<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *SSVM 2017: 6th International Conference on Scale Space and Variational Methods in Computer Vision*, Kolding, Denmark.

Citation for the original published paper:

Lindeberg, T. (2017)

Spatio-temporal scale selection in video data.

In: *Scale Space and Variational Methods in Computer Vision* (pp. 3-15). Springer-Verlag Tokyo Inc.

Springer Lecture Notes in Computer Science

https://doi.org/10.1007/978-3-319-58771-4_1

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-202415>

Spatio-temporal scale selection in video data^{*}

Tony Lindeberg

Computational Brain Science Lab
Department of Computational Science and Technology
School of Computer Science and Communication
KTH Royal Institute of Technology, Stockholm, Sweden

Abstract. We present a theory and a method 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. A theoretical analysis is given of the scale selection properties of six types of spatio-temporal interest point detectors, showing that five of them 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.

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 (Hubel and Wiesel [1]; DeAngelis *et al.* [2]; Zelnik-Manor and Irani [3]; Laptev and Lindeberg [4]; Simonyan and Zisserman [5]).

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. Figure 1 show illustrations of this problem by showing snapshots of spatio-temporal

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

Fig. 1. Time-causal spatio-temporal scale-space representation $L(x, y, t; s, \tau)$ with its second-order temporal derivative $L_{tt}(x, y, t; s, \tau)$ and the spatial Laplacian of the second-order temporal derivative $\nabla_{(x,y)}^2 L_{tt}$ 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) $\sigma_{s,1} = 10.6$ pixels and the temporal scales (left column) $\sigma_{\tau,1} = 40$ ms, (middle column) $\sigma_{\tau,2} = 160$ ms and (right column) $\sigma_{\tau,2} = 640$ ms 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 per second.)

receptive field responses over multiple spatial and temporal scales for a video sequence and 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. 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.

For purely spatial data, the problem of performing spatial scale selection is nowadays rather well understood. Given the spatial Gaussian scale-space concept (Koenderink [6]; Lindeberg [7, 8]; Florack [9]; ter Haar Romeny [10]), a general methodology for performing spatial scale selection has been developed based on local extrema over spatial scales of scale-normalized differential entities (Lindeberg [11]). This general methodology has in turn been successfully applied to develop robust methods for image-based matching and recognition that are able to handle large variations of the size of the objects in the image domain.

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 (Lindeberg [12]; Laptev and Lindeberg [13]; Willems *et al.* [14]), these methods suffer from either theoretical or practical limitations. The subject of this article is to present an extended theory for spatio-temporal scale selection in video data.

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

$$T(x_1, x_2, t; s, \tau; v, \Sigma) = g(x_1 - v_1 t, x_2 - v_2 t; s, \Sigma) h(t; \tau) \quad (1)$$

as previously derived, proposed and studied in Lindeberg [8, 15, 16], where we specifically here choose as temporal smoothing kernel over time either (i) the non-causal Gaussian kernel

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

or (ii) the time-causal limit kernel (Lindeberg [16, Equation (38)])

$$h(t; \tau) = \Psi(t; \tau, c) \quad (3)$$

defined via its Fourier transform of the form

$$\hat{\Psi}(\omega; \tau, c) = \prod_{k=1}^{\infty} \frac{1}{1 + i c^{-k} \sqrt{c^2 - 1} \sqrt{\tau} \omega}. \quad (4)$$

For simplicity, we consider space-time separable receptive fields with image velocity $v = (v_1, v_2) = (0, 0)$ and set the spatial covariance matrix to $\Sigma = I$.

