



Dynamic Texture Recognition Using Time-Causal and Time-Recursive Spatio-Temporal Receptive Fields

Ylva Jansson¹ · Tony Lindeberg¹

Received: 14 September 2017 / Accepted: 29 May 2018 / Published online: 20 June 2018
© The Author(s) 2018

Abstract

This work presents a first evaluation of using spatio-temporal receptive fields from a recently proposed time-causal spatio-temporal scale-space framework as primitives for video analysis. We propose a new family of video descriptors based on regional statistics of spatio-temporal receptive field responses and evaluate this approach on the problem of dynamic texture recognition. Our approach generalises a previously used method, based on joint histograms of receptive field responses, from the spatial to the spatio-temporal domain and from object recognition to dynamic texture recognition. The time-recursive formulation enables computationally efficient time-causal recognition. The experimental evaluation demonstrates competitive performance compared to state of the art. In particular, it is shown that binary versions of our dynamic texture descriptors achieve improved performance compared to a large range of similar methods using different primitives either handcrafted or learned from data. Further, our qualitative and quantitative investigation into parameter choices and the use of different sets of receptive fields highlights the robustness and flexibility of our approach. Together, these results support the descriptive power of this family of time-causal spatio-temporal receptive fields, validate our approach for dynamic texture recognition and point towards the possibility of designing a range of video analysis methods based on these new time-causal spatio-temporal primitives.

Keywords Dynamic texture · Receptive field · Spatio-temporal · Time-causal · Time-recursive · Video descriptor · Receptive field histogram · Scale space

1 Introduction

The ability to derive properties of the surrounding world from time-varying visual input is a key function of a general-purpose computer vision system and necessary for any artificial or biological agent that is to use vision for interpreting a dynamic environment. Motion provides additional cues for understanding a scene and some tasks by necessity require motion information, e.g. distinguishing between

events or actions with similar spatial appearance or estimating the speed of moving objects. Motion cues are also helpful when other visual cues are weak or contradictory. Thus, understanding how to represent and incorporate motion information for different video analysis tasks—including the question of what constitute meaningful spatio-temporal features—is a key factor for successful applications in areas such as action recognition, automatic surveillance, dynamic texture and scene understanding, video indexing and retrieval, autonomous driving, etc.

Challenges when dealing with spatio-temporal image data are similar to those present in the static case, that is high-dimensional data with large intraclass variability caused both by the diverse appearance of conceptually similar objects and by the presence of a large number of identity preserving visual transformations. From the spatial domain, we inherit the basic image transformations: translations, scalings, rotations, nonlinear perspective transformations and illumination transformations. In addition to these, spatio-temporal data will contain additional sources of variability: differences

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

✉ Ylva Jansson
yjansson@kth.se
Tony Lindeberg
tony@kth.se

¹ Computational Brain Science Lab, Department of Computational Science and Technology, KTH Royal Institute of Technology, Stockholm, Sweden

in motion patterns for conceptually similar motions, events occurring faster or slower, velocity transformations caused by camera motion and time-varying illumination. Moreover, the larger data dimensionality compared to static images presents additional computational challenges.

For biological vision, local image measurements in terms of receptive fields constitute the first processing layers [11,25]. In the area of scale-space theory, it has been shown that Gaussian derivatives and related operators constitute a canonical model for visual receptive fields [17,26,30–33,39,41,67,69,70,78,80]. In computer vision, spatial receptive fields based on the Gaussian scale-space concept have been demonstrated to be a powerful front-end for solving a large range of visual tasks. Theoretical properties of scale-space filters enable the design of methods invariant or robust to natural image transformations [41–44]. Also, early processing layers based on such “ideal” receptive fields can be shared among multiple tasks and thus free resources both for learning higher-level features from data and during online processing.

The most straightforward extension of the Gaussian scale-space concept to the spatio-temporal domain is to use Gaussian smoothing also over the temporal domain. However, for real-time processing or when modelling biological vision this would violate the fundamental *temporal causality* constraint present for real-time tasks: it is simply not possible to access the future. The ad hoc solution of using a time-delayed truncated Gaussian temporal kernel would instead imply unnecessarily long temporal delays, which would make the framework less suitable for time-critical applications.

The preferred option is to use truly time-causal visual operators. Recently, a new time-causal spatio-temporal scale-space framework, leading to a new family of *time-causal spatio-temporal receptive fields*, has been introduced by Lindeberg [44]. In addition to temporal causality, the time-recursive formulation of the temporal smoothing operation offers computational advantages compared to Gaussian filtering over time in terms of fewer computations and a compact recursively updated memory of the past.

The idealised spatio-temporal receptive fields derived within that framework also have a strong connection to biology in the sense of very well modelling both spatial and spatio-temporal receptive field shapes of neurons in the LGN and V1 [42,44]. This similarity further motivates designing algorithms based on these primitives. They provably work well in a biological system, which points towards their applicability also for artificial vision. An additional motivation, although not actively pursued here, is that designing methods based on primitives similar to biological receptive fields can enable a better understanding of information processing in biological systems.

The purpose of this study is a first evaluation of using this new family of time-causal spatio-temporal receptive fields as visual primitives for video analysis. As a first application, we have chosen the problem of dynamic texture recognition. A dynamic texture or spatio-temporal texture is an extension of texture to the spatio-temporal domain and can be naively defined as “texture + motion” or more formally as a spatio-temporal pattern that exhibits certain stationarity or self-similar properties over both space and time [6]. Examples of dynamic textures are windblown vegetation, fire, waves, a flock of flying birds or a flag flapping in the wind. Recognising different types of dynamic textures is important for visual tasks such as automatic surveillance (e.g. detecting forest fires), video indexing and retrieval (e.g. return all images set on the sea) and to enable artificial agents to understand and interact with the world by interpreting different environments.

In this work, we start by presenting a new family of video descriptors in the form of joint histograms of spatio-temporal receptive field responses. Our approach generalises a previous method by Linde and Lindeberg [38] from the spatial to the spatio-temporal domain and from object recognition to dynamic texture recognition. We subsequently perform an experimental evaluation on two commonly used dynamic texture datasets and present results on:

- (i) Qualitative and quantitative effects from varying model parameters including the spatial and the temporal scales and the number of principal components and the number of bins used in the histogram descriptor.
- (ii) A comparison between the performance of descriptors constructed from different sets of receptive fields.
- (iii) An extensive comparison with state-of-the-art dynamic texture recognition methods.

Our benchmark results demonstrate competitive performance compared to state-of-the-art dynamic texture recognition methods. Notably, we achieve very good results using a conceptually simple method that only makes use of local information and with the additional constraint of using time-causal operators. In particular, it is shown that binary versions of our dynamic texture descriptors achieve improved performance compared to a large range of similar methods using different primitives either handcrafted or learned from data. Further, our qualitative and quantitative investigation into parameter choices and the use of different sets of receptive fields highlights the robustness and flexibility of our approach.

Together, these results support the descriptive power of this family of time-causal spatio-temporal receptive fields, validate our approach for dynamic texture recognition and point towards the possibility of designing a range of video

analysis methods based on these new time-causal spatio-temporal primitives.

This paper is an extended version of a conference paper [27], in which only a single descriptor version was evaluated. We here present a more extensive experimental evaluation, including results on varying the descriptor parameters and of using different sets of receptive fields. We specifically present new video descriptors with improved performance compared to the previous video descriptor used in [27].

2 Related Work

For dynamic texture recognition, additional challenges compared to the static case include variabilities in motion patterns and speed and much higher data dimensionality. Although spatial texture descriptors can give reasonable performance also in the dynamic case (a human observer can typically distinguish dynamic textures based on a single frame), making use of dynamic information as well provides a consistent advantage over purely spatial descriptors (see, e.g. [1,3,24,56,87]). Thus, a large range of *dynamic* texture recognition methods have been developed, exploring different options for representing motion information and combining this with spatial information.

Some of the first methods for dynamic texture recognition were based on *optic flow*, see, e.g. the pioneering work by Nelson and Polana [52] and later work by Péteri and Chetverikov [54], Lu et al. [49], Fazekas and Chetverikov [16], Fazekas et al. [15] and Crivelli et al. [7].

Compared to the motion patterns of larger rigid objects, however, dynamic textures often feature chaotic non-smooth motions, multiple directions of motion at the same image point and intensity changes not mediated by rigid motion—consider, e.g. fluttering leaves, turbulent water and fire. Thus, the brightness constancy assumption, usually underlying optical flow estimation, is violated for many dynamic texture types. This implies difficulties estimating optic flow for dynamic textures, although alternative types of assumptions, such as brightness conservation and colour constancy, have later been explored, for example, in the context of dynamic texture detection and segmentation [5,15].

Another early approach, first proposed by Soatto et al. [66], is to model the time-evolving appearance of a dynamic texture as a *linear dynamical system* (LDS). Although originally designed for synthesis, recognition can be done using model parameters as features. A drawback of such global LDS models, however, is poor invariance to viewpoint and illumination conditions [81]. To circumvent the heavy dependence of the first LDS models on global spatial appearance, the *bags of dynamical systems* (BoS) approach was developed by Ravichandran et al. [60]. Here, the global model is replaced by a set of codebook LDS models, each describing

a small space–time cuboid, used as local descriptors in a bag-of-words framework. For additional LDS-based approaches, see also [4,51,57,63,75].

The use of *histograms of local image measurements* as visual descriptors was pioneered by Swain and Ballard [68], who proposed using 3D colour histograms for object recognition. Different types of histogram descriptors have subsequently proven highly useful for a large range of visual tasks [9,10,22,29,34,36–38,48,64,73,85].

A large group of dynamic texture recognition methods based on statistics of local image primitives are *local binary pattern* (LBP)-based approaches. The original spatial LBP descriptor captures the joint binarised distribution of pixel values in local neighbourhoods. The spatio-temporal generalisations of LBP used for dynamic texture recognition do the same either for 3D space–time volumes (VLBP) [86,87] or by applying a two-dimensional LBP descriptor but on *three orthogonal planes* (LBP-TOP) [87]. Several extensions/versions of LBP-TOP have subsequently been presented, e.g. utilising averaging and principal histogram analysis to get more reliable statistics (Enhanced LBP) [61] or multi-clustering of salient features to identify and remove outlier frames (AFS-TOP) [24]. See also the completed local binary pattern approach (CVLBP) [71] and multi-resolution edge-weighted local structure patterns (MEWLSP) [72].

Contrasting LBP-TOP with VLBP highlights a conceptual difference concerning the generalisation of spatial descriptors to the spatio-temporal domain: Whereas the VLBP descriptor is based on full space–time 2D+T features, the LBP-TOP descriptor applies 2D features originally designed for purely spatial tasks on several cross sections of a spatio-temporal volume. While the former is in some sense conceptually more appealing and implies more discriminative modelling of the space–time structure, this comes with the drawback of higher descriptor dimensionality and higher computational load.

For LBP-based methods, using 2D descriptors on three orthogonal planes has so far been proven more successful, and this approach is frequently used also by non-LBP-based methods [1–3,83]. A variant of this idea proposed by Norouzzhad et al. [53] is to replace the three orthogonal planes with *nine spatio-temporal symmetry planes* and apply a histogram of oriented gradients on these planes. Similarly, the directional number transitional graph (DNG) by Rivera et al. [62] is evaluated on nine spatio-temporal cross sections of a video.

A different group of approaches similarly based on gathering statistics of local space–time structure, but using different primitives, is *spatio-temporal filtering*-based approaches. Examples are the oriented energy representations by Wildes and Bergen [79] and Derpanis and Wildes [12], where the latter represents *pure dynamics* of spatio-temporal textures by capturing space–time orientation using 3D Gaussian deriva-

tive filters. Marginal histograms of spatio-temporal Gabor filters were proposed by Gonçalves et al. [20]. Our proposed approach, which is based on *joint histograms* of time-causal spatio-temporal receptive field responses, also fits into this category. However, in contrast to spatio-temporal filtering-based approaches using marginal histograms, the use of joint histograms, which can capture the co-variation of different features, enables distinguishing a larger number of local space–time patterns. Joint statistics of ideal scale-space filters have previously been used for spatial object recognition, see, e.g. [64] and the approach by Linde and Lindeberg [37,38], which we here generalise to the spatio-temporal domain. The use of *time-causal* filters in our approach also contrasts with other spatio-temporal filtering-based approaches, and it should also be noted that the time-causal limit kernel used in this paper is *time-recursive*, whereas no time-recursive formulation is known for the scale-time kernel previously proposed by Koenderink [31].

There also exist a number of related methods similarly based on joint statistics of local space–time structure but using *filters learned from data* as primitives. *Unsupervised learning*-based approaches are, e.g. multi-scale binarised statistical image features (MBSIF-TOP) by Arashloo and Kittler [3], which learn filters by means of independent component analysis and PCANet-TOP by Arashloo et al. [2], where two layers of hierarchical features are learnt by means of principal component analysis (PCA). Approaches instead based on *sparse coding* for learning filters are orthogonal tensor dictionary learning (OTD) [59], equiangular kernel dictionary learning (SKDL) [58] and manifold regularised slow feature analysis (MR-SFA) [50].

Recently, several *deep learning*-based approaches for dynamic texture recognition have been proposed. These are typically trained by *supervised learning* and learn complex and abstract high-level features from data. Andrearczyk and Whelan [1] train a dynamic texture convolutional neural network from scratch to extract features on three orthogonal planes (DT-CNN). Qi et al. [56] instead use a network pretrained on an object recognition task for feature extraction (TCof), whereas Hong et al. [23] use features from a pretrained network as the basis for a deep dual descriptor (D3). Additional approaches include the high-level feature approach by Wang and Hu [76], which is a hybrid method using deep learning in combination with chaotic features, and the early deep learning approach of Culibrk and Sebe [8] based on temporal dropout of changes. It should be noted that none of these approaches are based on full *spatio-temporal deep features*. They instead use features extracted from individual frames, differences between pairs of frames or on orthogonal space–time planes. The best deep learning approaches are among the best-performing dynamic texture recognition methods, but this comes at the price of a lack of understanding and interpretability. The properties of the

learned nonlinear mappings and why these prove successful are not fully understood, neither in general nor for a network trained on a specific task. These highly complex “black-box” methods may also suffer more from surprising and unintuitive failure modes. Compared to methods incorporating more prior information about the problem structure, deep learning also requires larger amounts of training data from the specific domain, or one similar enough for successful transfer learning.

Spatio-temporal transforms have also been applied for dynamic texture recognition. Ji et al. [28] propose a method based on wavelet domain fractal analysis (WMFS), whereas Dubois et al. [13] utilise the 2D+T curvelet transform and Smith et al. [65] use spatio-temporal wavelets for video texture indexing. *Fractal dimension*-based methods make use of the self-similarity properties of dynamic textures. Xu et al. [83] create a descriptor from the fractal dimension of motion features (DFS), whereas Xu et al. [82] instead extract the power-law behaviour of local multi-scale gradient orientation histograms (3D-OTF) (see also [18,65]).

Additional approaches include using average degrees of complex networks [21] and a total variation-based approach by El Moubtahij et al. [14]. Wang and Hu [77] instead create a descriptor from chaotic features. There are also approaches combining several different descriptor types or features, such as DL-PEGASOS by Ghanem and Ahuja [19], which combines LBP, PHOG and LDS descriptors with maximum margin distance learning, or Yang et al. [84] who use ensemble SVMs to combine LBP, shape-invariant co-occurrence patterns (SCOPs) and chromatic information with dynamic information represented by LDS models.

3 Spatio-Temporal Receptive Field Model

We here give a brief description of the time-causal spatio-temporal receptive field framework of Lindeberg [44]. This framework provides the time-causal primitives for describing local space–time structure—the *spatio-temporal receptive fields* (or equivalently 2D+T scale-space filters)—which our proposed family of video descriptors is based on.

In this framework, the axiomatically derived scale-space kernel at spatial scale s and temporal scale τ is of the form

$$T(x, y, t; s, \tau, u, v, \Sigma) = g(x - ut, y - vt; s, \Sigma) h(t; \tau) \quad (1)$$

where

- (x, y) denotes the image coordinates,
- t denotes time,
- $h(t; \tau)$ denotes a temporal smoothing kernel,

– $g(x - ut, y - vt; s, \Sigma)$ denotes a spatial affine Gaussian kernel with spatial covariance matrix Σ that moves with image velocity (u, v) .

Here, we restrict ourselves to rotationally symmetric Gaussian kernels over the spatial domain corresponding to $\Sigma = I$ and to smoothing kernels with image velocity $(u, v) = (0, 0)$ leading to space–time separable receptive fields. This set of receptive fields is closed under spatial and temporal scaling transformations, but not under the group of Galilean transformations.

A more general approach, and also more similar to biological vision, would be to include a family of velocity-adapted receptive fields over a range of image velocities v . Such a set of receptive fields would constitute a Galilean covariant representation enabling fully Galilean invariant recognition according to the general theory presented in [41, Section 4.1.3–4.1.4] and applied to motion recognition in [35]. Exploring these possibilities is, however, left for future work. In this study, we choose to evaluate how far we can get using space–time separable receptive fields only.

Notably, also the family of space–time separable receptive fields can represent image velocities, since a set of spatial and temporal derivatives of different orders with respect to space and time can implicitly encode optic flow. An interesting biological parallel is that the superior colliculus is able to perform basic visual tasks, although receiving its primary inputs from the lateral geniculate nucleus (LGN), where a majority of the receptive fields are space–time separable.

