of the first-order temporal derivative $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$ and (iii) the first-order temporal derivative of the determinant of the spatial Hessian $\partial_{t,\text{norm}} (\det \mathcal{H}_{(x,y),\text{norm}} L)$ for each one of these videos, and recorded the (i) the selected spatial scale $\hat{\sigma}_s$ in units of pixels, (ii) the selected temporal scale $\hat{\sigma}_\tau$ in units of milliseconds and (iii) a measure of the effective temporal delay $\delta = \hat{t} - t_{\max}$ defined as the time difference between the time moment $\hat{t}$ at which the spatio-temporal scale-space extremum is detected and the time moment $t_{\max}$ at which the spatio-temporal maximum of the spatio-temporal scale-space kernel at the same spatio-temporal scale $\sigma_{\tau,0}$ occurs.

The results of these experiments are given in Table 4 for two different settings of the temporal scale calibration parameter q. Note that (i) again the spatial scale estimates $\hat{\sigma}_s$ are highly accurate and that (ii) when using $q = 1$ the temporal scale estimates $\hat{\sigma}_\tau$ do also give good estimates of the temporal duration of the underlying spatio-temporal signals again considering the coarse sampling of the temporal scale levels induced by sparse sampling the temporal scale levels resulting from the distribution parameter of $c = 2$ for the time-causal limit kernel, which in turn means that the ratio between adjacent temporal scale levels is equal to two in units of dimension [time] and which again limits the effective resolution of the temporal scale estimates. For this problem of onset detection, the temporal delays are, however, longer than for the previous problem of detecting blinks. By instead choosing the temporal scale calibration parameter q to a lower value of $q = 3/4$, the effective temporal delay can be substantially reduced in some cases up to a reduction near 50 % for the spatial Laplacian of the first-order temporal derivative of $\nabla^2_{(x,y),\text{norm}} L_{t,\text{norm}}$ and the determinant of the spatial Hessian of the first-order temporal derivative $\det \mathcal{H}_{(x,y),\text{norm}} L_{t,\text{norm}}$ at the cost of less accurate but still not completely unreasonable estimates of the temporal duration of the underlying spatio-temporal image structures

7 Summary and Discussion

We have presented a general theory and methodology for performing simultaneous detection of local characteristic spatial and temporal scale estimates in video data. The theory comprises both (i) feature detection performed within a non-causal spatio-temporal scale-space representation computed for offline analysis of pre-recorded video data and (ii) feature detection performed from real-time image streams where the future cannot be accessed and memory requirements call for time-recursive algorithms based on only compact buffers of what has occurred in the past.

As a theoretical foundation for spatio-temporal scale selection, we have stated general sufficiency results regarding scale-covariant spatio-temporal scale estimates and complementary invariance properties of spatio-temporal features defined from video data in which there may be independent scaling transformations of the spatial and the temporal domains. For a wide class of homogeneous spatio-temporal differential expressions, the spatio-temporal scale estimates obtained from the presented theory and methodology have been shown to obey the basic property that they adaptively follow independent local spatial and temporal scaling transformations in the video data, which constitutes a basic requirement on a spatio-temporal scale selection mechanism. In other words, if the spatial size of the image structures changes by a factor S_s in the spatial domain and/or the temporal duration of the spatio-temporal image structures changes by a factor S_τ, then the spatial scale parameter in units of $\sigma_s = \sqrt{s}$ and the temporal scale parameter in units of $\sigma_\tau = \sqrt{\tau}$ of the detected spatio-temporal image features will change by corresponding factors. Additionally, we have shown that the magnitude estimates either are automatically invariant under spatio-temporal scaling transformations or can be compensated to become so by post-normalization, depending on the specific values of the scale normalization parameters γ_s and γ_τ. These properties together imply that the presented theory and methodology obeys the necessary properties to handle video data in which there may be large spatial and temporal scaling variations in the spatio-temporal image structures.

