On Space-Time Interest Points*

Ivan Laptev and Tony Lindeberg

Computational Vision and Active Perception Laboratory (CVAP),
Department of Numerical Analysis and Computer Science,
KTH, SE-100 44 Stockholm, Sweden

Email: {laptev, tony}@nada.kth.se

Technical report ISRN KTH/NA/P–03/12–SE

Shortened version in Proc. ICCV'03
Nice, France, October 2003, pp. 432–439.

Earlier version also in Proc. Scale Space '03
Isle of Skye, Scotland, June 2003, pp. 372–387.

Abstract

Local image features or interest points provide compact and abstract representations of patterns in an image. In this paper, we extend the notion of spatial interest points into the spatio-temporal domain and show how the resulting features capture interesting events in video and can be used for a compact representation and for interpretation of video data.

To detect spatio-temporal events, we build on the idea of the Harris and Förstner interest point operators and detect local structures in space-time where the image values have significant local variations in both space and time. We estimate the spatio-temporal extents of the detected events by maximizing a normalized spatio-temporal Laplacian operator over spatial and temporal scales. To represent the detected events, we then compute local, spatio-temporal, scale-invariant N-jets and classify each event with respect to its jet descriptor. For the problem of human motion analysis, we illustrate how a video representation in terms of local space-time features allows for detection of walking people in scenes with occlusions and dynamic cluttered backgrounds.

*The support from the Swedish Research Council and from the Royal Swedish Academy of Sciences as well as the Knut and Alice Wallenberg Foundation is gratefully acknowledged.

1 Introduction

Analyzing and interpreting video is a growing topic in computer vision and its applications. Video data contains information about changes in the environment and is highly important for many visual tasks including navigation, surveillance and video indexing.

Traditional approaches for motion analysis mainly involve the computation of optic flow (Barron, Fleet and Beauchemin, 1994) or feature tracking (Smith and Brady, 1995; Blake and Isard, 1998). Although very effective for many tasks, both of these techniques have limitations. Optic flow approaches mostly capture first-order motion and may fail when the motion has sudden changes. Interesting solutions to this problem have been proposed (Niyogi, 1995; Fleet, Black and Jepson, 1998; Hoey and Little, 2000). Feature trackers often assume a constant appearance of image patches over time and may hence fail when the appearance changes, for example, in situations when two objects in the image merge or split. Model-based solutions for this problem have been presented by (Black and Jepson, 1998).

Image structures in video are not restricted to constant velocity and/or constant appearance over time. On the contrary, many interesting events in video are characterized by strong variations in the data along both the spatial and the temporal dimensions. For example, consider a scene with a person entering a room, applauding hand gestures, a car crash or a water splash; see also the illustrations in figure 1.

More generally, points with non-constant motion correspond to accelerating local image structures that may correspond to accelerating objects in the world. Hence, such points can be expected to contain information about the forces acting in the physical environment and changing its structure.

In the spatial domain, points with a significant local variation of image intensities have been extensively investigated in the past (Förstner and Gülch, 1987; Harris and Stephens, 1988; Lindeberg, 1998; Schmid, Mohr and Bauckhage, 2000). Such image points are