The temporal smoothing kernel $h(t; \tau)$ used here is the time-causal kernel composed of truncated exponential functions coupled in cascade:

$$h_{\text{exp}}(t; \mu_k) = \begin{cases} \frac{1}{\mu_k} e^{-t/\mu_k} & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (2)$$

$$h(t; \tau) = *_{k=1}^K h_{\text{exp}}(-t; \mu_k) \quad (3)$$

where the symbol $*_{k=1}^K$ represents convolution of a set of K kernels, in our case K truncated exponentials with individual time constants μ_k . The temporal variance of the composed time-causal kernel (3) will depend on the time constants μ_k as $\tau_K = \sum_{k=1}^K \mu_k^2$. Specifically, we choose the time constants μ_k is such a way that the temporal variance τ_k of the intermediate kernels obeys a logarithmic distribution, $\tau_k = c^{k-K} \tau_K$ for some $c > 1$. Here we choose $c = 2$ which gives a reasonable trade-off between the sampling density over temporal scales and the temporal delays of the time-causal smoothing kernels [44].

In the limit of using an infinite number of temporal scale levels, the composed time-causal kernel (3) together with a logarithmic distribution of the intermediate scale levels leads

to a *scale-invariant limit kernel*, possessing true scale covariance and having a Fourier transform of the form [44, Eq. (38)]

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

For practical implementation, a finite number of intermediate scale levels need to be used. However, the composed kernel (3) converges rapidly to the limit kernel (4) with increasing K [44, Table 5]. The temporal smoothing kernels used in this work with temporal scale $\sigma_\tau = \sqrt{\tau}$ are such composed convolution kernels, where we use $K \geq 7$, which leads to very small deviations from the limit kernel.

The *spatio-temporal receptive fields* are in turn defined as partial derivatives of the spatio-temporal scale-space kernel T for different orders (m_1, m_2, n) of spatial and temporal differentiation computed at multiple spatio-temporal scales. The result of convolving a video $f(x, y, t)$ with one of these kernels

$$L_{x^{m_1} y^{m_2} t^n}(\cdot, \cdot, \cdot; s, \tau) = \partial_{x^{m_1} y^{m_2} t^n} (T(\cdot, \cdot, \cdot; s, \tau) * f(\cdot, \cdot, \cdot)) \quad (5)$$

is referred to as the *receptive field response*. A set of receptive field responses will comprise a *spatio-temporal N -jet* representation of the local space–time structure, in essence corresponding to a truncated Taylor expansion possibly at multiple spatial and temporal scales. Using this representation thus enables capturing more diverse information about the local space–time structure than what can be done using a single filter type.

One set of receptive fields considered in this work consists of combinations of the following sets of partial derivatives: (i) the first- and second-order spatial derivatives, (ii) the first- and second-order temporal derivatives of these and (iii) the first- and second-order temporal derivatives of the original smoothed video L :

$$\{ \{L_x, L_y, L_{xx}, L_{xy}, L_{yy}\}, \{L_{xt}, L_{yt}, L_{xxt}, L_{xyt}, L_{yyt}\}, \{L_{xtt}, L_{ytt}, L_{xxtt}, L_{xytt}, L_{yytt}\}, \{L_t, L_{tt}\} \}. \quad (6)$$

A second set consists of spatio-temporal invariants defined from combinations of these partial derivatives. A subset of the receptive fields/scale-space derivative kernels are shown in Fig. 1. (The kernels are three-dimensional over (x, y, t) , but are here illustrated using a single spatial dimension x .)

All video descriptors are expressed in terms of scale-normalised derivatives [44]:

$$\partial_\xi = s^{\gamma_s/2} \partial_x, \quad \partial_\eta = s^{\gamma_s/2} \partial_y, \quad \partial_\zeta = \tau^{\gamma_\tau/2} \partial_t \quad (7)$$

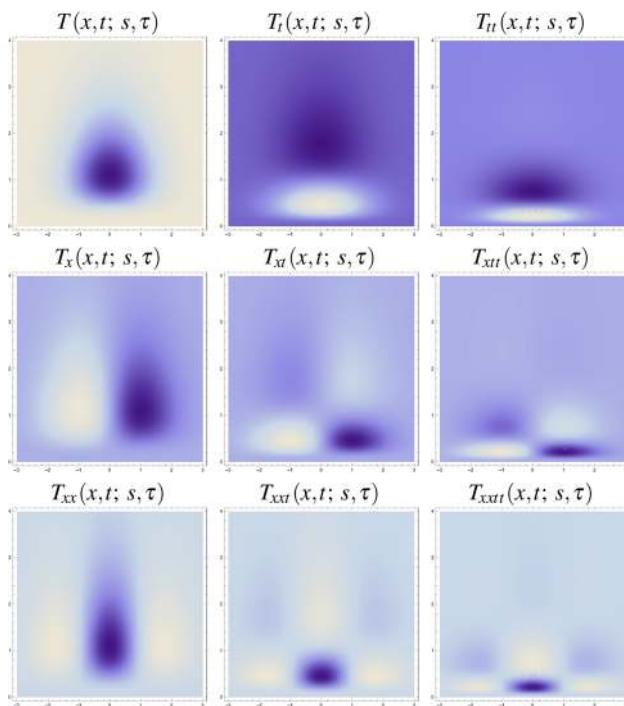


Fig. 1 Space–time separable kernels $T_{x^m t^n}(x, t; s, \tau) = \partial_{x^m t^n}(g(x; s)h(t; \tau))$ for all combinations of spatial and temporal derivatives up to order two as obtained from the composition of Gaussian kernels over the spatial domain x and the time-causal limit kernel over the temporal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

with the corresponding scale-normalised receptive fields computed as:

$$L_{\xi^{m_1} \eta^{m_2} \zeta^n} = s^{(m_1+m_2)\gamma_s/2} \tau^{n\gamma_\tau/2} L_{x^{m_1} y^{m_2} t^n} \quad (8)$$

where $\gamma_s > 0$ and $\gamma_\tau > 0$ are the scale normalisation powers of the spatial and the temporal domains, respectively. We here use $\gamma_s = 1$ and $\gamma_\tau = 1$ corresponding to the maximally scale-invariant case (as described in “Appendix A.1”).

It should be noted that the scale-space representation and receptive field responses are computed “on the fly” using only a compact temporal buffer. The time-recursive formulation for the temporal smoothing [44, Eq (56)] means that computing the scale-space representation for a new frame at temporal scale τ_k only requires information from the present moment t and the scale-space representation for the preceding frame $t - 1$ at temporal scales τ_k and τ_{k-1}

$$L(\cdot, \cdot, t; \cdot, \tau_k) = L(\cdot, \cdot, t - 1; \cdot, \tau_k) + \frac{1}{\mu_k} (L(\cdot, \cdot, t; \cdot, \tau_{k-1}) - L(\cdot, \cdot, t - 1; \cdot, \tau_{k-1})) \quad (9)$$

where t represents time and $\tau_k > \tau_{k-1}$ are two adjacent temporal scale levels, where $k = 0$ corresponds to the original signal. (The spatial coordinates and the spatial scale are here

suppressed for brevity.) This is a clear advantage compared to using a Gaussian kernel (possibly delayed and truncated) over the temporal domain, since it implies less computations and smaller temporal buffers.

These spatio-temporal receptive field responses can be computed efficiently by space–time separable filtering. The spatial smoothing is done in by separable discrete Gaussian filtering over the spatial dimensions, with the spatial extent proportional to the spatial scale parameter in units of the standard deviation $\sigma_s = \sqrt{s}$. The temporal smoothing is performed by recursive filtering requiring only two additions and one multiplication per pixel and spatio-temporal scale level. Then, scale-space derivatives are computed by applying small support discrete derivative approximations to the spatio-temporal scale-space representation.

For further details concerning the spatio-temporal scale-space representation and its discrete implementation, we refer to Lindeberg [44].

4 Video Descriptors

We here describe our proposed family of video descriptors. The histogram descriptor is based on regional statistics of time-causal spatio-temporal receptive field responses, and the process of computing the video descriptor can be divided into four main steps:

- (i) Computation of the spatio-temporal scale-space representation at the specified spatial and temporal scales (spatial and temporal smoothing).
- (ii) Computation of local spatio-temporal receptive field responses using discrete derivative approximation filters over space and time.
- (iii) Dimensionality reduction of the local spatio-temporal receptive field responses/feature vectors using principal component analysis (PCA).
- (iv) Aggregation of joint statistics of receptive field responses over a space–time region into a multi-dimensional histogram.

The resulting histogram video descriptor will describe a space–time region by the relative frequency of different local space–time patterns. Note that all computations are performed “on the fly”, one frame at a time. In the following, each of the above steps is described in more detail. A schematic illustration of our dynamic texture recognition workflow is given in Fig. 2.

4.1 Spatio-Temporal Scale-Space Representation

The first processing step is spatial and temporal smoothing to compute the time-causal spatio-temporal scale-space

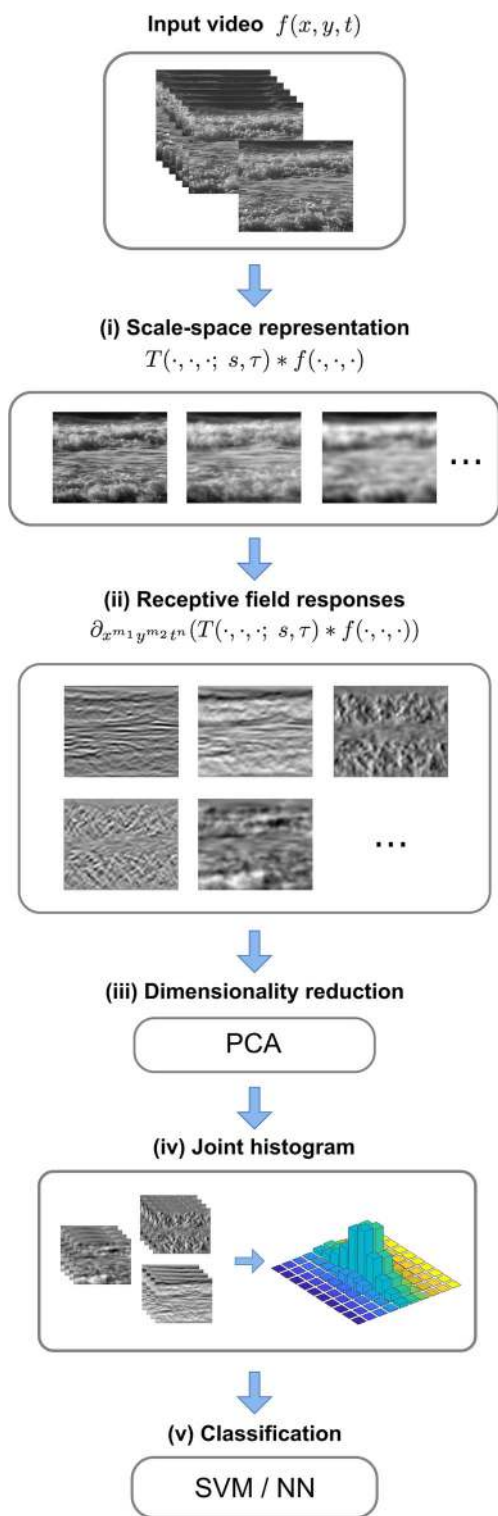


Fig. 2 Overview of the main steps in our dynamic texture recognition workflow: (i) Spatio-temporal scale-space representation of the current frame. (ii) Computation of local spatio-temporal receptive field responses. (iii) Dimensionality reduction of local receptive field responses using principal component analysis (PCA). (iv) Aggregation of joint statistics of receptive field responses over a space–time region into a joint multi-dimensional histogram (here illustrated in two dimensions). (v) Classification using an SVM or NN classifier

representation of the current frame at the chosen spatial and temporal scales. The spatial smoothing is done in a cascade over spatial scales, and the temporal smoothing is performed recursively over time and in a cascade over temporal scales according to [44]. The fully separable convolution kernels and the time-recursive implementation of the temporal smoothing imply that the smoothing can be done in a computationally efficient way.

4.2 Spatio-Temporal Receptive Field Responses

After the smoothing step, the spatio-temporal receptive field responses $F = [F_1, F_2, \dots, F_N]$ are computed *densely* over the current frame for a set of N chosen scale-space derivative filters. These filters will include partial derivatives over a set of different spatial and temporal differentiation orders at single or multiple spatial and temporal scales. Alternatively, for a video descriptor based on differential invariants, the features will correspond to such differential invariants computed from the partial derivatives. Using filters of multiple scales enables capturing image structures of different spatial extent and temporal duration. All derivatives are scale-normalised.

Previous methods utilising “ideal” (as opposed to learned) spatio-temporal filters have used families of filters created from a single filter type, e.g. by applying third-order filters in different directions in space–time [12] or using a Gabor filter extended by both spatial rotations and velocity transformations [20].

We here take a different approach, using a richer family of receptive fields encompassing different orders of spatial and temporal differentiation, specifically including mixed spatio-temporal derivatives. Using such a set enables representing local intensity variations of *different orders*. This is not possible if restricting filters to a single base filter, even if the filters would be extended by spatio-temporal transformations. The spatio-temporal receptive field responses from such a family can thus be used to separate patterns based on a combination of, for example, first-order and second-order intensity variations. It will also enable computing measures that place equal weight on the first- and second-order changes.

Biological vision does also comprise such richer sets of spatio-temporal receptive fields, which typically occur in pairs of odd-shaped and even-shaped receptive fields [74] and which can be well modelled by spatio-temporal scale-space derivatives for different orders of spatial and temporal differentiation [44]. As previously discussed in Sect. 3, a more general model would be to combine these spatio-temporal receptive fields with velocity adaptation over a set of Galilean transformations. This extension is, however, left for future work.

A more detailed discussion concerning the choice of the specific receptive field sets that our video descriptors are based on is given in Sect. 4.6.1.

4.3 Dimensionality Reduction with PCA

When combining a large number of local image measurements, most of the variability in a dataset will be contained in a lower-dimensional subspace of the original feature space. This opens up for Dimensionality reduction of the local feature vectors to enable summarising a larger set of receptive field responses without resulting in a prohibitively large joint histogram descriptor.

For this purpose, a dimensionality reduction step in the form of principal component analysis is added before constructing the joint histogram. The result will be a local feature vector $\tilde{F}(x, y, t) = [\tilde{F}_1, \tilde{F}_2, \dots, \tilde{F}_M] \in \mathbb{R}^M$ $M \leq N$, where the new components correspond to linear combinations of the filter responses from the original feature vector F .

The transformation properties of the spatio-temporal scale-space derivatives imply scale covariance for such linear combinations of receptive fields, as well as for the individual receptive field responses *if using the proper scale normalisation* with $\gamma_s = 1$ and $\gamma_\tau = 1$. A more detailed treatment of the scale covariance properties is given in “Appendix A.1”. The main reason for choosing PCA is simplicity and that it has empirically been shown to give good results for spatial images [38]. The number of principal components M can be adapted to requirements for descriptor size and need for detail in modelling the local image structure. The dimensionality reduction step can also be skipped if working with a smaller number of receptive fields.

4.4 Joint Receptive Field Histograms

The approach that we will follow to represent videos of dynamic textures is to use histograms of spatio-temporal receptive field responses. Notably, such a histogram representation will discard information about the spatial positions and the temporal moments that the feature responses originate from. Because the histograms are computed from spatio-temporal receptive field responses in terms of spatio-temporal derivatives, these feature responses will, however, implicitly code for partial spatial and temporal information, like pieces in a spatio-temporal jigsaw puzzle. By computing these receptive field responses over multiple spatial and temporal scales, we additionally capture primitives of different spatial size and temporal duration. For spatial recognition, related histogram approaches have turned out to be highly successful, leading to approaches such as SIFT [48], HOG [9] and HOF [10] and bag-of-words models. For the task of texture recognition, the loss of information caused by discarding spatial positions is also less critical, since many textures can be expected to possess certain stationarity properties. In this work, we use an extension of this paradigm of histogram-based image descriptors for building a conceptually simple system for video analysis.

The multi-dimensional histogram of receptive field responses will represent a discretised version of the joint distribution of local receptive field responses over a space–time region. To construct the histogram, each feature dimension is partitioned into n_{bins} number of equidistant bins in the range

$$[\text{mean}(\tilde{F}_i) - d \text{std}(\tilde{F}_i), \text{mean}(\tilde{F}_i) + d \text{std}(\tilde{F}_i)]$$

where the mean and the standard deviation are computed over the training set and d is a parameter controlling the number of standard deviations that are spanned for each dimension. We here use $d = 5$. If including M principal components, the result is a multi-dimensional histogram with $n_{\text{cells}} = (n_{\text{bins}})^M$ distinct histogram cells. Note the use of the parameter n_{bins} (histogram bins) to refer to the number of partitions along each feature dimension and the parameter n_{cells} (histogram cells) to refer to the number of cells in the joint histogram.

The receptive field responses up to a certain order will represent information corresponding to the coefficients of a Taylor expansion around each point in space time. Each histogram cell will correspond to a certain “template” local space–time structure, encoding joint conditions on the magnitudes of the spatio-temporal receptive field responses (here, image derivatives, differential invariants or PCA components). This is somewhat similar to, for example, the templates used in VLBP [87] but notably represented and computed using different primitives. The normalised histogram video descriptor then captures *the relative frequency* of such space–time structures in a space–time region. The number of different local “templates” will be decided by the number of receptive fields/principal components and the number of bins.