For seven specific spatio-temporal differential invariants: (i)–(ii) the spatial Laplacian of the first- and second-order temporal derivatives, (iii)–(iv) the determinant of the spatial Hessian of the first- and second-order temporal derivatives, (v) the determinant of the spatio-temporal Hessian matrix and (vi)–(vii) the first- and second-order temporal derivatives of the determinant of the spatial Hessian, we have performed an in-depth analysis of their theoretical scale selection properties and shown how scale calibration can be performed to determine the spatial and temporal scale normalization powers γ_s and γ_τ such that the selected spatio-temporal scale levels reflect the spatial extent and the temporal duration of the underlying spatio-temporal features that gave rise to the feature responses. These spatio-temporal differential invariants can all be used for formulating spatio-temporal interest point detectors. Theoretically and experimentally, we have described and illustrated their properties and shown that they lead to intuitively reasonable results.

For one spatio-temporal differential expression, an attempt to define a spatio-temporal Laplacian, we have on the other hand shown that this differential expression is not scale covariant under independent rescalings of the spatial and temporal domains, which explains a previously noted poor robustness of the scale selection step in the spatio-temporal interest point detector based on the spatio-temporal Harris operator [49].

Whereas the presented spatio-temporal scale selection theory is fully continuous over space and time, we have by quantitative experiments on model signals with ground truth shown that the numerical accuracy of the spatio-temporal scale estimates carries over to a carefully designed discrete implementation, based on the discrete analogue of the Gaussian over space and a cascade of first-order recursive filters over time.

To allow for different trade-offs between the temporal response properties of time-causal spatio-temporal feature detection (shorter temporal delays) in relation to signal detection theory, which would call for detection of image structures at the same spatial and temporal scales as they occur, we have specifically introduced a parameter q to regulate the temporal scale calibration to finer temporal scales $\hat{\tau} = q^2 \tau_0$ as opposed to the more common choice $\hat{s} = s_0$ over the spatial domain. According to the presented theoretical analysis of scale selection properties in non-causal spatio-temporal scale space, the results predict that this parameter should reduce the temporal delay by a factor of q: $\Delta t \mapsto q\,\Delta t$. Our numerical experiments with scale selection properties in time-causal spatio-temporal scale space confirm that a substantial decrease in temporal delay is obtained. The specific choice of the parameter q should be optimized with respect to the task that the spatio-temporal selection and the spatio-temporal features are to be used for and given specific requirements of the application domain.

We have also presented an explicit algorithm for detecting spatio-temporal interest points in a time-causal and time-recursive context in which the future cannot be accessed and memory requirements call for only compact buffers to store partial records of what has occurred in the past and presented experimental results of applying this algorithm to real-world video data for the different types of spatio-temporal interest point detectors that we have studied theoretically.

Experimentally, we have shown that four of the presented spatio-temporal interest operators: (i)–(ii) the spatial Laplacian of the first- and second-order temporal derivatives and (iii)–(iv) the determinant of the Hessian of the first- and second-order temporal derivatives, lead to significantly shorter temporal delays than (v) the determinant of the spatio-temporal Hessian matrix or (vi)–(vii) the first- and second-order temporal derivatives of the determinant of the spatial Hessian.

While the experimental results in this paper have been presented solely based on a time-causal and time-recursive spatio-temporal concept, the overall methodology can also be implemented based on a non-causal Gaussian spatio-temporal scale-space concept [67]. Such an implementation would, however, require more computations and larger temporal buffers compared to using the time-causal and time-recursive receptive fields based on first-order integrators coupled in cascade that constitute the temporal smoothing model underlying the implementation reported in this work. Additionally, an ad hoc use of time-delayed truncated Gaussian kernels instead would be expected to lead to less rapid temporal responses for time-critical applications compared to the truly time-causal scale-space kernels used for the experiments in this work. For offline analysis of pre-recorded data on an architecture where computational and memory resources do not constitute a bottle-neck, such a non-causal implementation would on the other hand have the potential of computing more accurate image features, since the method could then also make use of information from the future in relation to any pre-recorded time moment, which is not permitted for these time-causal operations.