3 Spatial-temporal scale covariance and scale invariance

In the corresponding spatio-temporal scale-space representation

$$L(x_1, x_2, t; s, \tau) = (T(\cdot, \cdot, \cdot; s, \tau) * f(\cdot, \cdot, \cdot))(x_1, x_2, t; s, \tau) \quad (5)$$

we define scale-normalized partial derivatives (Lindeberg [16, Equation (108)])

$$L_{x_1^{m_1} x_2^{m_2} t^n, norm} = s^{(m_1+m_2)\gamma_s/2} \tau^{n\gamma_\tau/2} L_{x_1^{m_1} x_2^{m_2} t^n} \quad (6)$$

and consider homogeneous spatio-temporal differential invariants of the form

$$\mathcal{D}L = \sum_{i=1}^I \prod_{j=1}^J c_i L_{x^{\alpha_{ij}} t^{\beta_{ij}}}, = \sum_{i=1}^I \prod_{j=1}^J c_i L_{x_1^{\alpha_{1ij}} x_2^{\alpha_{2ij}} t^{\beta_{ij}}}, \quad (7)$$

where the sum of the orders of spatial and temporal differentiation in a certain term $\sum_{j=1}^J |\alpha_{ij}| = \sum_{j=1}^J \alpha_{1ij} + \alpha_{2ij} = M$ and $\sum_{j=1}^J \beta_{ij} = N$ does not depend on the index i of that term. Consider next an independent scaling transformation of the spatial and the temporal domains of a video sequence

$$f'(x'_1, x'_2, t') = f(x_1, x_2, t) \quad \text{for} \quad (x'_1, x'_2, t') = (S_s x_1, S_s x_2, S_\tau t) \quad (8)$$

where S_s and S_τ denote the spatial and temporal scaling factors, respectively. Then, a homogeneous spatio-temporal derivative expression of the form (7) with the spatio-temporal derivatives $L_{x_1^{m_1} x_2^{m_2} t^n}$ replaced by scale-normalized spatio-temporal derivatives $L_{x_1^{m_1} x_2^{m_2} t^n, norm}$ according to (6) transforms according to

$$D'_{norm} L' = S_s^{M(\gamma_s-1)} S_\tau^{N(\gamma_\tau-1)} D_{norm} L. \quad (9)$$

This result follows from a combination and generalization of Equation (25) in (Lindeberg [11]) with Equations (10) and (104) in (Lindeberg [17]).

With the temporal smoothing performed by the scale-invariant limit kernel according to (3), the temporal scaling transformation property does, however, only hold for temporal scaling transformations that correspond to exact mappings between the discrete temporal scale levels 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.

Specifically, the scale covariance property (9) implies that if we can detect a spatio-temporal level $(\hat{s}, \hat{\tau})$ such that the scale-normalized expression $D'_{norm} L'$ assumes a local extremum at a point $(\hat{x}, \hat{y}, \hat{t}; \hat{s}, \hat{\tau})$ 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

$$(\hat{x}, \hat{y}, \hat{t}; \hat{s}, \hat{\tau}) \mapsto (S_s \hat{x}, S_s \hat{y}, S_\tau \hat{t}; S_s^2 \hat{s}, S_\tau^2 \hat{\tau}). \quad (10)$$

The properties (9) and (10) constitute a theoretical foundation for scale-covariant and scale-invariant spatio-temporal feature detection.

4 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 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 analyze 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 of the underlying spatio-temporal image structures.

4.1 Differential entities for spatio-temporal interest point detection

Inspired by the way neurons in the lateral geniculate nucleus (LGN) respond to visual input (DeAngelis et al [2]), which for many LGN cells can be modelled by idealized operations of the form (Lindeberg [15, Equation (108)])

$$h_{LGN}(x, y, t; s, \tau) = \pm(\partial_{xx} + \partial_{yy}) g(x, y; s) \partial_{t^n} h(t; \tau), \quad (11)$$

we consider the scale-normalized Laplacian of the first- and second-order temporal derivatives

$$\nabla_{(x,y),norm}^2 L_{t,norm} = s^{\gamma_s} \tau^{\gamma_\tau / 2} (L_{xxt} + L_{yyt}), \quad (12)$$

$$\nabla_{(x,y),norm}^2 L_{tt,norm} = s^{\gamma_s} \tau^{\gamma_\tau} (L_{xxtt} + L_{yytt}). \quad (13)$$