A *joint* histogram video descriptor explicitly models the *co-variation* of different types of image measurements, in contrast to the more common choice of descriptors based on marginal distributions or relative feature strength (see, e.g. [12,20]). A simple example of this is that a joint histogram over L_x and L_y will reflect the orientations of gradients over image space, which would not be sufficiently captured by the corresponding marginal histograms. Using joint histograms similarly implies the ability to represent other types of patterns that correspond to certain *relationships* between groups of features, such as how receptive field responses co-vary over different spatial and temporal scales.

In the general case, these histograms should be computed regionally, over different regions over space and/or time, e.g. to separate regions that contain different types of dynamic textures. For almost all experiments in this study, however, the *space–time region* for the histogram descriptor will be chosen as the entire video, leading to a *single global histogram descriptor* per dynamic texture video. The single exception is the experiment presented in Sect. 7.6, where

we compute histograms over a smaller number of video frames. The reason for primarily using global histograms is that the videos in the DynTex and UCLA benchmarks are presegmented to contain a single dynamic texture class per video. Thereby, we can make this conceptual simplification for the experimental evaluation of our different types of video descriptors.

It should be noted that, if represented naively, a subset of the histograms descriptors evaluated here would be prohibitively large. However, for such high-dimensional descriptors the number of *nonzero* histogram cells can be considerably lower than the maximum number of cells. This implies they can be efficiently represented using a computationally efficient sparse representation as outlined by Linde and Lindeberg [38].

4.5 Covariance and Invariance Properties of the Video Descriptor

The scale covariance properties of the spatio-temporal receptive fields and the PCA components, according to the theory presented in “Appendix A.1”, imply that a histogram descriptor constructed from these primitives will be *scale-covariant* over all nonzero spatial scaling factors and for temporal scaling factors that are integer powers of the distribution parameter c of the time-causal limit kernel. This means that our proposed video descriptors can be used as the building blocks of a *scale-invariant* recognition framework.

A straightforward option for this is to use the video descriptors in a multi-scale recognition framework, where each video is represented by a *set of descriptors* computed at multiple spatial and temporal scales, both during training and testing. A scale-covariant descriptor then implies that if training is performed for a video sequence at scale (s_1, τ_1) and if a corresponding video sequence is rescaled by spatial and temporal scaling factors S_s and S_τ , corresponding recognition can be performed at scale $(s_2, \tau_2) = (S_s^2 s_1, S_\tau^2 \tau_1)$. However, for the initial experiments performed in this work this option for scale-invariant recognition has not been explored. Instead, training and classification are performed at the same scale or the same set of scales, and the outlined scale-invariant extension is left for future work.

4.6 Choice of Receptive Fields and Descriptor Parameters

The basic video descriptor described above will give rise to a family of video descriptors when varying the set of receptive fields and the descriptor parameters. Here, we describe the different options investigated in this work considering: (i) the set of receptive fields, (ii) the number of bins and the number of principal components used for constructing the histogram and (iii) the spatial and temporal scales of the receptive fields.

Table 1 The video descriptors investigated in this paper and the receptive field sets they are based on

Name	Receptive field set
RF Spatial	$\{L_x, L_y, L_{xx}, L_{xy}, L_{yy}\}$
STRF N -jet	$\{L_x, L_y, L_{xx}, L_{xy}, L_{yy}\}, \{L_t, L_{tt}\},$ $\{L_{xt}, L_{yt}, L_{xxt}, L_{xyt}, L_{yyt}\},$ $\{L_{xtt}, L_{ytt}, L_{xxtt}, L_{xytt}, L_{yytt}\}$
STRF RotInv	$\{ \nabla L , \nabla L_t , \nabla L_{tt} \},$ $\{\nabla^2 L, \nabla^2 L_t, \nabla^2 L_{tt}\},$ $\{\det \mathcal{H} L, \det \mathcal{H} L_t, \det \mathcal{H} L_{tt}\}$
STRF N -jet (previous [27])	$\{L_x, L_y, L_{xx}, L_{xy}, L_{yy}\}, \{L_t, L_{tt}\},$ $\{L_{xt}, L_{yt}, L_{xxt}, L_{xyt}, L_{yyt}\}$

4.6.1 Receptive Field Sets

The set of receptive fields used as primitives for constructing the histogram will determine the type of information that is represented in the video descriptor. A straightforward example of this is that using rotationally invariant differential invariants will imply a rotationally invariant video descriptor. A second example is that including or excluding purely temporal derivatives will enable or disable capturing temporal intensity changes not mediated by spatial motion. We have chosen to compare video descriptors based on four different receptive field groups as summarised in Table 1.

First, note that all video descriptors, except STRF N -jet (previous), include first- and second-order spatial and temporal derivatives *in pairs* $\{L_x, L_{xx}\}, \{L_t, L_{tt}\}, \{L_{xt}, L_{xxt}\}, \{L_{xyt}, L_{xxtt}\}$, etc. The motivation for this is that first- and second-order derivatives provide complementary information, and by including both, equal weight is put on first- and second-order information. It has specifically been observed that biological receptive fields occur in pairs of odd-shaped and even-shaped receptive field profiles that can be well approximated by Gaussian derivatives [32,42,74]. In the following, we describe the four video descriptors in more detail and do further motivate the choice of their respective receptive field sets.

RF Spatial is a purely spatial descriptor based on the full spatial N -jet up to order two. This descriptor will capture the spatial patterns in “snapshots” of the scene (single frames) independent of the presence of movement. Using spatial derivatives up to order two means that each histogram cell template will represent a discretised second-order approximation of the local spatial image structure. An additional motivation for using this receptive field set is that this descriptor is one of the best-performing spatial descriptors for the receptive field-based object recognition method in [38]. This descriptor is primarily included as a baseline to compare the spatio-temporal descriptors against.

STRF N-jet is a directionally selective spatio-temporal descriptor, where the first- and second-order spatial derivatives are complemented with the first- and second-order temporal derivatives of these as well as the first- and second-order temporal derivatives of the smoothed video L . Including purely temporal derivatives means that the descriptor can capture intensity changes not mediated by spatial motion (flicker). The set of mixed spatio-temporal derivatives will on the other hand capture the interplay between changes over the spatial and temporal domains, such as movements of salient spatial patterns. An additional motivation for including mixed spatio-temporal derivatives is that they represent features that are well localised with respect to joint spatio-temporal scales. This implies that when using multiple scales, a descriptor including mixed spatio-temporal derivatives will have better ability to separate spatio-temporal patterns at different spatio-temporal scales.

STRF RotInv is a rotationally invariant video descriptor based on a set of rotationally invariant features over the spatial domain: the spatial gradient magnitude $|\nabla L| = \sqrt{L_x^2 + L_y^2}$, the spatial Laplacian $\nabla^2 L = L_{xx} + L_{yy}$ and the determinant of the spatial Hessian $\det \mathcal{H}L = L_{xx}L_{yy} - L_{xy}^2$.¹ These are evaluated on the smoothed video L directly and on the first- and second-order temporal derivatives of the scale-space representation L_t and L_{tt} . One motivation for choosing these spatial differential invariants is that they are functionally independent and span the space of rotationally invariant first- and second-order differential invariants over the spatial domain. This set of rotationally invariant features was also demonstrated to be the basis of one of the best-performing spatial descriptors in [38]. By applying these differential operators to the first- and second-order temporal derivatives of the video, the interplay between temporal and spatial intensity variations is captured.

STRF N-jet (previous) is our previously published [27] video descriptor. This descriptor is included for reference and differs from *STRF N-jet* by lacking the second-order temporal derivatives of the first- and second-order spatial derivatives. This descriptor was evaluated in [27] without parameter tuning, and we also here keep the original parameters.

It can be noted that none of these video descriptors makes use of the full spatio-temporal 4-jet. This reflects the philosophy of treating space and time as distinct dimensions, where the most relevant information lies in the interplay between spatial changes (here, of first- and second-order) with temporal changes (here, of first- and second-order). Third- and fourth-order information with respect to either the spatial

or the temporal domain is thus discarded. Receptive field responses for two videos of dynamic textures are shown for spatio-temporal partial derivatives in Fig. 3 and for rotational differential invariants in Fig. 4.

It should be noted that the recognition framework presented here also allows for using non-separable receptive fields with nonzero image velocities. Exploring this is, however, left for future work, and in this study we instead focus on evaluating different sets of space–time separable receptive fields (see also the discussion in Sect. 3).

4.6.2 Number of Bins and Principal Components

Different choices of the number of bins per feature dimension n_{bins} and the number of principal components n_{comp} will give rise to qualitatively different histogram descriptors. Using few principal components in combination with many bins will enable fine-grained recognition of a smaller number of similar pattern templates (separating patterns based on smaller magnitude differences in receptive field responses). On the other hand, using a larger number of principal components in combination with fewer bins will imply a histogram capturing a larger number of more varied but less “precise” patterns. The different options that we have considered in this work are:

$$n_{\text{comp}} \in [2, 17]$$

$$n_{\text{bins}} \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 25\}$$

After a set of initial experiments, where we varied the number of bins and the number of principal components (presented in Sect. 7.1), we noted that binary histograms with 10–17 principal components achieve highly competitive results for all benchmarks. Binary histogram descriptors also have an appeal in simplicity and one less parameter to tune. Therefore, the subsequent experiments (Sect. 7.2 and forward), were performed using binary histograms only.

4.6.3 Binary Histograms

When choosing $n_{\text{bins}} = 2$ equivalent to a joint binary histogram, the local image structure is described by only the sign of the different image measurements. This will make the descriptor invariant to uniform rescalings of the intensity values, such as multiplicative illumination transformations or indeed any change that does not affect the sign of the receptive field response. Binary histograms in addition enable combining a larger number of image measurements without a prohibitive large descriptor dimensionality and have proven an effective approach by a large number of LBP-inspired methods.

¹ To transform the determinant of the spatial Hessian having the same dimensionality in terms of [intensity] as the other spatial differential invariants, we transform the magnitude by a square root function while preserving its sign: $(\det \mathcal{H}L)_{\text{transf}} = \text{sign}(\det \mathcal{H}L) |\det \mathcal{H}L|^{1/2}$.

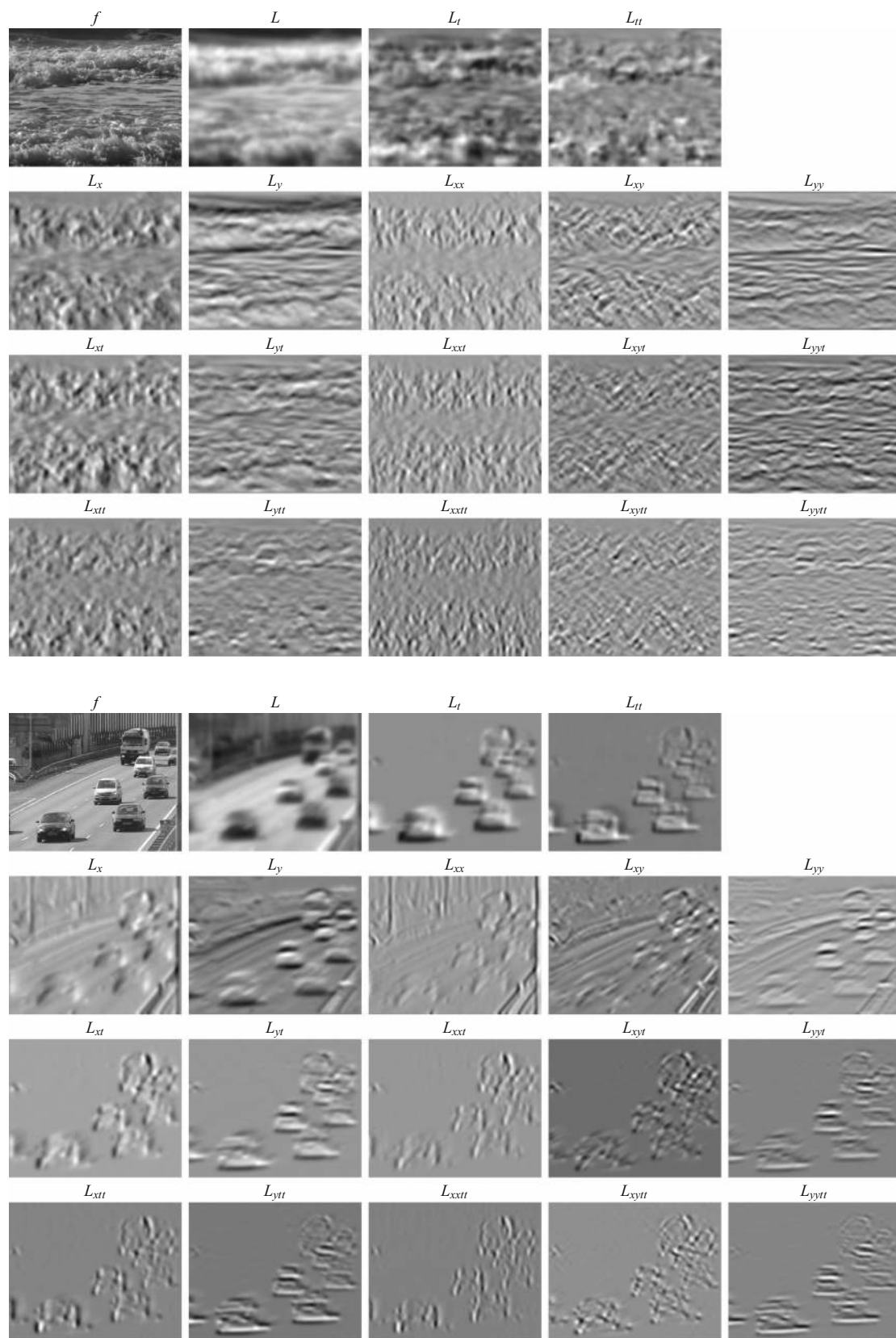


Fig. 3 Spatio-temporal receptive field responses for the set of spatio-temporal partial derivatives that the STRF N -jet descriptor is based on. Snapshots of receptive field responses are shown for two dynamic textures from the DynTex classes “waves” (top) and “traffic” (bottom) ($\sigma_s = 8$, $\sigma_\tau = 100$)

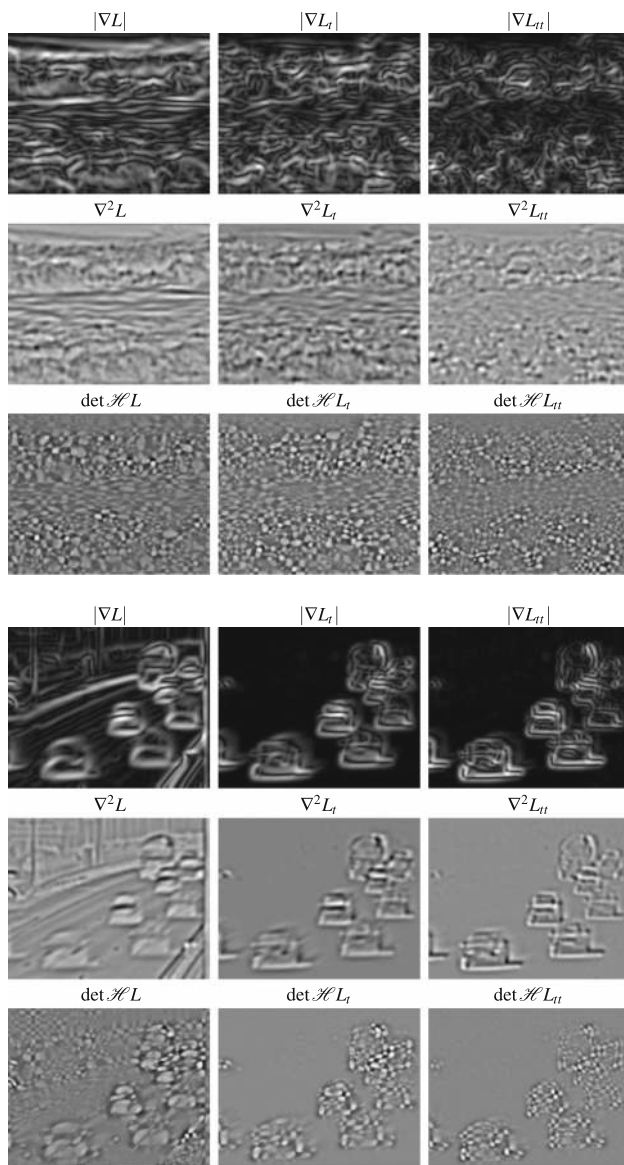


Fig. 4 Rotationally invariant spatio-temporal receptive field responses for the rotationally invariant differential invariants underlying the STRF RotInv video descriptor. Snapshots of receptive field responses are shown for two dynamic textures from the DynTex classes “waves” (top) and “traffic” (bottom) ($\sigma_s = 8$, $\sigma_\tau = 100$)

4.6.4 Spatial and Temporal Scales

Using receptive fields at multiple spatial and temporal scales in the descriptor makes it possible to capture image structures of different spatial extent and temporal duration. Such a multi-scale descriptor will also comprise more complex image primitives, since the corresponding local space–time templates will represent combinations of patterns at different scales. The spatial and temporal scales, i.e. the standard deviation for the respective scale-space kernels, considered in this work are:

$$\sigma_s \in \{1, 2, 4, 8, 16\} \text{ pixels}$$

$$\sigma_\tau \in \{50, 100, 200, 400\} \text{ ms}$$

The lowest temporal scale 50 ms has been chosen to be of the same order as the time difference between adjacent frames for regular video frame rates around 25 fps. This temporal scale is also of the same order as the temporal scales of spatio-temporal receptive fields observed in the lateral geniculate nucleus (LGN) and the primary visual cortex (V1), where examples of receptive fields have been well modelled using time constants over the range of 40–80 ms [44, Figs. 3–4].