We propose that the spatio-temporal scale selection mechanism presented in this paper should be far more general than the more specific applications developed here for detecting spatio-temporal interest points. Concerning extensions of the approach, a first natural extension concerns extending the sparse spatio-temporal scale selection into dense spatio-temporal scale selection, which is addressed in a companion paper [76]. A second natural extension is to extend the current use of a space–time separable spatio-temporal scale-space representation based on spatio-temporal receptive fields (1) with image velocity zero to incorporate mechanisms for velocity-adapted spatio-temporal receptive fields with nonzero image velocities and/or image stabilization.

A Spatial Scale-Space Extrema Detection Algorithm

This appendix describes an algorithm for detecting spatial scale-space extrema of a spatial differential expression $\mathcal{D}L$ as used in Lindeberg [59,65,74].

For a discrete signal, a point is defined as a discrete scale-space maximum (minimum) if its value is greater than (less than) the values of all its neighbours in scale space. For the three-dimensional scale-space representation of a two-dimensional image, comparisons will thus be made with respect to all its 26 neighbours or equivalently against all the non-central points in a $3 \times 3 \times \times 3$-neighbourhood.

This definition directly implies an operational method for scale-space extrema detection by:

(i) Gaussian smoothing $L(\cdot, \cdot;\ s) = g(\cdot, \cdot;\ s) * f$ to be approximated by some discrete approximation kernel $L(\cdot, \cdot;\ s) = T(\cdot, \cdot;\ s) * f$,
(ii) computation of discrete derivative approximations $L_{x^\alpha y^\beta} \approx \delta_{x^\alpha y^\beta} L$,
(iii) scale normalization of the derivative responses $L_{\xi^\alpha \eta^\beta} = C_{x^\alpha y^\beta}\, L_{x^\alpha y^\beta}$ using some discretization method for scale-normalized derivatives,
(iv) combination of the derivative approximations into a differential invariant $\mathcal{D}L$ for each point and scale,
(v) scale-space extrema detection by local comparisons in a neighbourhood around each point and
(vi) thresholding and/or sorting of all the feature responses obtained from the image.

For discrete scale-space smoothing, we use a scale-space concept especially developed for discrete image data corresponding to discrete convolution with a discrete analogue of the Gaussian kernel [56]. Discrete derivative approximations are then computed by applying difference operators to the image data, for which the discrete scale-space properties transfer also to the discrete implementation [57].

For discrete approximation of scale-normalized derivatives, one can perform either (i) variance normalization, by multiplying the discrete derivative approximations by the scale parameter s raised to an appropriate power γ or by (ii) l_p-normalization, by normalizing the equivalent discrete derivative approximation kernels to having the same l_p-norm as the L_p-norm of the corresponding scale-normalized Gaussian derivative kernels [65,80].

Preferably, the scale levels should be distributed such that the distribution is uniform when measured in terms of effective scale $s_{eff} = \log s$. For discrete signals, a natural definition of the notion of effective scale is $s_{eff} = A + B \log p(s)$, where $p(s)$ denotes the expected density of local extrema at scale s [58].

Conceptually, one may think of this algorithm as first generating a three-dimensional volume of data for a two-dimensional image. Computationally, however, it is usually faster to implement the algorithm as a moving window that keeps three adjacent scales in a buffer for local comparisons and then moves the buffer from finer to coarser scales. In this way, memory accesses can be confined to a smaller part of memory in a serial implementation.

The algorithm can also be made more efficient by introducing a threshold on the magnitude of the response $|\mathcal{D}_{\text{norm}} L| \geq C_{\mathcal{D}L}$ and only performing comparisons at points that satisfy this condition. Of course, we do not need to perform comparisons with all the 26 neighbours of each point. In a serial implementation, the comparisons can be stopped as soon as we know that the point cannot be a scale-space maximum or a scale-space minimum.

At the boundary cases with $s = s_{\min}$ or $s = s_{\max}$, boundary extrema can be accepted and be labelled as such, to allow for inclusions of image features whose true characteristic scales fall outside the given scale range while still giving rise to feature responses within this range.