Inspired by the way the determinant of the spatial Hessian matrix constitutes a better spatial interest point detector than the spatial Laplacian operator (Lindeberg [18]), we consider extensions of the spatial Laplacian of the first- and second-order temporal derivatives into the determinant of the spatial Hessian of the first- and second-order temporal derivatives

$$\det \mathcal{H}_{(x,y),norm} L_{t,norm} = s^{2\gamma_s} \tau^{\gamma_\tau} (L_{xxt} L_{yyt} - L_{xyt}^2), \quad (14)$$

$$\det \mathcal{H}_{(x,y),norm} L_{tt,norm} = s^{2\gamma_s} \tau^{2\gamma_\tau} (L_{xxtt} L_{yytt} - L_{xytt}^2). \quad (15)$$

Over three-dimensional joint space-time, we can also define the determinant of the spatio-temporal Hessian

$$\begin{aligned} \det \mathcal{H}_{(x,y,t),norm} L = s^{2\gamma_s} \tau^{\gamma_\tau} (& L_{xx} L_{yy} L_{tt} + 2L_{xy} L_{xt} L_{yt} \\ & - L_{xx} L_{yt}^2 - L_{yy} L_{xt}^2 - L_{tt} L_{xy}^2) \end{aligned} \quad (16)$$

and make an attempt to define a spatio-temporal Laplacian of the form

$$\nabla_{(x,y,t),norm}^2 L = s^{\gamma_s} (L_{xx} + L_{yy}) + \varkappa^2 \tau^{\gamma_\tau} L_{tt}, \quad (17)$$

where we have introduced a parameter $\varkappa^2$ to make explicit the arbitrary scaling factor between temporal *vs.* spatial derivatives that influences any attempt to add derivative expressions of different dimensionality.

4.2 Scale calibration for a Gaussian blink

Consider a spatio-temporal model signal defined as a Gaussian blink with spatial extent s_0 and temporal duration τ_0

$$f(x, y, t) = g(x, y; s_0) g(t; \tau_0) = \frac{1}{(2\pi)^{3/2} s_0 \sqrt{\tau_0}} e^{-(x^2+y^2)/2s_0} e^{-t^2/2\tau_0}, \quad (18)$$

for which the spatio-temporal scale-space representation is of the form

$$L(x, y, t; s, \tau) = g(x, y; s_0 + s) g(t; \tau_0 + \tau). \quad (19)$$

To calibrate the scale selection properties of the spatio-temporal interest point detectors $\nabla_{(x,y),norm}^2 L_{tt,norm}$, $\det \mathcal{H}_{(x,y),norm} L_{tt,norm}$ and $\det \mathcal{H}_{(x,y,t),norm} L$, we require that at the origin $(x, y, t) = (0, 0, 0)$ the scale-normalized spatio-temporal differential expression should assume its strongest extremum over spatial and temporal scales at spatio-temporal scale level

$$\hat{s} = s_0, \quad \hat{\tau} = q^2 \tau_0. \quad (20)$$

Here, we have introduced a parameter $q < 1$ to enforce temporal scale calibration to a finer temporal scale than τ_0 , to enable shorter temporal delays and thus faster responses for feature detection performed based on a time-causal spatio-temporal scale-space representation. By calculating the scale-normalized feature response at the origin for each one of the feature detectors, differentiating with respect to the spatial and temporal scale parameters and solving for the zero-crossings over spatial and temporal scales, we find that the spatial scale-normalization powers γ_s and γ_τ in the scale-normalized spatio-temporal derivatives (6) should be set to $\gamma_s = 1$ for the slice-based operators $\nabla_{(x,y)}^2 L_{tt}$ and $\det \mathcal{H}_{(x,y)} L_{tt}$, to $\gamma_s = 5/4$ for the genuine 3-D operator $\det \mathcal{H}_{(x,y,t)}$ and to

$$\gamma_{\tau, \nabla_{(x,y)}^2 L_{tt}} = \gamma_{\tau, \det \mathcal{H}_{(x,y)} L_{tt}} = \frac{3q^2}{2(q^2 + 1)}, \quad (21)$$

$$\gamma_{\tau, \det \mathcal{H}_{(x,y,t)}} = \frac{5q^2}{2(q^2 + 1)}. \quad (22)$$