In this study, we use receptive field responses at either a single spatial and temporal scale, or for a combination of pairs of adjacent spatial and temporal scales (2×2 scales):

$$(\sigma_{s_1}, \sigma_{s_2}) \in \{(1, 2), (2, 4), (4, 8), (8, 16)\} \text{ pixels}$$

$$(\sigma_{\tau_1}, \sigma_{\tau_2}) \in \{(50, 100), (100, 200), (200, 400)\} \text{ ms}$$

The reason why we do not include combinations of more than two temporal scales and two spatial scales is that an initial prestudy did not show any significant improvements from this on the benchmark datasets. The choice to use combinations of *adjacent scales* is done mainly to limit the complexity of the study. Thus, here, 20 combinations of a single spatial scale with a single temporal scale and 12 combinations of two spatial scales and two temporal scales are considered.

A more general approach than using fixed pairs of adjacent scales is to operate on a wider range of spatio-temporal scales in parallel. For example, for breaking water waves that roll onto a beach, the coarser scale receptive fields will respond to the gross motion pattern of the water waves, whereas the finer-scale receptive fields will respond to the detailed fine-scale motion pattern of the water surface. A general-purpose vision system should have the ability to dynamically operate over such different subsets of spatial and temporal scales, to extract maximum amount of relevant information about a dynamic scene. Specifically, there is interesting potential in determining local spatial and temporal scale levels adaptively from the video data, using recently developed methods for spatio-temporal scale selection [45,47]. We leave such extensions for future work.

5 Datasets

We evaluate our proposed approach on six standard dynamic texture recognition/classification benchmarks from two widely used dynamic texture datasets: UCLA [66] and DynTex [55]. We here give a brief description of the datasets and the benchmarks. Sample frames from the datasets are shown in Fig. 5 (UCLA) and Fig. 6 (DynTex).



Fig. 5 Sample frames from the UCLA dataset. The top and bottom rows show different dynamic texture instances from the same conceptual class as used in the UCLA8 and UCLA9 benchmarks. For the UCLA50 benchmark, these instances should instead be separated as

different classes. The examples are from the conceptual classes (from left to right): “fire”, “boiling”, “flowers”, “plants”, “sea”, “water” and “smoke”



Fig. 6 Sample frames from the DynTex dataset. The top and bottom rows show different dynamic texture instances from the same conceptual class. The examples are from the conceptual classes (from left to right): “sea”, “calm water”, “naked trees”, “fountains”, “flags”, “traffic” and “escalator”

5.1 UCLA

The UCLA dataset was introduced by Soatto et al. [66] and is composed of 200 videos (160×110 pixels, 15 fps) featuring 50 different dynamic textures with 4 samples from each texture. The **UCLA50** benchmark [66] divides the 200 videos into 50 classes with one class per individual texture/scene. It should be noted that this partitioning is not conceptual in the sense of the classes constituting different types of textures such as “fountains”, “sea” or “flowers” but instead targets *instance-specific* and *viewpoint-specific* recognition. This means that not only different individual fountains but also the same fountain seen from two different viewpoints should be separated from each other.

Since for many applications it is more relevant to recognise different dynamic texture categories, a partitioning of the UCLA dataset into *conceptual classes*, **UCLA9**, was introduced by Ravichandran et al. [60] with the following classes: boiling water (8), fire (8), flowers (12), fountains (20), plants (108), sea (12), smoke (4), water (12) and waterfall (16), where the numbers correspond to the number of samples from each class. Because of the large overrepresentation of plant videos for this benchmark, in the **UCLA8**

benchmark, those are excluded to get a more balanced dataset, leaving 92 videos from eight conceptual classes.

5.2 DynTex

A larger and more diverse dynamic texture dataset, **DynTex**, was introduced by Péteri et al. [55], featuring a larger variation of dynamic texture types recorded under more diverse conditions (720×576 pixels, 25 fps). From this dataset, three gradually larger and more challenging benchmarks have been compiled by Dubois et al. [13]. The **Alpha** benchmark includes 60 dynamic texture videos from three different classes: sea, grass and trees. There are 20 examples of each class and some variations in scale and viewpoint. The **Beta** benchmark includes 162 dynamic texture videos from ten classes: sea, vegetation, trees, flags, calm water, fountain, smoke, escalator, traffic and rotation. There are 7–20 examples of each class. The **Gamma** benchmark includes 264 dynamic texture videos from ten classes: flowers, sea, trees without foliage, dense foliage, escalator, calm water, flags, grass, traffic and fountains. There are 7–38 examples of each class, and this benchmark features the largest intraclass variability in terms of scale, orientation, etc.

6 Experimental Set-up

This section describes the cross-validation schemes used for the different benchmarks, the classifiers and the use of parameter tuning over the descriptor parameters.

6.1 Benchmark Cross-Validation Schemes

The standard test set-up for the UCLA50 benchmark, which we adopt also here, is fourfold cross-validation [66]. For each partitioning, three out of four samples from each dynamic texture instance are used for training, while the remaining one is held out for testing.

The standard test set-up for the UCLA8 and UCLA9 benchmarks is to report the average accuracy over 20 random partitions, with 50% data used for training and 50% for testing (randomly bisecting each class) [19]. We use the same set-up here, except that we report results as an average over 1000 trials to get more reliable statistics. This since we noted that, because of the small size of the dataset, the specific random partitioning will otherwise affect the result. For all the UCLA benchmarks, in contrast to the most common set-up of using manually extracted patches, we use the *non-cropped videos*; thus, our set-up could be considered a slightly harder problem.

For the DynTex benchmarks, the experimental set-up used is leave-one-out cross-validation as in [3,24,56,84]. We perform no subsampling of videos but use the full 720×576 pixels frames.

6.2 Classifiers

We present results of both using a support vector machine (SVM) classifier and a nearest neighbour (NN) classifier, the latter to evaluate the performance also of a simpler classifier without hidden tunable parameters. For NN we use the χ^2 -distance $d(x, y) = \sum_i (x_i - y_i)^2 / (x_i + y_i)$ to compute the distance between two histogram video descriptors, and for SVM we use the χ^2 -kernel $e^{-\gamma d(x,y)}$. Results are quite robust to the choice of the SVM hyperparameters γ and C . We here use $\gamma = 0.1$ and $C = 10,000$ for all experiments.

6.3 Descriptor Parameter Tuning

Comparisons with state of the art and between video descriptors based on different sets of receptive fields are made using binary descriptors. Parameter tuning is performed as a grid search over the number of principal components $n_{\text{comp}} \in [2, 17]$, spatial scales $\sigma_s \in \{1, 2, 4, 8, 16\}$ and temporal scales $\sigma_\tau \in \{50, 100, 200, 400\}$. For spatial and temporal scales, we consider both single scales and combinations of two adjacent spatial and temporal scales. The standard evaluation protocols for the respective benchmarks

(i.e. slightly different cross-validation schemes) are used for parameter selection, and results are reported for each video descriptor using the optimal set of parameters.

This is the standard method for parameter selection used on these benchmarks. The reason for following it also here is to be able to make a direct comparison with the already published results in the literature using the same experimental protocol. It should be noted, however, that this evaluation protocol, which does not fully separate the test and the train data, does introduce the risk for model selection bias. A better evaluation scheme would be to use a *full nested cross-validation scheme*, although this would be quite computationally expensive. To give transparency to the effect of parameter tuning on the performance of our descriptors, we therefore also present results from parameter tuning of the different descriptor parameters in (Sects. 7.1–7.2) showing how much the performance is affected by choosing non-optimal parameter values.

7 Experiments

Our first experiments consider a qualitative and quantitative evaluation of different versions of our video descriptors, where we present results on: (i) varying the number of bins and principal components, (ii) using different spatial and temporal scales for the receptive fields and (iii) comparing descriptors based on different sets of receptive fields. This is followed by (iv) a comparison with state-of-the-art dynamic texture recognition methods and finally (v) a qualitative analysis on reasons for errors.

7.1 Number of Bins and Principal Components

The classification performance of the STRF N -jet descriptor as function of the number of bins and the number of principal components used in the histogram descriptor are presented in Fig. 7 for the UCLA8 and UCLA50 benchmarks and in Fig. 8 for the Beta and Gamma benchmarks. A first observation is that, not surprisingly, when using a smaller number of principal components, each dimension needs to be divided into a larger number of bins to achieve good performance, e.g. for $n_{\text{comp}} \in \{2, 3\}$ the best performance is achieved for $n_{\text{bins}} \geq 15$ for all benchmarks. To discriminate between a large number of spatio-temporal patterns using only a few image measurements, these need to be more precisely recorded. A qualitative difference between using an odd or an even number of bins for $n_{\text{bins}} \leq 8$ can also be noted. This can be explained by a qualitative difference in the handling of feature values close to zero.

At the other end of the spectrum, it can be seen that when using a large number of principal components, fewer bins suffice. Using a large number of spatio-temporal primitives

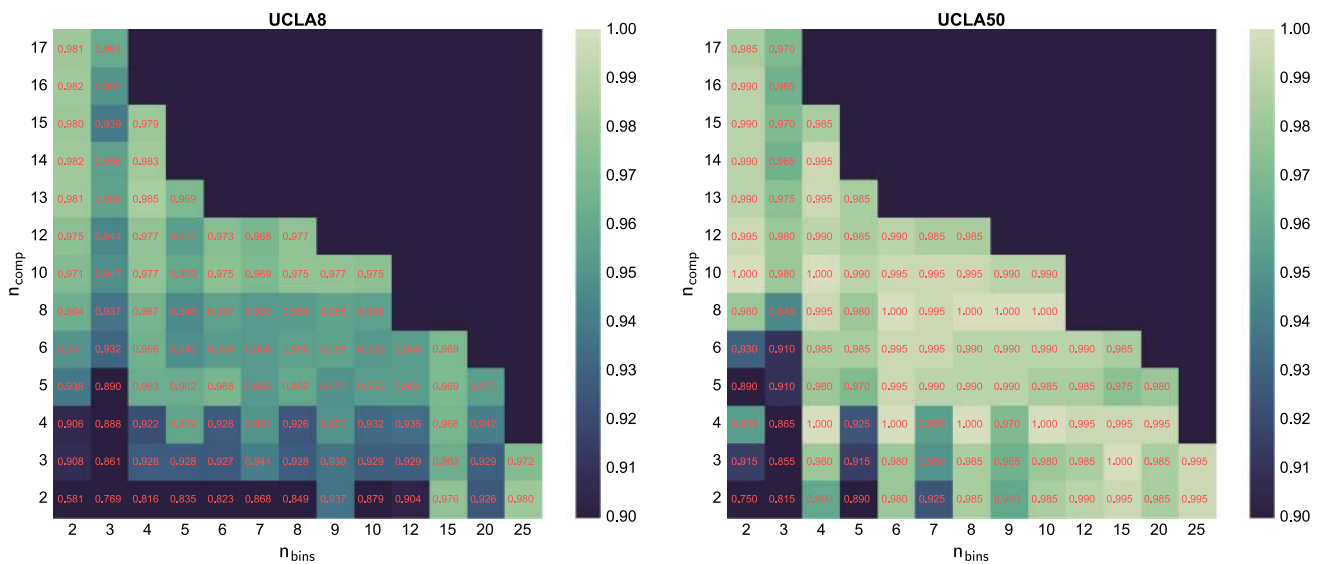


Fig. 7 The classification performance of the STRF N -jet descriptor as function of the number of bins and the number of principal components for the UCLA8 and UCLA50 benchmarks. For the UCLA8 benchmark, the best results are obtained when using more than eight principal components in combination with binary or ternary histograms.

For the UCLA50 benchmark, a large range of different configurations result in 0% error, indicating that the task of recognising dynamic texture instances is less challenging compared to separating dynamic texture categories. Spatial scales: $(\sigma_{s_1}, \sigma_{s_2}) = (1, 2)$ pixels. Temporal scales: $(\sigma_{t_1}, \sigma_{t_2}) = (50, 100)$ ms

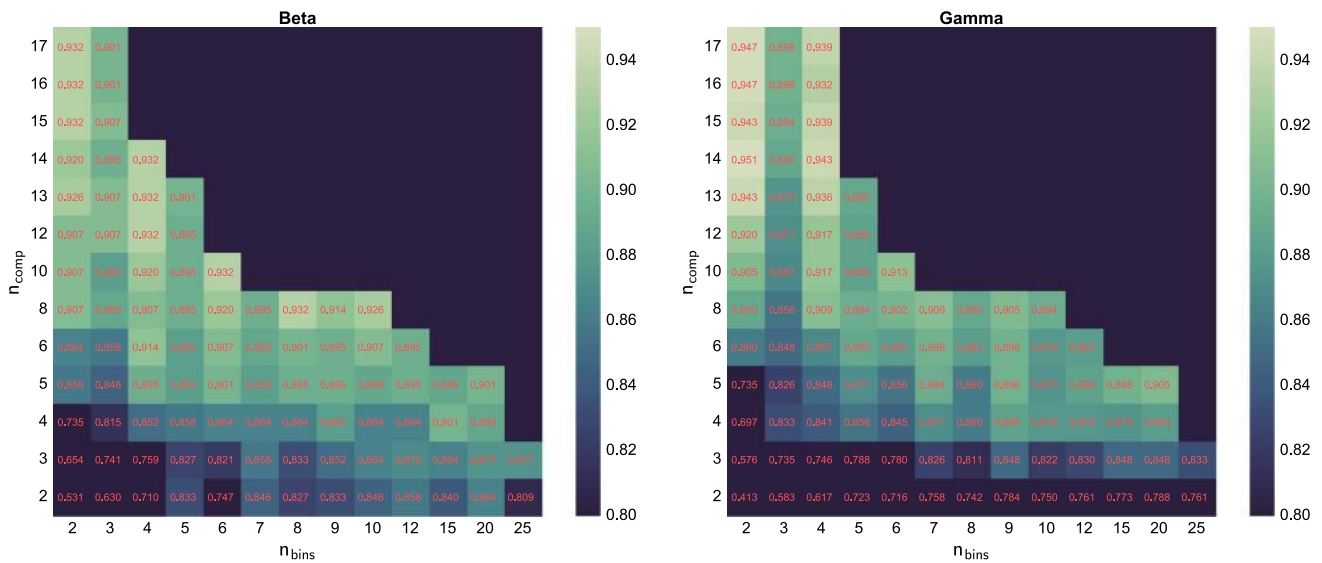


Fig. 8 The classification performance of the STRF N -jet descriptor as function of the number of bins and the number of principal components for the Beta and Gamma benchmarks. The best results are obtained when using more than eight principal components in combination with

binary or ternary histograms. Thus, for these more complex benchmarks, discarding too much information in the dimensionality reduction step impairs the performance. Spatial scales: $(\sigma_{s_1}, \sigma_{s_2}) = (4, 8)$ pixels. Temporal scales: $(\sigma_{t_1}, \sigma_{t_2}) = (100, 200)$ ms

in combination with a small number of bins means that the different qualitative “types” of patterns are more diverse, while at the same time being less “precise” in the sense of being unaffected by small changes in the magnitude of the filter responses. Binary or ternary descriptors are thus less sensitive to variations of the same rough type of space–time structure. Indeed, for binary descriptors only the sign of the receptive

field response is recorded and a binary descriptor thus gives full invariance to, for example, multiplicative illumination transformations.

For the larger Beta and Gamma benchmarks, it is clear that the descriptors that in this way combine a large number of image measurements with binary or ternary histograms achieve superior performance. This indicates that for these

larger more complex datasets, capturing the essence of a local space–time pattern rather than its more precise appearance is the right trade-off. In fact, here, the best results can for all benchmarks be achieved using $n_{\text{bins}} = 2$ and $n_{\text{comp}} \in [10, 17]$, with the single exception of 0.3 percentage points lower error on the UCLA8 dataset if instead using a ternary descriptor. A similar observation that binary histograms perform very well was made in previous work on spatial recognition [38]. Indeed, binary histograms based on the sign of the receptive field responses are independent on magnitude thresholds and invariant to multiplicative intensity transformations, which provides better robustness to illumination transformations.

We thus conclude that binary histogram descriptors are a very useful option, combining top performance with simplicity. Therefore, we in the following investigate the effect of varying the remaining descriptor parameters using binary descriptors only.

7.2 Spatial and Temporal Scales

Each dataset will have a set of scales that are better for describing the spatial patterns and the motion patterns present in the videos. The classification performance of the STRF N -jet descriptor as function of the spatial and the temporal scales of the receptive fields for different combinations of a single spatial scale $\sigma_s \in \{1, 2, 4, 8, 16\}$ and a single temporal scale $\sigma_t \in \{1, 2, 4, 8, 16\}$ is shown in Fig. 9 for the UCLA benchmarks and in Fig. 10 for the DynTex benchmarks. All results have been obtained with $n_{\text{comp}} = 15$ and $n_{\text{bins}} = 2$.