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. Abramowitz, M., Stegun, I.A. (eds.): Handbook of Mathematical Functions, 55th edn. National Bureau of Standards, Applied Mathematics Series (1964)
3. Adelson, E., Bergen, J.: Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A **2**, 284–299 (1985)
4. Alcantarilla, P.F., Bartoli, A., Davison, A.J.: KAZE features. In: Proceedings of European Conference on Computer Vision (ECCV 2012). Springer LNCS, vol. 7577, pp. 214–227 (2012)
5. Bay, H., Ess, A., Tuytelaars, T., van Gool, L.: Speeded up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
6. Bilinski, P., Bremond, F.: Evaluation of local descriptors for action recognition in videos. In: International Conference on Computer Vision Systems, pp. 61–70 (2011)
7. Brox, T., Weickert, J.: A TV flow based local scale measure for texture discrimination. In: Proceedings of European Conference on Computer Vision (ECCV 2004), pp. 578–590 (2004)
8. Brox, T., Weickert, J.: A TV flow based local scale estimate and its application to texture discrimination. J. Vis. Commun. Image Represent. **17**(5), 1053–1073 (2006)
9. Chakraborty, B., Holte, M.B., Moeslund, T.B., Gonzàlez, J.: Selective spatio-temporal interest points. Comput. Vis. Image Underst. **116**(3), 396–410 (2012)
10. Comaniciu, D., Ramesh, V., Meer, P.: The variable bandwidth mean shift and data-driven scale selection. In: Proceedings of International Conference on Computer Vision (ICCV 2001), pp. 438–445. Vancouver, Canada (2001)
11. Dawn, D.D., Shaikh, S.H.: A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. Vis. Comput. **32**(3), 289–306 (2016)
12. DeAngelis, G.C., Anzai, A.: A modern view of the classical receptive field: linear and non-linear spatio-temporal processing by V1 neurons. In: Chalupa, L.M., Werner, J.S. (eds.) The Visual Neurosciences, vol. 1, pp. 704–719. MIT Press (2004)
13. DeAngelis, G.C., Ohzawa, I., Freeman, R.D.: Receptive field dynamics in the central visual pathways. Trends Neurosci. **18**(10), 451–457 (1995)
14. de Geest, R., Tuytelaars, T.: Dense interest features for video processing. In: Proceedings of International Conference on Image Processing (ICIP 2014), pp. 5771–5775 (2014)
15. 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)
16. Derpanis, K.G., Wildes, R.P.: Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. IEEE Trans. Pattern Anal. Mach. Intell. **34**(6), 1193–1205 (2012)
17. Dickscheid, T., Schindler, F., Förstner, W.: Coding images with local features. Int. J. Comput. Vis. **94**(2), 154–174 (2011)
18. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Proceedings of 2nd Joint Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72. Beijing, China (2005)
19. Elder, J., Zucker, S.: Local scale control for edge detection and blur estimation. IEEE Trans. Pattern Anal. Mach. Intell. **20**(7), 699–716 (1998)
20. Everts, I., van Gemert, J.C., Gevers, T.: Evaluation of color STIPs for human action recognition. In: Proceedings of Computer Vision and Pattern Recognition (CVPR 2013), pp. 2850–2857 (2013)
21. Everts, I., van Gemert, J.C., Gevers, T.: Evaluation of color spatio-temporal interest points for human action recognition. IEEE Trans. Image Process. **23**(4), 1569–1580 (2014)
22. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. arXiv preprint arXiv:1604.06573 (2016)
23. Fleet, D.J., Langley, K.: Recursive filters for optical flow. IEEE Trans. Pattern Anal. Mach. Intell. **17**(1), 61–67 (1995)
24. Florack, L.M.J.: Image Structure. Series in Mathematical Imaging and Vision. Springer, Berlin (1997)
25. Förstner, W., Dickscheid, T., Schindler, F.: Detecting interpretable and accurate scale-invariant keypoints. In: Proceedings of International Conference on Computer Vision (ICCV 2009), pp. 2256–2263 (2009)
26. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. **32**(11), 1231–1237 (2013)
27. Guichard, F.: A morphological, affine, and Galilean invariant scale-space for movies. IEEE Trans. Image Process. **7**(3), 444–456 (1998)
28. Hassner, T., Mayzels, V., Zelnik-Manor, L.: On SIFTs and their scales. In: Proceedings of Computer Vision and Pattern Recognition (CVPR 2012), pp. 1522–1528. Providence, Rhode Island (2012)
29. Hassner, T., Filosof, S., Mayzels, V., Zelnik-Manor, L.: SIFTing through scales. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1431–1443 (2016)
30. Holte, M.B., Chakraborty, B., Gonzalez, J., Moeslund, T.B.: A local 3-D motion descriptor for multi-view human action recognition from 4-D spatio-temporal interest points. IEEE J. Sel. Top. Signal Process. **6**(5), 553–565 (2012)
31. Hong, B.W., Soatto, S., Ni, K., Chan, T.: The scale of a texture and its application to segmentation. In: Proceedings of Computer Vision and Pattern Recognition (CVPR 2008), pp. 1–8 (2008)
32. Hubel, D.H., Wiesel, T.N.: Receptive fields of single neurones in the cat's striate cortex. J. Physiol. **147**, 226–238 (1959)
33. Hubel, D.H., Wiesel, T.N.: Brain and Visual Perception: The Story of a 25-Year Collaboration. Oxford University Press, Oxford (2005)
34. Iijima, T.: Observation theory of two-dimensional visual patterns. Technical Report, Papers of Technical Group on Automata and Automatic Control, IECE, Japan (1962)