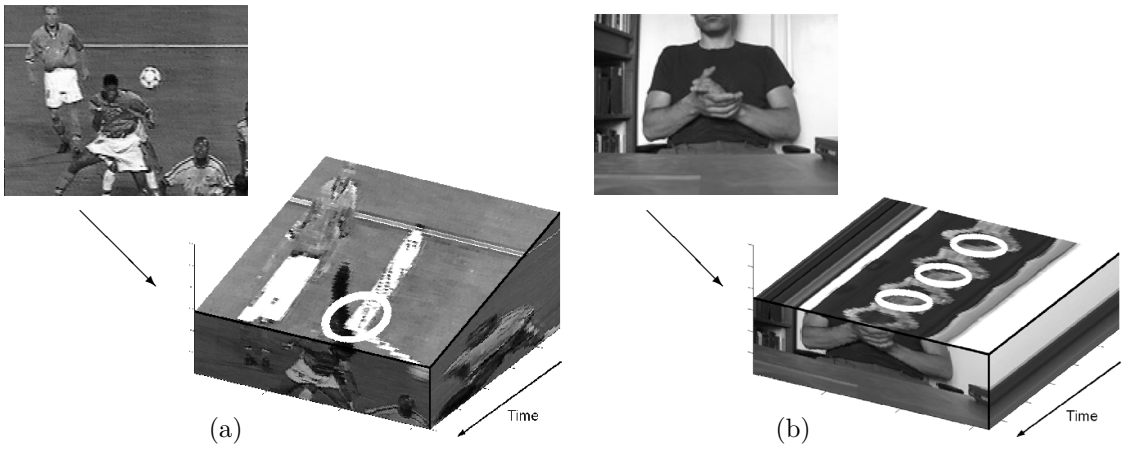


Figure 1: Result of detecting the strongest spatio-temporal interest points in a football sequence with a player heading the ball (a) and in a hand clapping sequence (b). From the temporal slices of space-time volumes shown here, it is evident that the detected events correspond to neighborhoods with high spatio-temporal variation in the image data or "space-time corners".

frequently referred to as "interest points" and are attractive due to their high information contents and relative stability with respect to perspective transformations of the data. Highly successful applications of interest points have been presented for image indexing (Schmid and Mohr, 1997), stereo matching (Tuytelaars and Van Gool, 2000; Mikolajczyk and Schmid, 2002; Tell and Carlsson, 2002), optic flow estimation and tracking (Smith and Brady, 1995; Bretzner and Lindeberg, 1998), and object recognition (Lowe, 1999; Hall, de Verdiere and Crowley, 2000; Fergus, Perona and Zisserman, 2003; Wallraven, Caputo and Graf, 2003).

In this paper, we extend the notion of interest points into the spatio-temporal domain and show that the resulting local space-time features often correspond to interesting events in video data (see figure 1). In particular, we aim at a direct scheme for event detection and interpretation that does not require feature tracking nor computation of optic flow. In the considered sample application we show that the proposed type of features can be used for sparse coding of video information that in turn can be used for interpreting video scenes such as human motion in situations with complex and non-stationary background.

To detect spatio-temporal interest points, we build on the idea of the Harris and Förstner interest point operators (Harris and Stephens, 1988; Förstner and Gülch, 1987) and describe the detection method in section 2. As events often have characteristic extents in both space and time (Koenderink, 1988; Lindeberg and Fagerström, 1996; Florack, 1997; Lindeberg, 1997; Chomat, Martin and Crowley, 2000b; Zelnik-Manor and Irani, 2001), we investigate the behavior of interest points in spatio-temporal scale-space and adapt both the spatial and the temporal scales of the detected features in section 3. In section 4, we show how the neighborhoods of interest points can be described in terms of spatio-temporal derivatives and then be used to distinguish different events in video. In section 5, we consider a video representation in terms of classified spatio-temporal interest points and demonstrate how this representation can be efficient for the task of video registration. In particular, we present an approach for detecting walking people in complex scenes with occlusions and dynamic background. Finally, section 6 concludes the paper with the summary and discussion.

2 Spatio-temporal interest point detection

2.1 Interest points in the spatial domain

In the spatial domain, we can model an image $f^{sp} : \mathbb{R}^2 \mapsto \mathbb{R}$ by its linear scale-space representation (Witkin, 1983; Koenderink and van Doorn, 1992; Lindeberg, 1994; Florack, 1997) $L^t \colon \mathbb{R}^2 \times \mathbb{R}_+ \mapsto \mathbb{R}$

$$L^{sp}(x, y; \sigma_l^2) = g^{sp}(x, y; \sigma_l^2) * f^{sp}(x, y), \tag{1}$$

defined by the convolution of f^{sp} with Gaussian kernels of variance σ_l^2

$$g^{sp}(x, y; \sigma_l^2) = \frac{1}{2\pi\sigma_l^2} \exp(-(x^2 + y^2)/2\sigma_l^2). \tag{2}$$