4.3 Scale calibration for a Gaussian onset blob

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

$$f(x, y, t) = g(x, y; s_0) \int_{u=0}^t g(u; \tau_0) du = \frac{1}{(2\pi)^{3/2} s_0 \sqrt{\tau_0}} e^{-(x^2+y^2)/2s_0} \int_{u=0}^t e^{-u^2/2\tau_0} du, \quad (23)$$

for which the spatio-temporal scale-space representation is of the form

$$L(x, y, t; s, \tau) = g(x, y; s_0 + s) \int_{u=0}^t g(u; \tau_0 + \tau) du. \quad (24)$$

To calibrate the scale selection properties of the spatio-temporal interest point detectors $\nabla_{(x,y),norm}^2 L_{t,norm}$ and $\det \mathcal{H}_{(x,y),norm} L_{t,norm}$, we require that at the origin $(x, y, t) = (0, 0, 0)$ the scale-normalized spatio-temporal differential entity should assume its strongest extremum at spatio-temporal scale level

$$\hat{s} = s_0, \quad \hat{\tau} = q^2 \tau_0. \quad (25)$$

By again calculating the scale-normalized response at the origin and differentiating with respect to the spatial and temporal scale parameters, we find that the spatial scale-normalization powers γ_s and γ_τ in the scale-normalized spatio-temporal derivatives (6) should for both operators be set to $\gamma_s = 1$ and

$$\gamma_\tau = \frac{q^2}{q^2 + 1}. \quad (26)$$

4.4 Lack of scale covariance for the spatio-temporal Laplacian

For the attempt to define a scale-normalized spatio-temporal Laplacian operator

$$\nabla_{(x,y,t),norm}^2 L = s^{\gamma_s} (L_{xx} + L_{yy}) + \varkappa^2 \tau^{\gamma_\tau} L_{tt}, \quad (27)$$

the equations that determine the spatial and temporal scale estimates are unfortunately hard to solve in closed form for general values of the scale normalization powers γ_s and γ_τ . For the specific case of $\gamma_s = 1$ and $\gamma_\tau = 1$, we can, however, note that the resulting scale estimates

$$\hat{s} = s_0 \left(\frac{5}{2 + \varkappa^2} - 1 \right) \quad \text{and} \quad \hat{\tau} = 2\tau_0 \left(2 - \frac{5}{2 + \varkappa^2} \right) \quad (28)$$

will be explicitly dependent on the relative scaling factor $\varkappa^2$ between the derivatives with respect to the temporal *vs.* the spatial dimensions. This situation is in clear contrast to the other spatio-temporal differential entities treated above, for which a multiplication of the temporal derivative by a factor $\varkappa$ does not affect the spatial or temporal scale estimates. This property does in turn imply that the spatio-temporal scale estimates will not be covariant under independent relative rescalings of the spatial and temporal dimensions. These theoretical arguments explain why the scale estimates from the spatial and temporal selection mechanisms in [13] were later empirically found to not be sufficiently robust.

5 Spatio-temporal interest points detected as spatio-temporal scale-space extrema over space-time

In this section, we shall use the above scale-normalized differential entities 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 $(\mathcal{D}_{norm} L)(x, y, t; s, \tau)$ simultaneously assumes local extrema with respect to both space-time (x, y, t) and spatio-temporal scales (s, τ) .