35. Jacobs, N., Pless, R.: Time scales in video surveillance. IEEE Trans. Circuits Syst. Video Technol. **18**(8), 1106–1113 (2008)
36. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: International Conference on Computer Vision (ICCV'07), pp. 1–8 (2007)
37. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
38. Jones, P.W., Le, T.M.: Local scales and multiscale image decompositions. Appl. Comput. Harmonic Anal. **26**(3), 371–394 (2009)
39. Kadir, T., Brady, M.: Saliency, scale and image description. Int. J. Comput. Vis. **45**(2), 83–105 (2001)
40. Kang, Y., Morooka, K., Nagahashi, H.: Scale invariant texture analysis using multi-scale local autocorrelation features. In: Proceedings of Scale Space and PDE Methods in Computer Vision (Scale-Space'05). Springer LNCS, vol. 3459, pp. 363–373 (2005). Springer
41. Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: Proceedings of Computer Vision and Pattern Recognition (CVPR'04), pp. II: 506–513. Washington, DC (2004)
42. Khan, N.Y., McCane, B., Wyvill, G.: SIFT and SURF performance evaluation against various image deformations on benchmark dataset. In: Proceedings of International Conference on Digital Image Computing Techniques and Applications (DICTA 2011), pp. 501–506 (2011)
43. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of British Machine Vision Conference, Leeds, UK (2008)
44. Koenderink, J.J.: The structure of images. Biol. Cybern. **50**, 363–370 (1984)
45. Koenderink, J.J.: Scale-time. Biol. Cybern. **58**, 159–162 (1988)
46. Koenderink, J.J., van Doorn, A.J.: Representation of local geometry in the visual system. Biol. Cybern. **55**, 367–375 (1987)
47. Koenderink, J.J., van Doorn, A.J.: Generic neighborhood operators. IEEE Trans. Pattern Anal. Mach. Intell. **14**(6), 597–605 (1992)
48. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: Proceedings of ECCV'04 Workshop on Spatial Coherence for Visual Motion Analysis, Springer LNCS, vol. 3667, pp. 91–103. Prague, Czech Republic (2004)
49. Laptev, I., Lindeberg, T.: Space-time interest points. In: Proceedings of International Conference on Computer Vision (ICCV 2003), pp. 432–439. Nice, France (2003)
50. Laptev, I., Lindeberg, T.: Velocity-adapted spatio-temporal receptive fields for direct recognition of activities. Image Vis. Comput. **22**(2), 105–116 (2004)
51. Laptev, I., Caputo, B., Schuldt, C., Lindeberg, T.: Local velocity-adapted motion events for spatio-temporal recognition. Comput. Vis. Image Underst. **108**, 207–229 (2007)
52. 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), *Springer LNCS*, vol. 7574, pp. III: 638–650. Springer (2012)
53. Li, Z., Gavves, E., Jain, M., Snoek, C.G.M.: VideoLSTM convolves, attends and flows for action recognition. arXiv preprint arXiv:1607.01794 (2016)
54. Li, Y., Tax, D.M.J., Loog, M.: Supervised scale-invariant segmentation (and detection). In: Proceedings of Scale Space and Variational Methods in Computer Vision (SSVM 2011), *Springer LNCS*, vol. 6667, pp. 350–361. Springer, Ein Gedi, Israel (2012)
55. Li, Y., Tax, D.M.J., Loog, M.: Scale selection for supervised image segmentation. Image Vis. Comput. **30**(12), 991–1003 (2012)
56. Lindeberg, T.: Scale-space for discrete signals. IEEE Trans. Pattern Anal. Mach. Intell. **12**(3), 234–254 (1990)