The idea of the Harris interest point detector is to find spatial locations where f^{sp} has significant changes in both directions. For a given scale of observation σ_l^2, such points can be found using a second moment matrix integrated over a Gaussian window with

variance σ_i^2 (Förstner and Gülch, 1987; Bigün, Granlund and Wiklund, 1991; Lindeberg and Garding, 1997):

$$\begin{aligned} \mu^{sp}(\cdot;\, \sigma_l^2, \sigma_i^2) &= g^{sp}(\cdot;\, \sigma_i^2) * \big((\nabla L(\cdot;\, \sigma_l^2))(\nabla L(\cdot;\, \sigma_l^2))^T\big) \\ &= g^{sp}(\cdot;\, \sigma_i^2) * \begin{pmatrix} (L_x^{sp})^2 & L_x^{sp} L_y^{sp} \\ L_x^{sp} L_y^{sp} & (L_y^{sp})^2 \end{pmatrix} \end{aligned} \tag{3}$$

where $'*'$ denotes the convolution operator, and L_x^{sp} and L_y^{sp} are Gaussian derivatives computed at local scale σ_l^2 according to $L_x^{sp} = \partial_x(g^{sp}(\cdot;\, \sigma_l^2) * f^{sp}(\cdot))$ and $L_y^{sp} = \partial_y(g^{sp}(\cdot;\, \sigma_l^2) * f^{sp}(\cdot))$. The second moment descriptor can be thought of as the covariance matrix of a two-dimensional distribution of image orientations in the local neighborhood of a point. Hence, the eigenvalues λ_1, λ_2, $(\lambda_1 \leq \lambda_2)$ of μ^{sp} constitute descriptors of variations in f^{sp} along the two image directions. Specifically, two significantly large values of λ_1, λ_2 indicate the presence of an interest point. To detect such points, Harris and Stephens (1988) proposed to detect positive maxima of the corner function

$$H^{sp} = \det(\mu^{sp}) - k\, \mathrm{trace}^2(\mu^{sp}) = \lambda_1 \lambda_2 - k(\lambda_1 + \lambda_2)^2. \tag{4}$$

At the positions of the interest points, the ratio of the eigenvalues $\alpha = \lambda_2/\lambda_1$ has to be high. From (4) it follows that for positive local maxima of H^{sp}, the ratio α has to satisfy $k \leq \alpha/(1+\alpha)^2$. Hence, if we set $k = 0.25$, the positive maxima of H will only correspond to ideally isotropic interest points with $\alpha = 1$, i.e. $\lambda_1 = \lambda_2$. Lower values of k allow us to detect interest points with more elongated shape, corresponding to higher values of α. A commonly used value of k in the literature is $k = 0.04$ corresponding to the detection of points with $\alpha < 23$.

The result of detecting Harris interest points in an outdoor image sequence of a walking person is presented at the bottom row of figure 8.

2.2 Interest points in the spatio-temporal domain

In this section, we develop an operator that responds to events in temporal image sequences at specific locations and with specific extents in space-time. The idea is to extend the notion of interest points in the spatial domain by requiring the image values in local spatio-temporal volumes to have large variations along both the spatial and the temporal directions. Points with such properties will correspond to spatial interest points with a distinct location in time corresponding to a local spatio-temporal neighborhood with non-constant motion.

To model a spatio-temporal image sequence, we use a function $f\colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ and construct its linear scale-space representation $L\colon \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}_+^2 \mapsto \mathbb{R}$ by convolution of f with an anisotropic Gaussian kernel[1] with distinct spatial variance σ_l^2 and temporal variance τ_l^2

$$L(\cdot;\, \sigma_l^2, \tau_l^2) = g(\cdot;\, \sigma_l^2, \tau_l^2) * f(\cdot), \tag{5}$$

[1] In general, convolution with a Gaussian kernel in the temporal domain violates causality constraints, since the temporal image data is available only for the past. For real-time implementation, time-causal scale-space filters thus have to be used (Koenderink, 1988; Lindeberg and Fagerström, 1996; Florack, 1997; Lindeberg, 2002). In this paper, however, we simplify this part of the investigation and assume that the data is available for a sufficiently long period of time and that the image sequence can be convolved with a Gaussian kernel over both space and time.

where the spatio-temporal separable Gaussian kernel is defined as

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}} \exp(-(x^2 + y^2)/2\sigma_l^2 - t^2/2\tau_l^2). \tag{6}$$

Using a separate scale parameter for the temporal domain is essential, since the spatial and the temporal extents of events are in general independent. Moreover, as will be illustrated in section 2.3, events detected using our interest point operator depend on both the spatial and the temporal scales of observation and, hence, require separate treatment of the corresponding scale parameters σ_l^2 and τ_l^2.