For all the UCLA benchmarks, an approximately unimodal maximum over scales is obtained. For the UCLA8

and UCLA9 benchmarks, the best performance is obtained when combining a smaller spatial scale with a shorter temporal scale. For the UCLA50 benchmark, the best results are instead achieved for shorter temporal scales in combination with larger spatial scales. The observation that a short temporal scale works well for all benchmarks could indicate that fast motions are discriminative and that the best spatial scales are *different* for UCLA50 is not strange, since this benchmark features instance recognition (e.g. separating 108 different plants) rather than generalising between classes. Although it might feel intuitive that small details should be useful for instance recognition, this will depend on the dataset. For example, plants with similar leaves but different global growth patterns could be easier to separate at a larger spatial scale.

For the DynTex benchmarks, the scale combinations that give the best results are scattered rather than showing a unimodal maximum. This could indicate that the different subsets of dynamic textures are best separated at different (and non-adjacent) scales. Since the DynTex dataset is quite diverse, this would not be strange. It should also be noted that the differences between the best and the second best results are here typically only one or two correctly classified videos. It is, however, clear that using the largest spatial scale in combination with the longest temporal scale gives markedly worse results.

When using 2×2 scales, we noted a similar performance pattern during scale tuning with unimodal maxima for the UCLA benchmarks and scattered maxima for the DynTex benchmarks (not shown). Comparing the absolute performance when using single versus multiple scales, it depends on the receptive field set if using multiple scales gives a consistent advantage. If inspecting the sets of optimal parameters

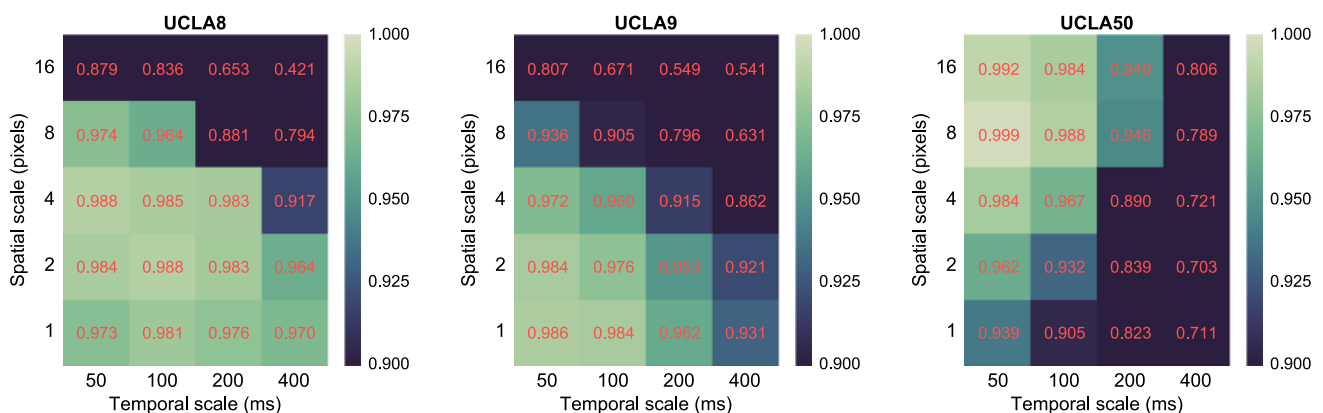


Fig. 9 The classification performance as function of the spatial and the temporal scales for the UCLA benchmarks. The maps show the classification accuracy for different combinations of a single spatial scale $\sigma_s \in \{1, 2, 4, 8, 16\}$ and a single temporal scale $\sigma_t \in \{50, 100, 200, 400\}$ for the descriptor STRF N -jet with $n_{\text{bins}} = 2$ and $n_{\text{comp}} = 17$. For

the UCLA8 and UCLA9 benchmarks, the best performance is obtained for a relatively small spatial scale in combination with a short temporal scale. For the UCLA50 benchmark, instead a combination of a relatively small spatial scale and a longer temporal scale gives the best results

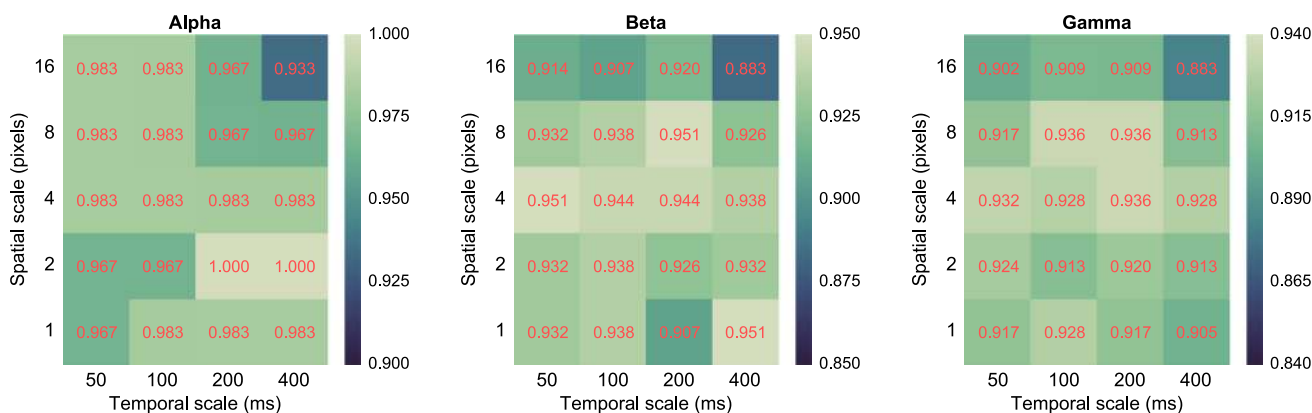


Fig. 10 The classification performance as function of the spatial and the temporal scales for the DynTex benchmarks. The maps show the classification accuracy for different combinations of a single spatial scale $\sigma_s \in \{1, 2, 4, 8, 16\}$ and a single temporal scale $\sigma_t \in \{50, 100, 200, 400\}$ for the descriptor STRF N -jet with $n_{bins} = 2$ and $n_{comp} = 17$. Note the

different ranges used for colour coding the maps. Several local optima are obtained for all the three benchmarks. This suggests that different dynamic texture classes are best separated at different (non-adjacent) scales, which could be due to the larger diversity of the dynamic texture types present in these benchmarks

found for the different benchmarks (presented in “Appendix A.2”), it can be noted that, for the STRF N -jet descriptor, the best results are sometimes achieved using a single scale and sometimes when using 2×2 scales. However, STRF N -jet includes a quite large number (17) of receptive fields and when using a smaller set of receptive fields, such as in RF Spatial (5 receptive fields) or STRF RotInv (9 receptive fields), video descriptors using multiple spatio-temporal scales consistently have the best performance. This shows that receptive fields at different scales can contain complementary information.

We conclude that, although results competitive with many state-of-the-art methods can be obtained for a heuristic choice of spatial and temporal scales, using parameter tuning to find an appropriate scale/set of scales may lead to improved performance of a few percentage points.

7.3 Receptive Field Sets

In this section, we present results on relative performance between our four proposed video descriptors constructed from different sets of receptive fields (see Sect. 4.6.1):

- (i) The new spatio-temporal descriptor *STRF N-jet*.
- (ii) The rotationally invariant descriptor *STRF RotInv*.
- (iii) The purely spatial *RF Spatial*.
- (iv) The previous spatio-temporal descriptor *STRF N-jet* (previous) [27].

A comparison of the classification performance of these four video descriptors across all benchmarks is shown in Fig. 11. The performance of all four video descriptors is also compared to state of the art in Tables 2 and 3.

7.3.1 STRF N -Jet (Previous) Versus STRF N -Jet

We note that parameter tuning and adding the second-order temporal derivatives of the spatial derivatives result in improved performance for our new STRF N -jet descriptor compared to the STRF N -jet (previous) descriptor [27]. The new descriptor shows improved accuracy for all the benchmarks. We have also observed an improvement from both these changes individually (not explicitly shown here).

7.3.2 Spatio-Temporal Versus Spatial Descriptors

A comparison between the STRF N -jet descriptor and RF Spatial reveals improved accuracy when including spatio-temporal receptive fields for the UCLA8, UCLA9, Alpha and Gamma benchmarks. Note that a comparison to the STRF N -jet (previous) descriptor is less relevant, since that descriptor is in contrast to the others not subject to parameter tuning.

The largest improvement is obtained for the Gamma benchmark, where adding spatio-temporal receptive fields reduces the error from 9.5 to 4.5% when using an SVM classifier. Smaller improvements are obtained for the UCLA8 and UCLA9 benchmarks, with a reduction in error from 2.2 to 1% and from 1.4 to 0.8 %, respectively. For the UCLA50 benchmarks, the performance saturates at 100% for both descriptor types (rather indicating the relative simplicity of this benchmark). The only exception where RF spatial shows better performance is for the Beta benchmark using a NN classifier. Here, the purely spatial descriptor achieves 5.6% error versus 6.2% error for STRF N -jet.

Competitive performance for purely spatial descriptors on the Beta benchmark has been reported previously [56], and we here make a similar observation. Thus, not surprisingly,

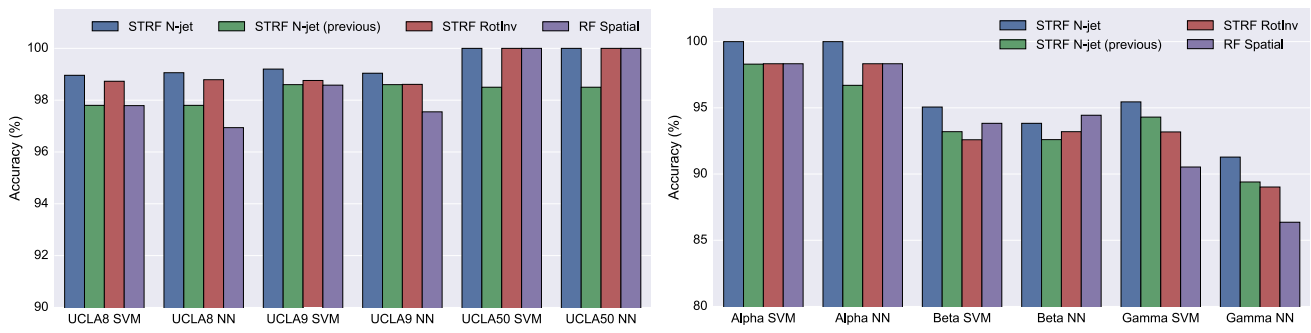


Fig. 11 The classification performance for video descriptors constructed from different sets of receptive fields. Left: For the UCLA benchmarks. Right: For the DynTex benchmarks. The STRF *N*-jet descriptor achieves improved performance compared to RF Spatial (with the single exception of the Beta NN benchmark). This shows that the spatio-temporal receptive fields provide complementary infor-

mation. The rotationally invariant STRF RotInv descriptor is also a competitive option. The new STRF *N*-jet achieves better results than STRF *N*-jet (previous) [27]. All descriptors are binary. Additional parameters for each of these benchmark results are given in “Appendix A.2”

Table 2 Comparison to state of the art for the UCLA benchmarks

	UCLA8		UCLA9		UCLA50		
	SVM	NN	SVM	NN	SVM	NN	
<i>STRF N-jet</i>	99.0	99.1	99.2	99.0	100	100	greyscale
DNGP [64]	99.4	97.0	99.6	98.1	–	–	greyscale
OTD [61]	99.5	97.0	98.2	97.5	99.8	98.5	greyscale
DT-CNN [1]	99.0	–	98.4	–	99.5	–	colour, deep learning
<i>STRF RotInv</i>	98.7	98.8	98.8	98.6	100	100	greyscale
3D-OTF [82]	99.5	95.8	97.2	96.3	87.1	99.25	greyscale
Ensemble SVMs [84]	–	–	–	–	100	–	greyscale
Enhanced LBP [63]	–	–	–	98.2	–	100	greyscale
MBSIF-TOP [3]	–	97.8	–	98.8	–	99.5	greyscale
<i>STRF N-jet (previous)</i> [29]	97.8	97.5	98.6	98.3	98.5	97.0	greyscale
MEWLSP [73]	–	98.0	–	98.6	–	96.5	greyscale
<i>RF Spatial</i>	97.8	96.9	98.6	97.6	100	100	greyscale
SKDL [60]	98.6	–	–	–	–	–	greyscale
HOG-NSP [55]	98.7	–	98.1	–	97.2	–	greyscale
WMFS [30]	97.0	97.2	97.1	97.0	99.8	99.1	greyscale
PCA-net TOP [2]	–	–	–	–	99.5	–	greyscale
CVLBP [86], from [73]	–	95.7	–	96.9	–	93.0	greyscale
DL-PEGASOS [19]	–	–	–	95.6	–	99.0	greyscale
Temporal dropout CNN [8]	–	–	–	–	–	98.0	greyscale, deep learning
VLBP [86], from [73]	–	92.0	–	96.3	–	89.5	greyscale
Oriented energy rep. [12]	–	–	–	–	–	81.0	greyscale

Our proposed *STRF N-jet* descriptor shows consistently very competitive results for all these benchmarks, achieving the highest mean accuracy averaged over all benchmarks as well as the single best result on four out of the six benchmarks. All the STRF and RF descriptors are binary ($n_{bins} = 2$), and parameter tuning has been performed except for *STRF N-jet (previous)*. The parameter values are given in “Appendix A.2”. Bold font = highest accuracy for each single benchmark. Highlighted rows = our proposed descriptors

for some settings genuine spatio-temporal information is of greater importance than for others. Here, the largest gain is indeed obtained for the most complex task.

7.3.3 Rotationally Invariant Descriptors

The rotationally invariant STRF *N*-jet RotInv descriptor does not achieve fully as good performance as the directionally dependent STRF *N*-jet descriptor for the tested benchmarks. The difference in classification accuracy in favour of the

directionally selective descriptor is most pronounced for the more complex DynTex benchmarks: STRF RotInv achieves 7.4 and 6.8% error on the Beta and Gamma benchmarks using an SVM classifier, compared to STRF *N*-jet with 4.9 and 4.5% error, respectively. However, a comparison with state of the art in Tables 2 and 3 reveals that the STRF *N*-jet RotInv descriptor still achieves competitive performance compared to other dynamic texture recognition approaches.

It is of conceptual interest that these good results can be obtained also when disregarding orientation informa-

Table 3 Comparison to state of the art for the DynTex benchmarks

	Alpha		Beta		Gamma		
	SVM	NN	SVM	NN	SVM	NN	
DT-CNN [1]	100	–	100	–	99.6	–	colour, deep-learning
st-TCoF [58]	100	98.3	100	98.1	98.1	98.1	colour, deep-learning
Deep Dual (D3) [25]	100	–	100	–	98.1	–	colour, deep-learning
Ensemble SVMs [84]	–	–	–	–	99.5	–	colour
MR-SFA [52]	–	–	99.0	98.1	–	–	greyscale
<i>STRF N-jet</i>	100	100	95.1	93.8	95.5	91.2	greyscale
<i>STRF N-jet (previous)</i> [29]	98.3	96.7	93.2	92.6	94.3	89.4	greyscale
<i>STRF RotInv</i>	98.3	98.3	92.6	93.2	93.2	89.0	greyscale
<i>RF Spatial</i>	98.3	98.3	93.8	94.4	90.5	86.4	greyscale
SoB + Align [65]	–	98.3	–	90.1	–	79.9	greyscale
PCANet-TOP [2]	–	96.7	–	90.7	–	89.4	greyscale
MBSIF-TOP [3]	–	90.0	–	90.7	–	91.3	greyscale
AFS-TOP [26]	98.3	91.7	90.1	86.4	94.3	89.4	greyscale
LBP-TOP [87], from [58]	98.3	96.7	88.9	85.8	94.2	84.9	greyscale
ELM [75]	–	–	93.8*	–	88.3*	–	greyscale
SKDL [60]	88.8*	–	77.4*	–	75.6*	–	greyscale
2D+T curvelet [13]	–	88.0 [†]	–	70.0 [†]	–	68.0 [†]	greyscale
OTD [61]	87.8*	86.6 [†]	76.7*	69.0 [†]	74.8*	64.2 [†]	greyscale
DFS [83]	85.2*	–	76.9*	–	74.8*	–	greyscale

Our proposed *STRF N-jet* descriptor ranks at the very top among the grey-scale methods showing better performance than a large range of similar methods using different spatio-temporal primitives. All the STRF and RF descriptors are binary ($n_{\text{bins}} = 2$), and parameter tuning has been performed except for *STRF N-jet (previous)*. The parameter values are given in “Appendix A.2”. Bold font = best greyscale descriptor for each single benchmark, italics font = best colour descriptor, *indicates a different train/test partitioning for SVM and [†] the use of a nearest centroid classifier. Highlighted rows = our proposed descriptors

tion completely. Indeed, if considering marginal histograms of receptive field responses, the most striking differences between texture classes such as waves, grass and foliage are the typical directions of change. (Waves show a stronger gradient in the vertical directions grass in the horizontal and foliage in both.) A qualitative conclusion is that directional information is not the main mode of recognition here; instead, the local space–time structure *independent of orientation* is highly discriminative. We conclude that our proposed STRF RotInv descriptor could be a viable option for tasks where rotation invariance is of greater importance than for these benchmarks. However, the possible gain from enabling recognition of textures at orientations not present in the training data will have to be balanced against the possible gain from discriminative directional information.