57. 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)
58. Lindeberg, T.: Effective scale: a natural unit for measuring scale-space lifetime. IEEE Trans. Pattern Anal. Mach. Intell. **15**(10), 1068–1074 (1993)
59. Lindeberg, T.: Scale-Space Theory in Computer Vision. Springer, Berlin (1993)
60. Lindeberg, T.: Scale-space theory: a basic tool for analysing structures at different scales. J. Appl. Stat. **21**(2), 225–270 (1994)
61. Lindeberg, T.: Linear spatio-temporal scale-space. In: ter Haar Romeny, B.M., Florack, L.M.J., Koenderink, J.J., Viergever, M.A. (eds.) Proceedings of International Conference on Scale-Space Theory in Computer Vision (Scale-Space'97), *Springer LNCS*, vol. 1252, pp. 113–127. Springer, Utrecht, The Netherlands (1997)
62. Lindeberg, T.: Principles for automatic scale selection. In: Handbook on Computer Vision and Applications, pp. 239–274. Academic Press, Boston, USA (1999). http://www.csc.kth.se/cvap/abstracts/cvap222.html
63. 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, Springer LNCS, vol. 1315, pp. 94–113. Kiel, Germany (1997)
64. Lindeberg, T.: Edge detection and ridge detection with automatic scale selection. Int. J. Comput. Vis. **30**(2), 117–154 (1998)
65. Lindeberg, T.: Feature detection with automatic scale selection. Int. J. Comput. Vis. **30**(2), 77–116 (1998)
66. Lindeberg, T.: A scale selection principle for estimating image deformations. Image Vis. Comput. **16**(14), 961–977 (1998)
67. 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)
68. Lindeberg, T.: Scale invariant feature transform. Scholarpedia **7**(5), 10,491 (2012)
69. Lindeberg, T.: A computational theory of visual receptive fields. Biol. Cybern. **107**(6), 589–635 (2013)
70. 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)
71. Lindeberg, T.: Invariance of visual operations at the level of receptive fields. PLoS ONE **8**(7), e66,990 (2013)
72. Lindeberg, T.: Scale selection properties of generalized scale-space interest point detectors. J. Math. Imaging Vis. **46**(2), 177–210 (2013)
73. Lindeberg, T.: Scale selection. In: Ikeuchi, K. (ed.) Computer Vision: A Reference Guide, pp. 701–713. Springer, Berlin (2014)
74. Lindeberg, T.: Image matching using generalized scale-space interest points. J. Math. Imaging Vis. **52**(1), 3–36 (2015)
75. Lindeberg, T.: Time-causal and time-recursive spatio-temporal receptive fields. J. Math. Imaging Vis. **55**(1), 50–88 (2016)
76. Lindeberg, T.: Dense scale selection over space, time and space-time. arXiv preprint arXiv:1709.08603 (2017)
77. Lindeberg, T.: Temporal scale selection in time-causal scale space. J. Math. Imaging Vis. **58**(1), 57–101 (2017)
78. Lindeberg, T.: Normative theory of visual receptive fields. arXiv preprint arXiv:1701.06333 (2017)
79. Lindeberg, T.: Spatio-temporal scale selection in video data. In: Proceedings of Scale-Space and Variational Methods for Computer Vision (SSVM 2017), *Springer LNCS*, vol. 10302, pp. 3–15. Kolding, Denmark (2017)
80. Lindeberg, T., Bretzner, L.: Real-time scale selection in hybrid multi-scale representations. In: Griffin, L., Lillholm, M. (eds.) Proc. Scale-Space Methods in Computer Vision (Scale-