Similar to the spatial domain, we consider a spatio-temporal second-moment matrix, which is a 3-by-3 matrix composed of first order spatial and temporal derivatives averaged using a Gaussian weighting function $g(\cdot; \sigma_i^2, \tau_i^2)$

$$\mu = g(\cdot; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}, \tag{7}$$

where we here relate the integration scales σ_i^2 and τ_i^2 to the local scales σ_l^2 and τ_l^2 according to $\sigma_i^2 = s\sigma_l^2$ and $\tau_i^2 = s\tau_l^2$. The first-order derivatives are defined as

$$L_x(\cdot; \sigma_l^2, \tau_l^2) = \partial_x(g * f), \quad L_y(\cdot; \sigma_l^2, \tau_l^2) = \partial_y(g * f), \quad L_t(\cdot; \sigma_l^2, \tau_l^2) = \partial_t(g * f).$$

To detect interest points, we search for regions in f having significant eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of μ. Among different approaches to find such regions, we propose here to extend the Harris corner function (4) defined for the spatial domain into the spatio-temporal domain by combining the determinant and the trace of μ as follows:

$$H = \det(\mu) - k \operatorname{trace}^3(\mu) = \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3. \tag{8}$$

To show how positive local maxima of H correspond to points with high values of $\lambda_1, \lambda_2, \lambda_3$ ($\lambda_1 \leq \lambda_2 \leq \lambda_3$), we define the ratios $\alpha = \lambda_2/\lambda_1$ and $\beta = \lambda_3/\lambda_1$ and re-write H as

$$H = \lambda_1^3(\alpha\beta - k(1 + \alpha + \beta)^3).$$

From the requirement $H \geq 0$, we get $k \leq \alpha\beta/(1 + \alpha + \beta)^3$ and it follows that k assumes its maximum possible value $k = 1/27$ when $\alpha = \beta = 1$. For sufficiently large values of k, positive local maxima of H correspond to points with high variation in the image values along both the spatial and the temporal directions. In particular, if we set the maximum value of α, β to 23 as in the spatial domain, the value of k to be used in H (8) will then be $k \approx 0.005$. Thus, spatio-temporal interest points of f can be found by detecting local positive spatio-temporal maxima in H.

2.3 Experimental results for synthetic data

In this section, we illustrate the detection of spatio-temporal interest points on synthetic image sequences. For clarity of presentation, we show the spatio-temporal data as 3-D space-time plots, where the original signal is represented by a threshold surface, while the detected interest points are illustrated by ellipsoids with positions corresponding to the

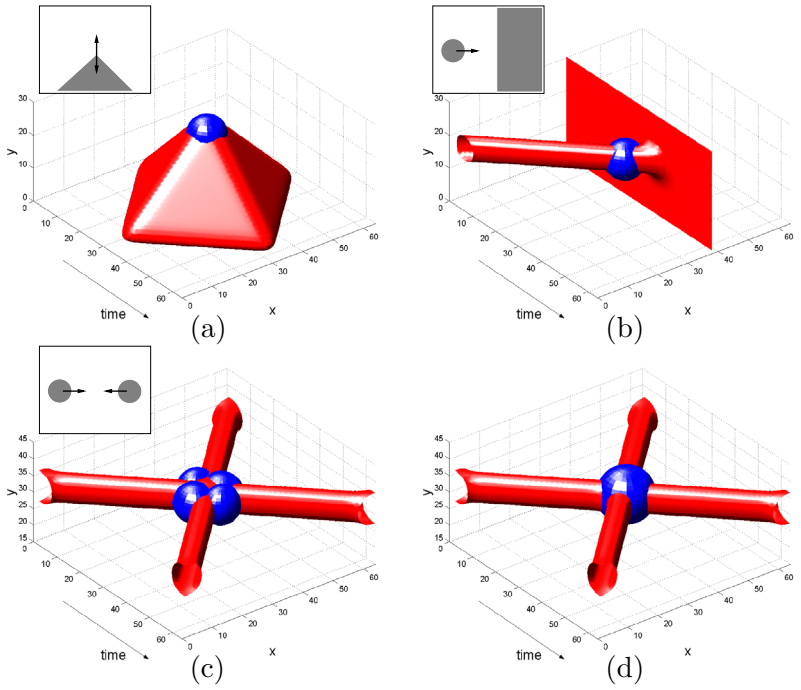


Figure 2: Results of detecting spatio-temporal interest points on synthetic image sequences. (a): A moving corner; (b) A merge of a ball and a wall; (c): Collision of two balls with interest points detected at scales $\sigma_l^2 = 8$ and $\tau_l^2 = 8$; (d): the same signal as in (c) but with the interest points detected at coarser scales $\sigma_l^2 = 16$ and $\tau_l^2 = 16$.

space-time location of the interest point and the length of the semi-axes proportional to the local scale parameters σ_l and τ_l used in the computation of H.