7.4 Comparison to State of the Art

This section presents a comparison between our proposed approach and state-of-the-art dynamic texture recognition methods. We include video descriptors constructed from four different sets of receptive fields (see Table 1) and compare against the best-performing methods found in the literature for each benchmark. We also aim to include a range of *different types* of approaches with an extra focus on methods similar to ours, i.e. different LBP versions and relatively shallow (max 2 layers) spatio-temporal filtering-based

approaches using either handcrafted filters or filters learned from data. Results for all the other methods are taken from the literature, where the relevant references are indicated in the table.

7.4.1 UCLA Datasets

The UCLA benchmark results are presented in Table 2. Our proposed STRF *N-jet* descriptor shows highly competitive performance compared to all the other methods, achieving the highest mean accuracy averaged over all the benchmarks and either the single best or the shared best result on four out of the six benchmarks.

For the UCLA50 benchmark, our three new video descriptors achieve 0% error using both an SVM and a NN classifier. The main difference between these descriptors and the untuned STRF *N-jet (previous)* is the use of a larger spatial scale, which is seen in Sect. 7.2 to be more adequate for this benchmark. Enhanced LBP [61] and Ensemble SVMs [84] also achieve 0% error rate, and there are several methods with error rates below 0.5%. The main conclusions we draw from the UCLA50 results are that recognising the same dynamic texture instance from the same viewpoint is (not surprisingly) an in comparison easier task than separating conceptual classes and that our approach performs on par with the best state-of-the-art methods on this task.

For the conceptual UCLA8 and UCLA9 benchmarks using an NN classifier, our STRF N -jet descriptor achieves 0.9 and 1.0% error, respectively, which are the single best results among all methods. This demonstrates that our approach is stable and works well with a simple classifier also for a quite high-dimensional descriptor. For the UCLA8 benchmark together with a NN classifier, the second best-performing approach is our rotational-invariant descriptor STRF RotInv with 1.2% error and after that MEWLSP [72] with 2% error. For UCLA9, the second best-performing approach is MBSIF-TOP [3] with 1.2% error followed by STRF RotInv and MEWLSP, which both achieve 1.4% error.

For the UCLA8 benchmark combined with an SVM classifier, the best-performing approaches are OTD [59] and 3D-OTF [82] both with 0.5% error. For UCLA9, the best method using an SVM classifier is DNGP [62], which achieves 0.4% error. Our STRF N -jet descriptor achieves 1.0% error on the UCLA8 benchmark and 0.8% error on the UCLA9 benchmark. It should be noted that OTD, 3D-OTF and DNGP simultaneously show considerably worse results on the NN benchmarks and that the standard UCLA protocol (average over 20 trials) can give quite variable results because of the limited number of samples in the benchmarks. Averaging over 1000 trials means that our results are more stable and less likely to include “outliers” for some of the benchmarks.

Our approach shows improved results on all the UCLA benchmarks compared to a large range of similar methods also based on gathering statistics of local space–time structure but using *different spatio-temporal primitives*. This includes methods that are more complex in the sense of combining several different descriptors or a larger number of feature extracting steps (MEWLSP [72], HOG-NSP [53]), methods learning higher-level hierarchical features (PCANet-TOP [2], SKDL [58]), temporal dropout DL, DT-CNN [1]) and improved and extended LBP-based methods (Enhanced LBP [61], MBSIF-TOP [3], MEWLSP [72], CVLBP [71]) as well as the standard LBP-TOP [87] and VLBP [86] descriptors. An interesting observation is also that compared to VLBP and CVLBP, which is similar to our approach use binary histograms and full 2D+T primitives, the performance of our approach is 2.1–10.5 percentage points better for all the benchmarks. The most important difference between these methods and our approach is indeed the spatio-temporal primitives used for computing the histogram.

7.4.2 DynTex Datasets

The DynTex benchmark results are presented in Table 3. For this larger and more complex dataset, it can be seen that utilising colour information and supervised hierarchical feature learning seems to give a clear advantage with three deep learning approaches on top. DT-CNN, trained from scratch

to extract features on three orthogonal planes, demonstrates the best performance with 0% error on the Alpha and Beta benchmarks and 0.4% error on the Gamma benchmark. Two deep learning methods based on feature extraction using pre-trained networks (Deep Dual descriptor and st-TCof) also obtain very good results. However, although included for reference, we do not directly aim here to compete with these more conceptually complex methods. The main focus of our work is instead to evaluate the usefulness of the time-causal spatio-temporal primitives without entanglement with a more complicated hierarchical framework.

Our proposed STRF N -jet descriptor achieves 0% error on the Alpha benchmark using both an SVM and a NN classifier, 4.9% (SVM) and 6.2% (NN) error on the Beta benchmark and 4.5% (SVM) and 8.8% error (NN) on the Gamma benchmark. This means that we achieve better results than all other non-deep learning methods utilising only grey-scale information except one: MR-SFA [50] which achieves 1% error on the Beta SVM benchmark and 1.9% error on the Beta NN benchmark. (This method has not been tested on the Alpha and Gamma benchmarks.) It should, however, be noted that MR-SFA uses regional descriptors capturing the relative location of image structures and a bag-of-words framework on top of the histogram descriptor. This approach is thus significantly more complex compared to our method. Both these extensions would be relatively straightforward to implement also using our proposed video descriptors.

Our results can also be compared to the LBP-TOP extension AFS-TOP, which shows the most competitive results using an SVM classifier with 1.7, 9.9 and 5.7% error, respectively, on the Alpha, Beta and Gamma benchmarks. Our approach thus achieves better performance on all the tested benchmarks, although AFS-TOP includes several added features, such as removing outlier frames. Improvements compared to the basic LBP-TOP descriptor are larger, and this can be considered a more fair benchmark, since we are testing an early version of our approach. Compared to MBSIF-TOP and PCANet-TOP, which both learn 2D filters from data and apply those on three orthogonal planes, our approach also achieves better results on all the DynTex benchmarks.

We also show notably better results (in the order of 10–20 percentage points) than those reported from using DFS [83], OTD [59], SKDL [58] and the 2D+T curvelet transform [13]. However, those use a nearest centroid classifier and a different SVM train-test partition, which means that a direct comparison is not possible. We also note that although STRF N -jet achieves the best results, the rotationally invariant descriptor version STRF RotInv, the RF Spatial descriptor and the untuned STRF N -jet (previous) descriptor also achieve competitive performance. This demonstrates the robustness and flexibility of our approach.

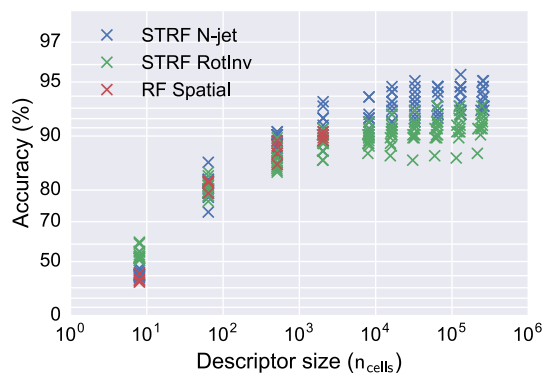


Fig. 12 Descriptor size versus classification accuracy for the binary STRF *N*-jet, STRF RotInv and RF Spatial video descriptors evaluated on the Gamma benchmark. The performance for each descriptor version is plotted against the average descriptor size for all different combinations of two spatial and two temporal scales (see Sect. 6.3) and for different numbers of principal components (see Table 4). The accuracy is presented as $\log(\text{error}) = \log(100 - \text{accuracy})$ with the vertical axis reversed

In conclusion, our approach shows highly competitive performance for this larger and more complex benchmark, even though our proposed approach is a conceptually simple method utilising only local information. The STRF *N*-jet descriptor achieves better performance than all other grey-scale methods of similar conceptual complexity and better performance compared to both several more complex methods and two methods learning filters from data. We believe this should be considered as strong validation that these time-causal spatio-temporal receptive fields indeed capture useful spatio-temporal information.

7.5 Descriptor Size Versus Performance

The proposed family of video descriptors may vary substantially in descriptor size, depending on the parameter choices made. We here illustrate the trade-off between descriptor size and classification performance for binary descriptors on the most challenging Gamma benchmark. Size is here defined as the average number of non-empty cells for a certain descriptor type and benchmark. This also gives an estimate on the sparsity of the histograms.

The classification accuracy together with the descriptor size for increasing numbers of principal components n_{comp} for the binary STRF *N*-jet, STRF RotInv and RF Spatial descriptors is shown in Fig. 12. We include all combinations of 2×2 scales (see Sect. 4.6.4) for each value of n_{comp} to illustrate the overall trend rather than focusing on one specific choice of spatial and temporal scales. Table 4 shows the best classification performance (after scale tuning) for the STRF *N*-jet descriptor on the Gamma benchmark for each value of n_{comp} together with the average number of non-empty histogram cells.

Table 4 Descriptor size versus classification accuracy for the binary STRF *N*-jet descriptor evaluated on the Gamma benchmark

n_{comp}	n_{cells} (filled)	Accuracy (%)
2	4.0×10^0	45.5
5	3.2×10^1	86.0
8	2.6×10^2	90.5
10	1.0×10^3	93.6
12	4.1×10^3	93.9
13	8.2×10^3	94.7
14	1.6×10^4	95.1
15	3.3×10^4	94.7
16	6.5×10^4	95.5
17	1.3×10^5	95.1

The table shows the average descriptor size (the number of nonzero histogram cells) together with the top accuracy obtained for video descriptors using different numbers of principal components

For all video descriptors, it can be seen that the performance first increases fast with the descriptor size (here, equivalently the number of principal components used) up to around $n_{\text{comp}} = 10$. For the STRF *N*-jet descriptor, some additional gains are then obtained, primarily between 10 and 14 principal components, after which the performance saturates. For the STRF RotInv descriptor, the performance saturates at a lower level than for the STRF *N*-jet descriptor, indicating that by using only rotationally invariant receptive fields some discriminative information is lost. For the RF spatial descriptor, the best performance is obtained using all principal components (5 receptive fields \times 2 spatial scales).

From Table 4, it can further be seen that the best results are achieved for $8200 \leq n_{\text{cells}} \leq 130000$, with accuracy in the range 94.7–95.5%. However, also an even smaller descriptor using $n_{\text{cells}} = 1024$ gives reasonably good performance. We conclude that good performance is not contingent on a very large descriptor size and that our approach can be viable also in settings where a descriptor with a small memory footprint is required.

Comparing the number of filled histogram cells with the corresponding powers of two, it can be seen that these binary histograms are almost filled. We found this to be true for the other binary histogram types tested here as well. In contrast to this result, we have noted as low as 0.01% filled histogram cells for some of the larger non-binary histograms tested in Sect. 7.1.

Indeed, for non-binary histograms the more precise conditions on the magnitude of the receptive field responses will imply a larger set of very “rare” patterns. Thus, for the larger non-binary histograms there can be large benefits from using a sparse representation, while for the binary histograms tested here, a non-sparse representation is most advantageous.

7.6 Recognition Using Fewer Frames

In most real-world scenarios, an artificial or biological agent will get advantages from being able to interpret, and thus react, to its surroundings as fast as possible. We here investigate how the recognition performance of our proposed approach is affected by gathering statistics over a smaller interval of frames as opposed to over the full video. The experimental set-up is chosen to reflect the most likely real-world scenario—full videos (or scenes) are available during training but classification of a new dynamic texture is made from a single shorter sequence.

It should be noted that the stationarity properties of dynamic textures will often manifest over timescales longer than a few frames. Thus, it is important that local (in time) parts of motion patterns can be matched to other local (in time) patterns, rather than forcing also a short sequence to match the statistics of a full video. We, here, extract a collection of shorter sequences of length $n_{\text{frames}} \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ from each video. Then, during training, the full set of shorter sequences for each value of n_{frames} is used as training examples, while for testing the recognition is based on a single shorter sequence.

We study the performance of the binary STRF N -jet descriptor with $n_{\text{comp}} \in \{10, 13\}$ on the Beta and Gamma benchmarks. The results are presented in Fig. 13. It can be seen that for $n_{\text{comp}} = 13$ performance close to the baseline can be achieved also using very few frames. For the Beta

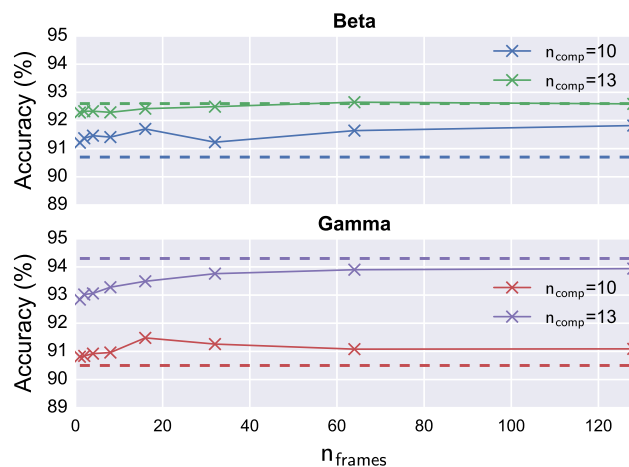


Fig. 13 Recognition from fewer frames for the Beta and Gamma benchmarks using the binary STRF N -jet descriptor. The plots show the classification accuracy as a function of the number of frames used to gather statistics for the video descriptor. The dashed line represents the performance obtained when using the full video. For $n_{\text{comp}} = 13$, results close to baseline performance can be obtained also when using very few frames. For $n_{\text{comp}} = 10$, the performance is improved when training and testing using shorter sequences for all values of n_{frames} . Spatial scales: $(\sigma_{s_1}, \sigma_{s_2}) = (4, 8)$ pixels. Temporal scales: $(\sigma_{\tau_1}, \sigma_{\tau_2}) = (100, 200)$ ms

benchmark, there is 0.3% difference between using a single frame and the entire video (92.3 vs 92.6 %), while for the Gamma benchmark this difference is 1.1% (92.8 vs 93.9%). When using a smaller descriptor with $n_{\text{comp}} = 10$, we note that the performance is instead *improved* when training and testing using shorter sequences, where the best values of n_{frames} give accuracy 1% above the baseline for both benchmarks. Note that, in a real-time scenario, also the temporal delay of the time-causal filters will influence the reaction time and that at least three frames are needed to compute a discrete second-order temporal derivative approximation.

The reason why using less data/fewer frames might give a drop in performance is clear—there is simply less information that can be used for making a decision. The reason why using shorter sequences can improve the performance is most likely related to that the matching becomes more flexible—each shorter sequence of a video only needs to have a similar representation to *at least one* shorter sequence from another video in the same class, rather than requiring the full videos to match.

In conclusion, these results demonstrate that our approach is robust to using descriptors computed from shorter video sequences and that it should thus be a viable option also for real-time scenarios where a quick decision is needed.

7.7 Qualitative Results

To gain more insight into the qualitative behaviour of our proposed family of video descriptors, we inspected the confusions and the closest neighbours to correctly classified and misclassified samples. Confusion matrices for the UCLA9, the Beta and the Gamma benchmarks for the STRF N -jet descriptor are presented in Fig. 14. We note that the main cause of error for UCLA9 is confusing fire with smoke. UCLA8 (not shown here) shares a similar pattern. There is indeed a similarity in dynamics between these textures in the presence of temporal intensity changes not mediated by spatial movements. Confusions between flowers and plants and between fountain and waterfall are most likely caused by similarities in the spatial appearance and the motion patterns of these dynamic texture classes.

When inspecting the confusions between the different classes for the Beta and Gamma benchmarks, there is no clear pattern visible. This is probably partly due to the fact that these benchmarks contain larger intraclass variabilities. Misclassifications seem to be caused by a single video in one class having some similarity to a single video of different class rather than certain classes being consistently mixed up. We note the largest ratio of misclassified samples for the escalator and traffic classes, which are also the classes with the fewest samples.

A bit more light is shed on the reasons for misclassifications for the DynTex benchmarks when inspecting the closest

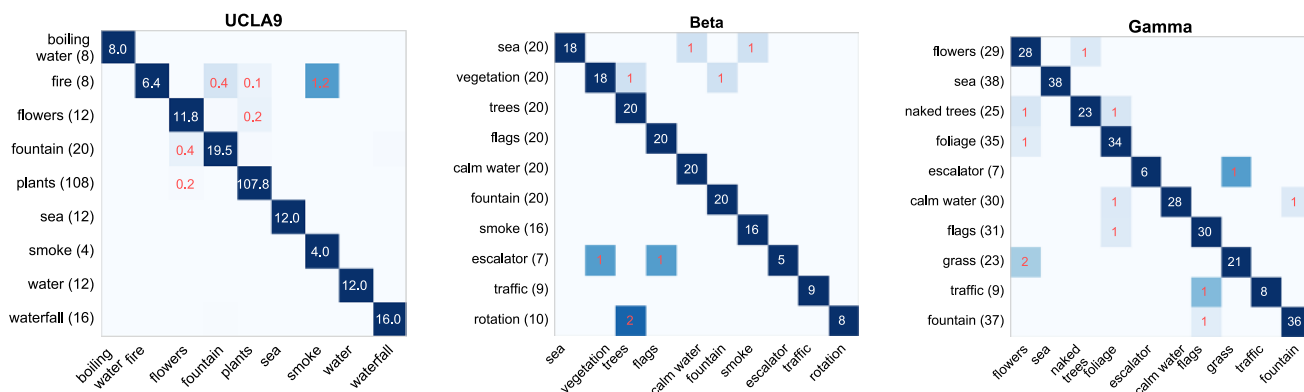


Fig. 14 Confusion matrices for the STRF *N*-jet descriptor combined with an SVM classifier. Left: For the UCLA9 benchmark (averaged over trials). Parameters: $\sigma_s = (4, 8)$, $\sigma_\tau = (50, 100)$, $n_{comp} = 14$. Here, the largest confusion is between fire and smoke. Middle: For the

Beta benchmark. Parameters: $\sigma_s = 8$, $\sigma_\tau = 200$, $n_{comp} = 17$. Right: For the Gamma benchmark. Parameters: $\sigma_\tau = (50, 100)$, $n_{comp} = 16$. Here, no single class is dominating among the confusions. See also Fig. 16

neighbours in feature space for misclassified videos (Fig. 15) and for correctly classified videos (Fig. 16). We note that a frequent feature of the misclassified videos is the presence of multiple textures, such as a flag with light foliage in front of it (misclassified as foliage), or a fountain flowing into a pool of calm water (misclassified as calm water). There are also examples of confusion caused by similarity in either the spatial appearance or the temporal dynamics between specific instances of different classes, such as calm water reflecting a small tree being misclassified as foliage or a field of grass waving in the wind misclassified as sea.