Space'03), Springer LNCS, vol. 2695, pp. 148–163. Springer, Isle of Skye, Scotland (2003)

81. Lindeberg, T., Fagerström, D.: Scale-space with causal time direction. In: Proceedings of European Conference on Computer Vision (ECCV'96), Springer LNCS, vol. 1064, pp. 229–240. Cambridge, UK (1996)
82. Liu, X.M., Wang, C., Yao, H., Zhang, L.: The scale of edges. In: Proceedings of Computer Vision and Pattern Recognition (CVPR 2012), pp. 462–469 (2012)
83. Loog, M., Li, Y., Tax, D.: Maximum membership scale selection. In: Multiple Classifier Systems, Springer LNCS, vol. 5519, pp. 468–477. Springer (2009)
84. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
85. Luo, B., Aujol, J.F., Gousseau, Y.: Local scale measure from the topographic map and application to remote sensing images. Multiscale Model. Simul. **8**(1), 1–29 (2009)
86. Mainali, P., Lafruit, G., Yang, Q., Geelen, B., Gool, L.V., Lauwereins, R.: SIFER: Scale-invariant feature detector with error resilience. Int. J. Comput. Vis. **104**(2), 172–197 (2013)
87. Mainali, P., Lafruit, G., Tack, K., van Gool, L., Lauwereins, R.: Derivative-based scale invariant image feature detector with error resilience. IEEE Trans. Image Process. **23**(5), 2380–2391 (2014)
88. Maninis, K., Koutras, P., Maragos, P.: Advances on action recognition in videos using an interest point detector based on multiband spatio-temporal energies. In: International Conference on Image Processing (ICIP 2014), pp. 1490–1494 (2014)
89. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. Int. J. Comput. Vis. **60**(1), 63–86 (2004)
90. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005)
91. Mrázek, P., Navara, M.: Selection of optimal stopping time for nonlinear diffusion filtering. Int. J. Comput. Vis. **52**(2–3), 189–203 (2003)
92. Ng, J., Bharath, A.A.: Steering in scale space to optimally detect image structures. In: Proceedings of European Conference on Computer Vision (ECCV 2004), Springer LNCS, vol. 3021, pp. 482–494 (2004)
93. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. Int. J. Comput. Vis. **79**(3), 299–318 (2008)
94. Oikonomopoulos, A., Patras, I., Pantic, M.: Spatiotemporal salient points for visual recognition of human actions. IEEE Trans. Syst. Man Cybern. Part B **36**(3), 710–719 (2005)
95. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
96. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. arXiv preprint arXiv:1611.00850 (2016)
97. Rapantzikos, K., Avrithis, Y., Kollias, S.: Dense saliency-based spatiotemporal feature points for action recognition. In: Proceedings of Computer Vision and Pattern Recognition (CVPR 2009), pp. 1454–1461 (2009)
98. Rivero-Moreno, C.J., Bres, S.: Spatio-temporal primitive extraction using Hermite and Laguerre filters for early vision video indexing. In: Image Analysis and Recognition. Springer LNCS , vol.**3211**, pp. 825–832 (2004)
99. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional SIFT descriptor and its application to action recognition. In: Proceedings of ACM International Conference on Multimedia, pp. 357–360 (2007)
100. Shabani, A.H., Clausi, D.A., Zelek, J.S.: Evaluation of local spatio-temporal salient feature detectors for human action recognition. In: Proceedings of Computer and Robot Vision (CRV 2012), pp. 468–475 (2012)
101. Shabani, A.H., Clausi, D.A., Zelek, J.S.: Improved spatio-temporal salient feature detection for action recognition. In: British Machine Vision Conference (BMVC'11), pp. 1–12. Dundee, UK (2011)
102. Shao, L., Mattivi, R.: Feature detector and descriptor evaluation in human action recognition. In: Proceedings of ACM International Conference on Image and Video Retrieval (CIVR'10), pp. 477–484. Xian, China (2010)
103. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems (NIPS 2014), pp. 568–576 (2014)
104. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human action classes from videos in the wild. Tech. Rep. CRCV-TR-12-01, Center for Research in Computer Vision, University of Central Florida (2012). arXiv preprint arXiv:1212.0402
105. Sporring, J., Colios, C.J., Trahanias, P.E.: Generalized scale selection. In: Proceedings of International Conference on Image Processing (ICIP'00), pp. 920–923. Vancouver, Canada (2000)
106. Sporring, J., Nielsen, M., Florack, L., Johansen, P. (eds.): Gaussian Scale-Space Theory: Proceedings of PhD School on Scale-Space Theory. Series in Mathematical Imaging and Vision. Springer, Copenhagen, Denmark (1997)
107. Stöttinger, J., Hanbury, A., Sebe, N., Gevers, T.: Sparse color interest points for image retrieval and object categorization. IEEE Trans. Image Process. **21**(5), 2681–2692 (2012)
108. Tamrakar, A., Ali, S., Yu, Q., Liu, J., Javed, O., Divakaran, A., Cheng, H., Sawhney, H.: Evaluation of low-level features and their combinations for complex event detection in open source videos. In: Proceedings of Computer Vision and Pattern Recognition (CVPR 2012), pp. 3681–3688 (2012)
109. Tau, M., Hassner, T.: Dense correspondences across scenes and scales. IEEE Trans. Pattern Anal. Mach. Intell. **38**(5), 875–888 (2016)
110. ter Haar Romeny, B., Florack, L., Nielsen, M.: Scale-time kernels and models. In: Proceedings of International Conference on Scale-Space and Morphology in Computer Vision (Scale-Space'01), Springer LNCS. Springer, Vancouver, Canada (2001)
111. ter Haar Romeny, B.: Front-End Vision and Multi-scale Image Analysis. Springer, Berlin (2003)
112. Tuytelaars, T., Mikolajczyk, K.: A Survey on Local Invariant Features, Foundations and Trends in Computer Graphics and Vision, vol. 3(3). Now Publishers (2008)
113. Tuytelaars, T., van Gool, L.: Matching widely separated views based on affine invariant regions. Int. J. Comput. Vis. **59**(1), 61–85 (2004)
114. 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)
115. Vanhamel, I., Mihai, C., Sahli, H., Katartzis, A., Pratikakis, I.: Scale selection for compact scale-space representation of vector-valued images. Int. J. Comput. Vis. **84**(2), 194–204 (2009)
116. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Proceedings of Computer Vision and Pattern Recognition (CVPR 2011), pp. 3169–3176 (2011)
117. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), pp. 4305–4314 (2015)
118. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of International Conference on Computer Vision (ICCV 2013), pp. 3551–3558 (2013)
119. Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: Proceedings of British Machine Vision Conference (BMVC 2009). London, UK (2009)