Figure 2a shows a sequence with a moving corner. The interest point is detected at the moment in time when the motion of the corner changes direction. This type of event occurs frequently in natural sequences, such as sequences of articulated motion. Note that according to the definition of spatio-temporal interest points, image structures with constant motion do not give rise to responses of the detector. Other typical types of events that can be detected by the proposed method are splits and unifications of image structures. In figure 2b, the interest point is detected at the moment and the position corresponding to the collision of a ball and a wall. Similarly, interest points are detected at the moment of collision and bouncing of two balls as shown in figure 2c-d. Note, that different types of events are detected depending on the scale of observation.

To further emphasize the importance of the spatial and the temporal scales of observation, let us consider an oscillating signal with different spatial and temporal frequencies defined by $f(x, y, t) = \mathrm{sgn}(y - \sin(x^4)\sin(t^4))$, where $\mathrm{sgn}(u) = 1$ for $u > 0$ and $\mathrm{sgn}(u) = -1$ for $u < 0$ (see figure 3). As can be seen from the illustration, the result of detecting the strongest interest points highly depends on the choice of the scale parameters σ_l^2 and τ_l^2. We can observe that space-time structures with long temporal extents are detected for large values of τ_l^2 while short events are preferred by the detector with small values of

τ_l^2. Similarly, the spatial extent of the events is related to the value of the spatial scale parameter σ_l^2.

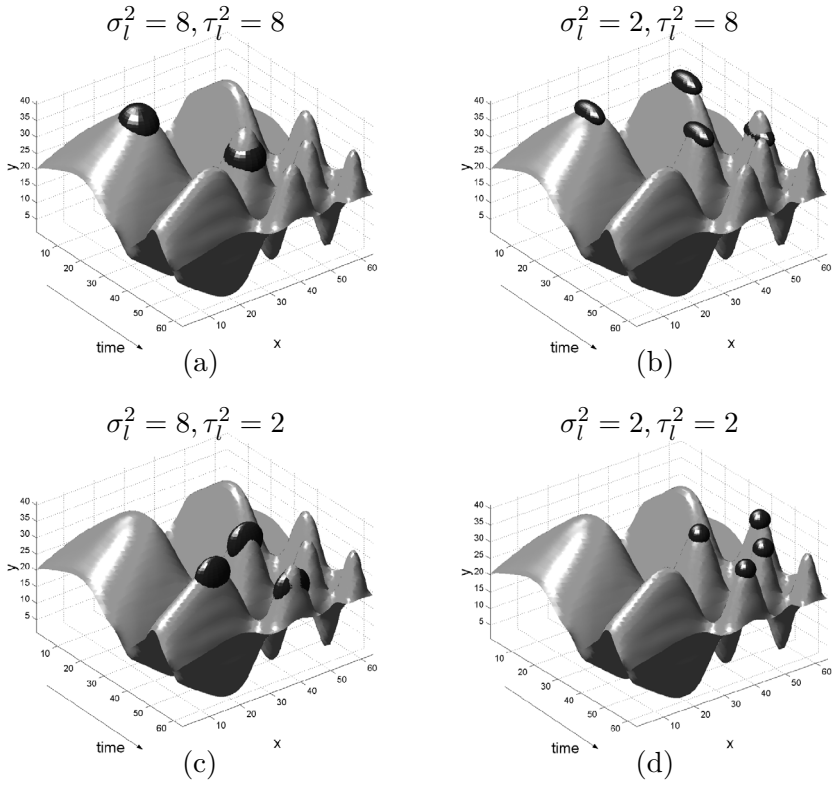


Figure 3: Results of detecting interest point at different spatial and temporal scales for a synthetic sequence with impulses having varying extents in space and time. The extents of the detected events roughly corresponds to the scale parameters σ_l^2 and τ_l^2 used for computing H (8).