5.1 Time-causal and time-recursive algorithm for spatio-temporal scale-space extrema detection

By approximating the spatial smoothing operation by convolution with the discrete analogue of the Gaussian kernel over the spatial domain [7], which obeys a semi-group property over spatial scales, and approximating the time-causal limit kernel by a cascade of first-order recursive filters [16, Equation (56)], we can state an algorithm for computing the time-causal and time-recursive spatio-temporal scale-space representation and detecting spatio-temporal scale-space extrema of scale-normalized differential invariants from it as follows:

1. Determine a set of 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 $\mu_k = (\sqrt{1 + 4r^2(\tau_k - \tau_{k-1})} - 1)/2$ according to (Lindeberg [16, Equations (58) and (55)]) for approximating the time-causal limit kernel by a cascade of recursive filters, where r denotes the frame rate and the temporal scale levels τ_k are given in units of [seconds]².
3. For each temporal scale level, initiate a temporal buffer for temporal scale-space smoothing at this temporal scale $B(x, y, k) = f(x, y, 0)$.
4. 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.)
5. Loop forwards over time t (in units of time steps):
 - (a) Loop over the temporal scale levels k in ascending order:
 - i. Perform temporal smoothing according to (with $B(x, y, 0) = f(x, y, t)$)

$$B(x, y, k) := B(x, y, k) + \frac{1}{1 + \mu_k} (B(x, y, k - 1) - B(x, y, k)). \quad (29)$$

- ii. Initiate spatio-temporal scale-space representation at current temporal scale $L(x, y, t; s_1, \tau_k) = T(x, y; s_1) * B(x, y, k)$.
- iii. Loop over the spatial scale levels s_l in ascending order:
 - A. Compute spatio-temporal representations at coarser spatial scales using the semi-group property over spatial scales

$$L(\cdot, \cdot, t; s_l, \tau_k) = T(\cdot, \cdot; s_l - s_{l-1}) * L(\cdot, \cdot, t; s_{l-1}, \tau_k). \quad (30)$$

- (b) For all temporal and spatial scales, compute temporal derivatives using backward differences over the short temporal buffers of past frames.
- (c) For all temporal and spatial scales, compute the scale-normalized differential entity $(\mathcal{D}_{norm} L)(x, y, t; s_l, \tau_k)$ at that spatio-temporal scale.
- (d) 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 $|(\mathcal{D}_{postnorm} L)(x, y, t; s_l, \tau_k)| \geq \theta$, determine if the point is either a positive maximum or a negative minimum in comparison to 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 [17, 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 order the comparisons with respect to the nearest neighbours in each dimension and 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 or a local minimum.

5.2 Experimental results

Figure 2 shows the result of detecting spatio-temporal interest points in this way from the same video sequence as used for the illustrations in figure 1. For this experiment, 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 distribution parameter $c = 2$ for the time-causal limit kernel. To obtain comparable numbers of features from the different feature detectors, we adapted the thresholds on the scale-normalized differential invariants $\nabla_{(x,y),norm}^2 L_{t,norm}$, $\nabla_{(x,y),norm}^2 L_{tt,norm}$, $\det \mathcal{H}_{(x,y),norm} L_{t,norm}$, $\det \mathcal{H}_{(x,y),norm} L_{tt,norm}$, $\det \mathcal{H}_{(x,y,t),norm} L$ and $\nabla_{(x,y,t),norm}^2 L$ such that the average number of features from each feature detector was 50 features per frame.

As can be seen from the results, all of these 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 $\nabla_{(x,y),norm}^2 L_{t,norm}$ and $\nabla_{(x,y),norm}^2 L_{tt,norm}$ 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 $\det \mathcal{H}_{(x,y),norm} L_{t,norm}$ and $\det \mathcal{H}_{(x,y),norm} L_{tt,norm}$. The more corner-detector-inspired feature detector $\det \mathcal{H}_{(x,y,t),norm} L$ responds more to image features where there are simultaneously rapid variations over all the spatial and temporal dimensions.