120. Weickert, J., Ishikawa, S., Imiya, A.: Linear scale-space has first been proposed in Japan. J. Math. Imaging Vis. **10**(3), 237–252 (1999)
121. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Comput. Vis. Image Underst. **115**(2), 224–241 (2011)
122. Willems, G., Tuytelaars, T., van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Proceedings og European Conference on Computer Vision (ECCV 2008), *Springer LNCS*, vol. 5303, pp. 650–663. Marseille, France (2008)
123. Witkin, A.P.: Scale-space filtering. In: Proceedings of 8th International Joint Conference on Artificial Intelligence, pp. 1019–1022. Karlsruhe, Germany (1983)
124. Wong, S.F., Cipolla, R.: Extracting spatiotemporal interest points using global information. In: International Conference on Computer Vision (ICCV 2007), pp. 1–8 (2007)
125. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. In: Proceedings of Computer Vision and Pattern Recognition (CVPR'01), pp. II: 123–130 (2001)
126. Zhen, X., Shao, L.: Action recognition via spatio-temporal local features: a comprehensive study. Image Vis. Comput. **50**, 1–13 (2016)
127. Zhu, Y., Chen, W., Guo, G.: Evaluating spatiotemporal interest point features for depth-based action recognition. Image Vis. Comput. **32**(8), 453–464 (2014)

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 and 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.