From the presented examples, we can conclude that a correct selection of temporal and spatial scales is crucial when capturing events with different spatial and temporal extents. Moreover, estimating the spatio-temporal extents of events is important for their further interpretation. In the next section, we will present a mechanism for simultaneous estimation of spatio-temporal scales. This mechanism will then be combined with the interest point detector resulting in scale-adapted interest points in section 3.2.

3 Spatio-temporal scale adaptation

3.1 Scale selection in space-time

During recent years, the problem of automatic scale selection has been addressed in several different ways, based on the maximization of normalized derivative expressions over scale, or the behavior of entropy measures or error measures over scales (see Lindeberg and Bretzner (2003) for a review). To estimate the spatio-temporal extent of an event in

space-time, we follow works on local scale selection proposed in the spatial domain by Lindeberg (1998) as well as in the temporal domain (Lindeberg, 1997). The idea is to define a differential operator that assumes simultaneous extrema over spatial and temporal scales that are characteristic for an event with a particular spatio-temporal location.

For the purpose of analysis, we will first study a prototype event represented by a spatio-temporal Gaussian blob

$$f(x, y, t;\, \sigma_0^2, \tau_0^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}} \exp(-(x^2 + y^2)/2\sigma_0^2 - t^2/2\tau_0^2)$$

with spatial variance σ_0^2 and temporal variance τ_0^2 (see figure 4a). Using the semi-group property of the Gaussian kernel, it follows that the scale-space representation of f is

$$L(\cdot;\, \sigma^2, \tau^2) = g(\cdot;\, \sigma^2, \tau^2) * f(\cdot;\, \sigma_0^2, \tau_0^2) = g(\cdot;\, \sigma_0^2 + \sigma^2, \tau_0^2 + \tau^2).$$

To recover the spatio-temporal extent (σ_0, τ_0) of f, we consider second-order derivatives of L normalized by the scale parameters as follows

$$L_{xx,norm} = \sigma^{2a} \tau^{2b} L_{xx}, \quad L_{yy,norm} = \sigma^{2a} \tau^{2b} L_{yy}, \quad L_{tt,norm} = \sigma^{2c} \tau^{2d} L_{tt}. \tag{9}$$

All of these entities assume local extrema over space and time at the center of the blob f. Moreover, depending on the parameters a, b and c, d, they also assume local extrema over scales at certain spatial and temporal scales, $\tilde{\sigma}^2$ and $\tilde{\tau}^2$.

The idea of scale selection we follow here is to determine the parameters a, b, c, d such that $L_{xx,norm}$, $L_{yy,norm}$ and $L_{tt,norm}$ assume extrema at scales $\tilde{\sigma}^2 = \sigma_0^2$ and $\tilde{\tau}^2 = \tau_0^2$. To find such extrema, we differentiate the expressions in (9) with respect to the spatial and the temporal scale parameters. For the spatial derivatives we obtain the following expressions at the center of the blob

$$\frac{\partial}{\partial \sigma^2} \Big[L_{xx,norm}(0, 0, 0;\, \sigma^2, \tau^2) \Big] = -\frac{a\sigma^2 - 2\sigma^2 + a\sigma_0^2}{\sqrt{(2\pi)^3 (\sigma_0^2 + \sigma^2)^6 (\tau_0^2 + \tau^2)}}\, \sigma^{2(a-1)} \tau^{2b} \tag{10}$$

$$\frac{\partial}{\partial \tau^2} \Big[L_{xx,norm}(0, 0, 0;\, \sigma^2, \tau^2) \Big] = -\frac{2b\tau_0^2 + 2b\tau^2 - \tau^2}{\sqrt{2^5 \pi^3 (\sigma_0^2 + \sigma^2)^4 (\tau_0^2 + \tau^2)^3}}\, \tau^{2(b-1)} \sigma^{2a}. \tag{11}$$