A subset of the misclassifications can likely be resolved if utilising colour information, since colour can be highly discriminative for dynamic textures. When considering the state-of-the-art results summarised in Table 3, these also indicate that a “deeper” framework, extracting more high-level abstract features would most likely improve the performance. Options that may be considered are adding additional layers of abstraction before or after the histograms, for example, by learning hierarchical features from data or using bag-of-words models.

8 Summary and Discussion

We have presented a new family of video descriptors based on joint histograms of spatio-temporal receptive field responses and evaluated several members in this family on the problem of dynamic texture recognition. This is, to our knowledge, the first video descriptor that uses *joint statistics* of a set of “ideal” (in the sense of derived on the basis of pure mathematical reasoning) spatio-temporal scale-space filters for video analysis and the first quantitative performance evaluation of using the family of *time-causal scale-space filters* derived by Lindeberg [44] as primitives for video analy-

sis. Our proposed approach generalises a previous method by Linde and Lindeberg [37,38], based on joint histograms of receptive field responses, from the spatial to the spatio-temporal domain and from object recognition to dynamic texture recognition.

Our experimental evaluation on several benchmarks from two widely used dynamic texture datasets demonstrates competitive performance compared to state-of-the-art dynamic texture recognition methods. The best results are achieved for our binary STRF *N*-jet descriptor, which combines directionally selective spatial and spatio-temporal receptive field responses. For the UCLA benchmarks, the STRF *N*-jet descriptor achieves the highest mean accuracy averaged over all benchmarks as well as the shared best or single best result for several benchmarks. In addition, our approach achieves improved results on all the UCLA benchmarks compared to a large range of similar methods also based on gathering statistics of local space–time structure but using *different spatio-temporal primitives*.

For the larger more complex DynTex benchmarks, deep learning approaches come out on top. However, also here our proposed video descriptor achieves improved performance compared to a range of similar methods, such as local binary pattern-based methods including recent extensions and improvements [24,61,86,87] and two methods learning filters by means of unsupervised learning [2,3]. The improved performance compared to approaches that learn filters from data (where for the UCLA benchmarks this also includes two deep learning-based approaches), shows that designing filters based on structural priors of the world can for these tasks be as effective as learning.

It should be noted that we believe an extension of our framework, for example, by complementing the current single layer of receptive fields with hierarchical features, would, indeed, be beneficial to address more complicated visual



Fig. 15 A subset of correctly classified samples (left in each pair) for the Gamma benchmark together with their closest neighbours (right in each pair) when using the STRF N -jet descriptor together with an SVM classifier. Parameters: $\sigma_s = (4, 8)$, $\sigma_\tau = (50, 100)$, $n_{\text{comp}} = 17$



Fig. 16 A subset of misclassified samples (left in each pair with class name in black) for the Gamma benchmark together with their closest neighbours (right in each pair with class name in red) when using the STRF N -jet descriptor together with an SVM classifier. We note that sources of error are the presence of several dynamic textures in the same video (top row image pairs one, three and four and bottom row image

pair four) and the presence of large static image regions. The remaining cases appear to be based on spatial or temporal similarity between individual texture instances, such as a bush reflected in the water being classified as foliage (bottom leftmost pair) or an escalator misclassified as grass (top row second image pair). Parameters: $\sigma_s = (4, 8)$, $\sigma_\tau = (50, 100)$, $n_{\text{comp}} = 17$

tasks. Therefore, the presented approach is not intended as a final point, but rather as a first conceptually simple approach for performing video analysis based on these new spatio-temporal primitives. Using a set of localised histograms, which capture the relative locations of different image structures, or learning or designing higher-level hierarchical features on top of such local spatio-temporal derivative responses, will most likely improve the performance. A single global histogram would of course not be appropriate for videos that are not presegmented or do not have the spatial and temporal stationarity properties of dynamic textures. To handle such scenes, regional histograms should instead be computed. Also, in the presence of global camera motion, velocity adaptation would be beneficial. However, the objective here has primarily been to use these new time-causal spatio-temporal primitives within a simple framework, to focus more on a first evaluation of the spatio-temporal primitives and less on what is to be built on top. Thus, the presented framework is not aimed at directly competing with such conceptually more complex methods. Our approach does not include combinations of different feature types, ensemble classifiers, regional pooling to capture the relative location of features or learned or handcrafted mid-level features. How-

ever, constructing a more complex framework on top of these spatio-temporal primitives is certainly possible and would with high probability result in additional performance gains.

In summary, our conceptually simple video descriptor achieves highly competitive performance across all benchmarks compared to other grey-scale “shallow” methods and improved performance compared to all other methods of similar conceptual complexity using different spatio-temporal primitives, either handcrafted or learned from data. We consider this as strong validation that these time-causal spatio-temporal receptive fields are highly descriptive for modelling the local space–time structure and as evidence in favour of their general applicability as primitives for other video analysis tasks.

Our approach could also be implemented using non-causal Gaussian spatio-temporal scale-space kernels. This might give somewhat improved results, since at each point in time, additional information from the future could then also be used. However, a time-delayed Gaussian kernel would imply longer temporal delays—thereby making it less suited for time-critical applications—in addition to requiring more computations and larger temporal buffers. The computational advantages of these time-causal filters imply that they can, in

fact, be preferable also in settings where the temporal causality might not be of primary interest.

The scale-space filters used in this work have a strong connection to biology in the sense that a subset of these receptive fields very well model both spatial and spatio-temporal receptive fields in the LGN and V1 [11,42,44]. The receptive fields in V1, V2, V4 and MT serve as input to a large number of higher visual areas. It does indeed appear attractive to be able to use similar filters for early visual processing over a wide range of visual tasks. This study can be seen as a first step in a more general investigation into what can be done with spatio-temporal features similar to those in the primate brain.

Directions for future work include: (i) A multi-scale recognition framework, where each video is represented by a set of descriptors computed at multiple spatial and temporal scales, both during training and testing. This would enable scale-invariant recognition according to the theory presented in “Appendix A.1”. (ii) An extension to colour by including spatio-chromo-temporal receptive fields would with high probability improve the performance for tasks where colour information is discriminative. (iii) Including non-separable spatio-temporal receptive fields with nonzero image velocities in the video descriptor according to the more general receptive field model in [44]. Such velocity-adapted receptive fields, in fact, constitute a dominant portion of the receptive fields in areas V1 and MT [11]. (iv) Using position-dependent histograms to take into account the relative locations of features. (v) Learning or designing higher-level hierarchical features based on these spatio-temporal primitives.

We propose that the spatio-temporal receptive field framework should be of more general applicability to other video analysis tasks. Time-causal spatio-temporal receptive fields are indeed the visual primitives used for solving a large range of visual tasks for biological agents. The theoretical properties of these spatio-temporal receptive fields imply that they can be used to design methods that are provably invariant or robust to different types of natural image transformations, where such invariances will reduce the sample complexity for learning. We thus see the possibility to both integrate time-causal spatio-temporal receptive fields into current video analysis methods and to design new types of methods based on these primitives.

Acknowledgements We would like to thank Oskar Linde for providing access to his code for computing joint receptive field histograms for spatial object recognition, which has influenced our implementation for spatio-temporal recognition.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

A Appendix

A.1 Covariance Properties of Partial Derivatives, PCA Components and Differential Invariants Over Spatio-Temporal Scales

The video descriptors that we construct are based on combinations of partial derivatives

$$L_{x^{m_1} y^{m_2} t^n} \tag{10}$$

of different orders (m_1, m_2, n) computed at multiple spatio-temporal scales (s, τ) , where s denotes the spatial scale parameter and τ the temporal scale parameter. Specifically, all video descriptors are expressed in terms of scale-normalized derivatives [40,44]

$$\partial_\xi = s^{\gamma_s/2} \partial_x, \quad \partial_\eta = s^{\gamma_s/2} \partial_y, \quad \partial_\zeta = \tau^{\gamma_\tau/2} \partial_t \tag{11}$$

written

$$L_{\xi^{m_1} \eta^{m_2} \zeta^n} = s^{(m_1+m_2)\gamma_s/2} \tau^{n\gamma_\tau/2} L_{x^{m_1} y^{m_2} t^n} \tag{12}$$

where $\gamma_s > 0$ and $\gamma_\tau > 0$ are scale normalisation powers of the spatial and the temporal domains, respectively.

Consider an independent scaling transformation of the spatial and the temporal domains

$$f'(x'_1, x'_2, t') = f(x_1, x_2, t) \tag{13}$$

for

$$(x'_1, x'_2, t') = (S_s x_1, S_s x_2, S_\tau t) \tag{14}$$

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

$$L(x_1, x_2, t; s, \tau) = (T(\cdot, \cdot, \cdot; s, \tau) * f(\cdot, \cdot, \cdot))(x_1, x_2, t; s, \tau) \tag{15}$$

$$L'(x'_1, x'_2, t'; s', \tau') = (T(\cdot, \cdot, \cdot; s', \tau') * f'(\cdot, \cdot, \cdot))(x'_1, x'_2, t'; s', \tau') \tag{16}$$

Then, for the spatio-temporal scale-space kernel defined from the Gaussian kernel $g(x, y; s)$ over the spatial domain and the time-causal limit kernel $\Psi(t; \tau, c)$ over the temporal domain

$$T(x, y, t; s, \tau) = g(x, y; s) \Psi(t; \tau, c) \tag{17}$$

these spatio-temporal scale-space representations will be closed under independent scaling transformations of the spatial and the temporal domains

$$L'(x'_1, x'_2, t'; s', \tau') = L(x_1, x_2, t; s, \tau) \tag{18}$$

provided that the spatio-temporal scale levels are appropriately matched [45, Equation (26)]

$$s' = S_s^2 s, \quad \tau' = S_\tau^2 \tau \tag{19}$$

This property holds for all nonzero spatial scaling factors S_s . Because of the discrete nature of the temporal scale levels $\tau_k = \tau_0 c^{2k}$ in the time-causal temporal scale-space representation obtained by convolution with the time-causal limit kernel, this closedness property does, however, only hold for temporal scaling factors S_τ that are integer powers of the distribution parameter c of the time-causal limit kernel

$$S_\tau = c^j \quad \text{for } j \in \mathbb{Z} \tag{20}$$

Specifically, under an independent scaling transformation of the spatial and the temporal domains (14), the partial derivatives transform according to [46, Equation (9)]

$$\begin{aligned} L'_{\xi^{m_1} \eta^{m_2} t^n} (x', y', t'; s', \tau') \\ = S_s^{(m_1+m_2)(\gamma_s-1)} S_\tau^{n(\gamma_\tau-1)} L_{\xi^{m_1} \eta^{m_2} \zeta^n} (x, y, t; s, \tau) \end{aligned} \tag{21}$$

In particular, for the specific choice of $\gamma_s = 1$ and $\gamma_\tau = 1$ these partial derivatives will be equal

$$L'_{\xi^{m_1} \eta^{m_2} t^n} (x', y', t'; s', \tau') = L_{\xi^{m_1} \eta^{m_2} \zeta^n} (x, y, t; s, \tau) \tag{22}$$

when computed at matching spatio-temporal scales (19).

Concerning the dimensionality reduction step, where a large set of scale-normalised partial derivatives $L_{\xi^{m_1} \eta^{m_2} \zeta^n}$ is replaced by a lower-dimensional set of PCA components

$$\tilde{F}_i = \sum_j w_{ij} L_{\xi^{m_{1,i,j}} \eta^{m_{2,i,j}} \zeta^{n_{i,j}}} \tag{23}$$

this implies that also the PCA components will be equal at matching spatio-temporal scales (19), although constituting linear combinations of partial derivatives of a different spatial orders $(m_{1,i,j}, m_{2,i,j})$ and temporal orders $n_{i,j}$.

This property does also extend to scale-normalised spatio-temporal differential invariants computed for $\gamma_s = 1$ and $\gamma_\tau = 1$ as well as to PCA components defined from lin-

Table 5 The parameter values selected for the UCLA benchmark results. Note that for the STRF N -jet, STRF RotInv and RF Spatial descriptors, parameter tuning has been performed according to Sect. 6.3,

while the STRF N -jet descriptor has been computed with the heuristically chosen parameters used in [27]

Descriptor	Benchmark	Classifier	n_{bins}	n_{comp}	Spatial scales (pixels)	Temporal scales (ms)
STRF N -jet	UCLA8	SVM	2	16	(4, 8)	(100, 200)
STRF N -jet	UCLA8	NN	2	16	(4)	(50)
STRF N -jet	UCLA9	SVM	2	14	(1, 2)	(50, 100)
STRF N -jet	UCLA9	NN	2	14	(1)	(50)
STRF N -jet	UCLA50	SVM	2	13	(4, 8)	(50, 100)
STRF N -jet	UCLA50	NN	2	12	(8, 16)	(50, 100)
STRF RotInv	UCLA8	SVM	2	14	(4, 8)	(50, 100)
STRF RotInv	UCLA8	NN	2	13	(4, 8)	(50, 100)
STRF RotInv	UCLA9	SVM	2	12	(4, 8)	(50, 100)
STRF RotInv	UCLA9	NN	2	13	(1, 2)	(100, 200)
STRF RotInv	UCLA50	SVM	2	6	(4, 8)	(50, 100)
STRF RotInv	UCLA50	NN	2	5	(4, 8)	(50, 100)
RF Spatial	UCLA8	SVM	2	10	(8, 16)	–
RF Spatial	UCLA8	NN	2	9	(8, 16)	–
RF Spatial	UCLA9	SVM	2	9	(4, 8)	–
RF Spatial	UCLA9	NN	2	6	(4, 8)	–
RF Spatial	UCLA50	SVM	2	5	(4, 8)	–
RF Spatial	UCLA50	NN	2	5	(4, 8)	–
STRF N -jet (previous)	UCLA8	SVM	2	15	(1, 2)	(50, 100)
STRF N -jet (previous)	UCLA8	NN	2	15	(1, 2)	(50, 100)
STRF N -jet (previous)	UCLA9	SVM	2	15	(1, 2)	(50, 100)
STRF N -jet (previous)	UCLA9	NN	2	15	(1, 2)	(50, 100)
STRF N -jet (previous)	UCLA50	SVM	2	15	(1, 2)	(50, 100)
STRF N -jet (previous)	UCLA50	NN	2	15	(1, 2)	(50, 100)

Table 6 The parameter values selected for the DynTex benchmark results. Note that for the STRF *N*-jet, STRF RotInv and RF Spatial descriptors, parameter tuning has been performed according to Sect. 6.3,while the STRF *N*-jet descriptor has been computed with the heuristically chosen parameters used in [27]

Descriptor	Benchmark	Classifier	n_{bins}	n_{comp}	Spatial scales (pixels)	Temporal scales (ms)
STRF <i>N</i> -jet	Alpha	SVM	2	17	(2)	(200)
STRF <i>N</i> -jet	Alpha	NN	2	11	(1)	(400)
STRF <i>N</i> -jet	Beta	SVM	2	17	(8)	(200)
STRF <i>N</i> -jet	Beta	NN	2	13	(2, 4)	(200, 400)
STRF <i>N</i> -jet	Gamma	SVM	2	16	(4, 8)	(50, 100)
STRF <i>N</i> -jet	Gamma	NN	2	13	(2, 4)	(200, 400)
STRF RotInv	Alpha	SVM	2	5	(8, 16)	(100, 200)
STRF RotInv	Alpha	NN	2	5	(8, 16)	(100, 200)
STRF RotInv	Beta	SVM	2	15	(4, 8)	(200, 400)
STRF RotInv	Beta	NN	2	17	(8, 16)	(100, 200)
STRF RotInv	Gamma	SVM	2	15	(2, 4)	(100, 200)
STRF RotInv	Gamma	NN	2	16	(2, 4)	(50, 100)
RF Spatial	Alpha	SVM	2	8	(8, 16)	–
RF Spatial	Alpha	NN	2	5	(4, 8)	–
RF Spatial	Beta	SVM	2	10	(2, 4)	–
RF Spatial	Beta	NN	2	10	(4, 8)	–
RF Spatial	Gamma	SVM	2	10	(2, 4)	–
RF Spatial	Gamma	NN	2	10	(2, 4)	–
STRF <i>N</i> -jet (previous)	Alpha	SVM	2	15	(2, 4)	(200, 400)
STRF <i>N</i> -jet (previous)	Alpha	NN	2	15	(2, 4)	(200, 400)
STRF <i>N</i> -jet (previous)	Beta	SVM	2	15	(2, 4)	(200, 400)
STRF <i>N</i> -jet (previous)	Beta	NN	2	15	(2, 4)	(200, 400)
STRF <i>N</i> -jet (previous)	Gamma	SVM	2	15	(2, 4)	(200, 400)
STRF <i>N</i> -jet (previous)	Gamma	NN	2	15	(2, 4)	(200, 400)

ear combinations of such scale-normalised spatio-temporal differential invariants.