5.3 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 [11]) in combination with the temporal scale selection properties of the temporal scale selection theory in (Lindeberg [17]) with the scale covariance of the underlying spatio-temporal derivative expressions $\nabla_{(x,y),norm}^2 L_{t,norm}$, $\nabla_{(x,y),norm}^2 L_{tt,norm}$, $\det \mathcal{H}_{(x,y),norm} L_{t,norm}$, $\det \mathcal{H}_{(x,y),norm} L_{tt,norm}$ and $\det \mathcal{H}_{(x,y,t),norm} L$ described in (Lindeberg [16]), it follows that these spatio-temporal interest point detectors

Fig. 2. 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 $c = 2$ based on recursive filters coupled in cascade 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 per second.)

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 $g(t; \tau)$ over the temporal domain or the time-causal limit kernel $\Psi(t; \tau, c)$. From the general proof in Section 3, it follows that the detected spatio-temporal interest points transform in a scale-covariant way under independent scaling transformations of the spatial and the temporal domains.

6 Summary and discussion

We have presented a theory and a method for performing simultaneous detection of local 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 results regarding covariance and invariance properties of spatio-temporal features defined from video data with independent scaling transformations of the spatial and the temporal domains. For five 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 and (v) the determinant of the spatio-temporal Hessian matrix, we have analysed the theoretical scale selection properties of these feature detectors and shown how scale calibration of these feature detectors can be performed to make the spatio-temporal scale estimates reflect the spatial extent and the temporal duration of the underlying spatio-temporal features that gave rise to the feature responses.

For 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 [13].

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 a 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 : $\Delta t \mapsto q \Delta t$. An experimental quantification of the scale selection properties in time-causal spatio-temporal scale space in a longer version of this paper 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.

References

1. Hubel, D.H., Wiesel, T.N.: Brain and Visual Perception: The Story of a 25-Year Collaboration. Oxford University Press (2005)
2. DeAngelis, G.C., Ohzawa, I., Freeman, R.D.: Receptive field dynamics in the central visual pathways. *Trends in Neuroscience* **18** (1995) 451–457
3. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. In: Proc. Computer Vision and Pattern Recognition (CVPR’01). (2001) II:123–130
4. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: Proc. ECCV’04 Workshop on Spatial Coherence for Visual Motion Analysis. Volume 3667 of Springer LNCS., Prague, Czech Republic (2004) 91–103
5. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems. (2014) 568–576
6. Koenderink, J.J.: The structure of images. *Biol. Cyb.* **50** (1984) 363–370
7. Lindeberg, T.: Scale-Space Theory in Computer Vision. Springer (1993)
8. Lindeberg, T.: Generalized Gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space. *Journal of Mathematical Imaging and Vision* **40** (2011) 36–81
9. Florack, L.M.J.: Image Structure. Springer (1997)
10. ter Haar Romeny, B.: Front-End Vision and Multi-Scale Image Analysis. Springer (2003)
11. Lindeberg, T.: Feature detection with automatic scale selection. *Int. J. Comp. Vis.* **30** (1998) 77–116
12. Lindeberg, T.: On automatic selection of temporal scales in time-causal scale-space. In: Proc. AFPAC’97. Volume 1315 of Springer LNCS. (1997) 94–113
13. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV. (2003) 432–439
14. Willems, G., Tuytelaars, T., van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: ECCV’08. Volume 5303 of Springer LNCS. (2008) 650–663
15. Lindeberg, T.: A computational theory of visual receptive fields. *Biological Cybernetics* **107** (2013) 589–635
16. Lindeberg, T.: Time-causal and time-recursive spatio-temporal receptive fields. *Journal of Mathematical Imaging and Vision* **55** (2016) 50–88
17. Lindeberg, T.: Temporal scale selection in time-causal scale space. *Journal of Mathematical Imaging and Vision* (2017) doi:10.1007/s10851-016-0691-3.
18. Lindeberg, T.: Image matching using generalized scale-space interest points. *Journal of Mathematical Imaging and Vision* **52** (2015) 3–36