In these respects, our video descriptors are truly scale-covariant under independent scaling transformations of the spatial and the temporal domains.

Corresponding spatio-temporal scale covariance properties do also hold for video descriptors defined from the non-causal Gaussian spatio-temporal scale-space representation computed based on spatial smoothing with the regular two-dimensional Gaussian kernel and temporal smoothing with the one-dimensional non-causal temporal Gaussian kernel. Then, the scale covariance properties hold for all nonzero spatial scaling constants S_s and all nonzero temporal scaling constants S_τ .

A.2 Parameters for Benchmark Results

This section presents the parameter settings used for comparing descriptors constructed from different sets of receptive fields (Sect. 7.3) and for comparing our approach with state of the art (Sect. 7.4). The parameters are shown for the UCLA

benchmarks in Table 5 and for the DynTex benchmarks in Table 6. For the STRF *N*-jet, STRF RotInv and RF Spatial descriptors, the parameters have been determined by cross-validation (see Sect. 6.3), while the STRF *N*-jet (previous) descriptor is tested with the heuristically chosen parameters used in [27] (and included here for completeness). In the quite frequent case that several sets of parameters give the same classification performance, only one of these parameter settings is reported here.

References

1. Andrearczyk, V., Whelan, P.F.: Convolutional neural network on three orthogonal planes for dynamic texture classification. *Pattern Recognit.* **76**, 36–49 (2018)
2. Arashloo, S.R., Amirani, M.C., Noroozi, A.: Dynamic texture representation using a deep multi-scale convolutional network. *J. Vis. Commun. Image Represent.* **43**, 89–97 (2017)
3. Arashloo, S.R., Kittler, J.: Dynamic texture recognition using multiscale binarized statistical image features. *IEEE Trans. Multimed.* **16**(8), 2099–2109 (2014)

4. Chan, A.B., Vasconcelos, N.: Classifying video with kernel dynamic textures. In: *Proceedings of Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 1–6. IEEE (2007)
5. Chetverikov, D., Fazekas, S., Haindl, M.: Dynamic texture as foreground and background. *Mach. Vis. Appl.* **22**(5), 741–750 (2011)
6. Chetverikov, D., Péteri, R.: A brief survey of dynamic texture description and recognition. In: Kurzyński, M., Puchała, E., Woźniak, M., Żołnierek, A. (eds.) *Computer Recognition Systems*, pp. 17–26. Springer, Berlin, Heidelberg (2005)
7. Crivelli, T., Cernuschi-Frias, B., Bouthemy, P., Yao, J.F.: Motion textures: modeling, classification, and segmentation using mixed-state Markov random fields. *SIAM J. Imaging Sci.* **6**(4), 2484–2520 (2013)
8. Culibrk, D., Sebe, N.: Temporal dropout of changes approach to convolutional learning of spatio-temporal features. In: *Proceedings of 22nd ACM International Conference on Multimedia*, pp. 1201–1204 (2014)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *Proceedings of Computer Vision and Pattern Recognition (CVPR 2005)*, vol. 1, pp. 886–893 (2005)
10. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *Proceedings of European Conference on Computer Vision (ECCV 2006)*, Lecture Notes in Computer Science, vol. 3952, pp. 428–441. Springer, Berlin (2006)
11. DeAngelis, G.C., Ohzawa, I., Freeman, R.D.: Receptive field dynamics in the central visual pathways. *Trends Neurosci.* **18**(10), 451–457 (1995)
12. 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)
13. Dubois, S., Péteri, R., Ménard, M.: Characterization and recognition of dynamic textures based on the 2D+T curvelet transform. *Signal Image Video Process.* **9**(4), 819–830 (2015)
14. El Moubtahij, R., Augereau, B., Fernandez-Maloigne, C., Tairi, H.: A polynomial texture extraction with application in dynamic texture classification. In: *Twelfth International Conference on Quality Control by Artificial Vision 2015* (2015)
15. Fazekas, S., Amiaz, T., Chetverikov, D., Kiryati, N.: Dynamic texture detection based on motion analysis. *Int. J. Comput. Vis.* **82**(1), 48–63 (2009)
16. Fazekas, S., Chetverikov, D.: Analysis and performance evaluation of optical flow features for dynamic texture recognition. *Signal Process. Image Commun.* **22**(7), 680–691 (2007)
17. Florack, L.M.J.: *Image Structure*. Springer, Berlin (1997)
18. Ghanem, B., Ahuja, N.: Phase based modelling of dynamic textures. In: *Proceedings of International Conference on Computer Vision (ICCV 2007)*, pp. 1–8. IEEE (2007)
19. Ghanem, B., Ahuja, N.: Maximum margin distance learning for dynamic texture recognition. In: *European Conference on Computer Vision (ECCV 2010)*, Springer LNCS, vol. 6312, pp. 223–236 (2010)
20. Gonçalves, W.N., Machado, B.B., Bruno, O.M.: Spatiotemporal Gabor filters: A new method for dynamic texture recognition. *arXiv preprint arXiv:1201.3612* (2012)
21. Gonçalves, W.N., Machado, B.B., Bruno, O.M.: A complex network approach for dynamic texture recognition. *Neurocomputing* **153**, 211–220 (2015)
22. Hernandez, J.A.R., Crowley, J.L., Lux, A., Pietikäinen, M.: Histogram-tensorial gaussian representations and its applications to facial analysis. In: *Local Binary Patterns: New Variants and Applications*, pp. 245–268. Springer (2014)
23. Hong, S., Ryu, J., Im, W., Yang, H.S.: D3: Recognizing dynamic scenes with deep dual descriptor based on key frames and key segments. *Neurocomputing* **273**, 611–621 (2018)
24. Hong, S., Ryu, J., Yang, H.S.: Not all frames are equal: Aggregating salient features for dynamic texture classification. *Multidimens. Syst. Signal Process.* (2016). <https://doi.org/10.1007/s11045-016-0463-7>
25. Hubel, D.H., Wiesel, T.N.: *Brain and Visual Perception: The Story of a 25-Year Collaboration*. Oxford University Press, Oxford (2005)
26. Iijima, T.: Observation theory of two-dimensional visual patterns. Technical report, Papers of Technical Group on Automata and Automatic Control, IECE, Japan (1962) (in Japanese)
27. Jansson, Y., Lindeberg, T.: Dynamic texture recognition using time-causal spatio-temporal scale-space filters. In: *Proceedings of Scale Space and Variational Methods for Computer Vision (SSVM 2017)*, Springer LNCS, vol. 10302, pp. 16–28. Kolding, Denmark (2017)
28. Ji, H., Yang, X., Ling, H., Xu, Y.: Wavelet domain multifractal analysis for static and dynamic texture classification. *IEEE Trans. Image Process.* **22**(1), 286–299 (2013)
29. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: *Proceedings of British Machine Vision Conference Leeds, UK* (2008)
30. Koenderink, J.J.: The structure of images. *Biol. Cybernet.* **50**, 363–370 (1984)
31. Koenderink, J.J.: Scale-time. *Biol. Cybernet.* **58**, 159–162 (1988)
32. Koenderink, J.J., van Doorn, A.J.: Representation of local geometry in the visual system. *Biol. Cybernet.* **55**, 367–375 (1987)
33. Koenderink, J.J., van Doorn, A.J.: Generic neighborhood operators. *IEEE Trans. Pattern Anal. Mach. Intell.* **14**(6), 597–605 (1992)
34. 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)
35. Laptev, I., Lindeberg, T.: Velocity adaptation of spatio-temporal receptive fields for direct recognition of activities: an experimental study. *Image Vis. Comput.* **22**(2), 105–116 (2004)
36. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *Proceedings of Computer Vision and Pattern Recognition (CVPR'06)*, pp. 2169–2178. Washington, DC, USA (2006)
37. Linde, O., Lindeberg, T.: Object recognition using composed receptive field histograms of higher dimensionality. In: *International conference on pattern recognition*, vol. 2, pp. 1–6. Cambridge (2004)
38. Linde, O., Lindeberg, T.: Composed complex-cue histograms: an investigation of the information content in receptive field based image descriptors for object recognition. *Comput. Vis. Image Underst.* **116**, 538–560 (2012)
39. Lindeberg, T.: *Scale-Space Theory in Computer Vision*. Springer, Berlin (1993)
40. Lindeberg, T.: Feature detection with automatic scale selection. *Int. J. Comput. Vis.* **30**(2), 77–116 (1998)
41. 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)
42. Lindeberg, T.: A computational theory of visual receptive fields. *Biol. Cybernet.* **107**(6), 589–635 (2013)
43. Lindeberg, T.: Invariance of visual operators at the level of receptive fields. *PLOS ONE* **8**(7), e66,990 (2013)
44. Lindeberg, T.: Time-causal and time-recursive spatio-temporal receptive fields. *J. Math. Imaging Vis.* **55**(1), 50–88 (2016)
45. Lindeberg, T.: Spatio-temporal scale selection in video data. *J. Math. Imaging Vis.* (2017). <https://doi.org/10.1007/s10851-017-0766-9>
46. Lindeberg, T.: Spatio-temporal scale selection in video data. In: *Proceedings of Scale Space and Variational Methods in Computer Vision (SSVM 2017)*, Springer LNCS, vol. 10302, pp. 3–15 (2017)

47. Lindeberg, T.: Dense scale selection over space, time and space-time. *SIAM J. Imaging Sci.* **11**(1), 407–441 (2018)
48. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
49. Lu, Z., Xie, W., Pei, J., Huang, J.: Dynamic texture recognition by spatio-temporal multiresolution histograms. In: *Proceedings of Seventh IEEE Workshop Motion and Video Computing*, vol. 2, pp. 241–246. IEEE (2005)
50. Miao, J., Xu, X., Xing, X., Tao, D.: Manifold regularized slow feature analysis for dynamic texture recognition. *arXiv preprint arXiv:1706.03015* (2017)
51. Mumtaz, A., Coviello, E., Lanckriet, G.R., Chan, A.B.: A scalable and accurate descriptor for dynamic textures using bag of system trees. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(4), 697–712 (2015)
52. Nelson, R.C., Polana, R.: Qualitative recognition of motion using temporal texture. *CVGIP Image Underst.* **56**(1), 78–89 (1992)
53. Norouznezhad, E., Harandi, M.T., Bigdeli, A., Baktash, M., Postula, A., Lovell, B.C.: Directional space-time oriented gradients for 3D visual pattern analysis. In: *European Conference on Computer Vision*. Springer LNCS, vol. 7574, pp. 736–749 (2012)
54. Péteri, R., Chetverikov, D.: Dynamic texture recognition using normal flow and texture regularity. In: *Pattern Recognition and Image Analysis*, pp. 9–23 (2005)
55. Péteri, R., Fazekas, S., Huiskes, M.J.: DynTex: a comprehensive database of dynamic textures. *Pattern Recognit. Lett.* **31**(12), 1627–1632 (2010)
56. Qi, X., Li, C., Guoying, Z., Hong, X., Pietikäinen, M.: Dynamic texture and scene classification by transferring deep image features. *Neurocomputing* **171**, 1230–1241 (2016)
57. Qiao, Y., Weng, L.: Hidden markov model based dynamic texture classification. *IEEE Signal Process. Lett.* **22**(4), 509–512 (2015)
58. Quan, Y., Bao, C., Ji, H.: Equiangular kernel dictionary learning with applications to dynamic texture analysis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 308–316 (2016)
59. Quan, Y., Huang, Y., Ji, H.: Dynamic texture recognition via orthogonal tensor dictionary learning. In: *IEEE International Conference on Computer Vision*, pp. 73–81 (2015)
60. Ravichandran, A., Chaudhry, R., Vidal, R.: View-invariant dynamic texture recognition using a bag of dynamical systems. In: *Computer Vision and Pattern Recognition*, pp. 1651–1657 (2009)
61. Ren, J., Jiang, X., Yuan, J.: Dynamic texture recognition using enhanced LBP features. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2400–2404 (2013)
62. Rivera, A.R., Chae, O.: Spatiotemporal directional number transitional graph for dynamic texture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(10), 2146–2152 (2015)
63. Sagel, A., Kleinstueber, M.: Alignment distances on systems of bags. *IEEE Trans. Circuits Syst. Video Technol.* (2017). <https://doi.org/10.1109/TCSVT.2017.2715851>
64. Schiele, B., Crowley, J.: Recognition without correspondence using multidimensional receptive field histograms. *Int. J. Comput. Vis.* **36**(1), 31–50 (2000)
65. Smith, J.R., Lin, C.Y., Naphade, M.: Video texture indexing using spatio-temporal wavelets. In: *Proceedings of IEEE International Conference on Image Processing*, vol. 2, pp. 437–440 (2002)
66. Soatto, S., Doretto, G., Wu, Y.N.: Dynamic textures. *IEEE Int. Conf. Comput. Vis.* **2**, 439–446 (2001)
67. Sporring, J., Nielsen, M., Florack, L., Johansen, P. (eds.): *Gaussian Scale-Space Theory: Proc. PhD School on Scale-Space Theory. Series in Mathematical Imaging and Vision*. Springer, Copenhagen (1997)
68. Swain, M.J., Ballard, D.H.: Color indexing. *Int. J. Comput. Vis.* **7**(1), 11–32 (1991)
69. ter Haar Romeny, B.: *Front-End Vision and Multi-Scale Image Analysis*. Springer, Berlin (2003)
70. 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, vol. 2106. Vancouver, Canada (2001)
71. Tiwari, D., Tyagi, V.: Dynamic texture recognition based on completed volume local binary pattern. *Multidimens. Syst. Signal Process.* **27**(2), 563–575 (2016)
72. Tiwari, D., Tyagi, V.: Dynamic texture recognition using multiresolution edge-weighted local structure pattern. *Comput. Electr. Eng.* (2016). <https://doi.org/10.1016/j.compeleceng.2016.11.008>
73. 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)
74. Valois, R.L.D., Cottaris, N.P., Mahon, L.E., Elfer, S.D., Wilson, J.A.: Spatial and temporal receptive fields of geniculate and cortical cells and directional selectivity. *Vis. Res.* **40**(2), 3685–3702 (2000)
75. Wang, L., Liu, H., Sun, F.: Dynamic texture video classification using extreme learning machine. *Neurocomputing* **174**, 278–285 (2016)
76. Wang, Y., Hu, S.: Exploiting high level feature for dynamic textures recognition. *Neurocomputing* **154**, 217–224 (2015)
77. Wang, Y., Hu, S.: Chaotic features for dynamic textures recognition. *Soft Comput.* **20**(5), 1977–1989 (2016)
78. Weickert, J., Ishikawa, S., Imiya, A.: Linear scale-space has first been proposed in Japan. *J. Math. Imaging Vis.* **10**(3), 237–252 (1999)
79. Wildes, R.P., Bergen, J.R.: Qualitative spatiotemporal analysis using an oriented energy representation. *European Conference on Computer Vision*, Springer LNCS, vol. 1843, pp. 768–784 (2000)
80. Witkin, A.P.: Scale-space filtering. In: *Proceedings of 8th International Joint Conference on Artificial Intelligence*, pp. 1019–1022. Karlsruhe, Germany (1983)
81. Woolfe, F., Fitzgibbon, A.: Shift-invariant dynamic texture recognition. *Eur. Conf. Comput. Vis.* **3952**, 549–562 (2006)
82. Xu, Y., Huang, S., Ji, H., Fermüller, C.: Scale-space texture description on SIFT-like textons. *Comput. Vis. Image Underst.* **116**(9), 999–1013 (2012)
83. Xu, Y., Quan, Y., Zhang, Z., Ling, H., Ji, H.: Classifying dynamic textures via spatiotemporal fractal analysis. *Pattern Recognit.* **48**(10), 3239–3248 (2015)
84. Yang, F., Xia, G.S., Liu, G., Zhang, L., Huang, X.: Dynamic texture recognition by aggregating spatial and temporal features via ensemble SVMs. *Neurocomputing* **173**, 1310–1321 (2016)
85. 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)
86. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using volume local binary patterns. In: *Proceedings of the Workshop on Dynamical Vision WDV*, Springer LNCS, vol. 4358, pp. 165–177 (2006)
87. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 915–928 (2007)



Ylva Jansson obtained her MSc in Engineering Physics from KTH Royal Institute of Technology, Stockholm, Sweden, in 2015. She is currently a second-year PhD student at the Computational Brain Science Lab at the Department of Computational Science and Technology at KTH Royal Institute of Technology. Her current research interests include video analysis and hybrid approaches for spatial and spatio-temporal recognition in the intersection between scale-space methods, deep learning and bio-

logical vision. She also has an interest in interpretability and visualisation for machine learning methods.



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 MSc degree in 1987 and his PhD 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, video analysis, 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*.