By setting these expressions to zero, we obtain the following simple relations for a and b

$$a\sigma^2 - 2\sigma^2 + a\sigma_0^2 = 0, \quad 2b\tau_0^2 + 2b\tau^2 - \tau^2 = 0$$

which after substituting $\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$ lead to $a = 1$ and $b = 1/4$. Similarly, differentiating the second-order temporal derivative

$$\frac{\partial}{\partial \sigma^2} \Big[L_{tt,norm}(0, 0, 0;\, \sigma^2, \tau^2) \Big] = -\frac{c\sigma^2 - \sigma^2 + c\sigma_0^2}{\sqrt{(2\pi)^3 (\sigma_0^2 + \sigma^2)^4 (\tau_0^2 + \tau^2)^3}}\, \sigma^{2(c-1)} \tau^{2d} \tag{12}$$

$$\frac{\partial}{\partial \tau^2} \Big[L_{tt,norm}(0, 0, 0;\, \sigma^2, \tau^2) \Big] = -\frac{2d\tau_0^2 + 2d\tau^2 - 3\tau^2}{\sqrt{2^5 \pi^3 (\sigma_0^2 + \sigma^2)^2 (\tau_0^2 + \tau^2)^5}}\, \tau^{2(d-1)} \sigma^{2c} \tag{13}$$

leads to the expressions

$$c\sigma^2 - 2\sigma^2 + c\sigma_0^2 = 0, \quad 2d\tau_0^2 + 2d\tau^2 - \tau^2 = 0$$

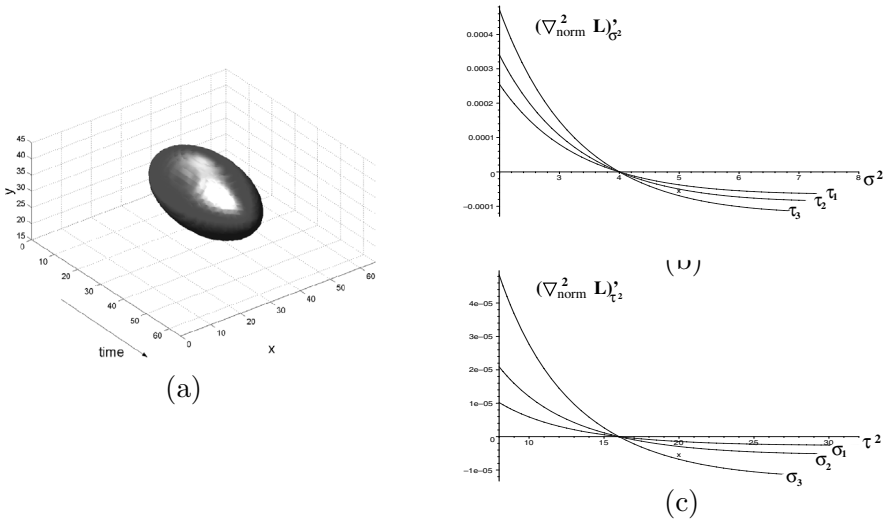


Figure 4: (a): A Spatio-temporal Gaussian blob with spatial variance $\sigma_0^2 = 4$ and temporal variance $\tau_0^2 = 16$; (b)-(c) derivatives of $\nabla^2_{norm} L$ with respect to scales. The zero-crossings of $(\nabla^2_{norm} L)'_{\sigma^2}$ and $(\nabla^2_{norm} L)'_{\tau^2}$ indicate extrema of $\nabla^2_{norm} L$ at scales corresponding to the spatial and the temporal extents of the blob.

which after substituting $\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$ result in $c = 1/2$ and $d = 3/4$.

The normalization of derivatives in (9) guarantees that all these partial derivative expressions assume local space-time-scale extrema at the center of the blob f and at scales corresponding to the spatial and the temporal extents of f, i.e. $\sigma = \sigma_0$ and $\tau = \tau_0$. From the sum of these partial derivatives, we then define a normalized spatio-temporal Laplace operator according to

$$\begin{aligned} \nabla^2_{norm} L &= L_{xx,norm} + L_{yy,norm} + L_{tt,norm} \\ &= \sigma^2 \tau^{1/2} (L_{xx} + L_{yy}) + \sigma \tau^{3/2} L_{tt}. \end{aligned} \tag{14}$$

Figures 4b-c show derivatives of this operator with respect to the scale parameters evaluated at the center of a spatio-temporal blob with spatial variance $\sigma_0^2 = 4$ and temporal variance $\tau_0^2 = 16$. The zero-crossings of the curves verify that $\nabla^2_{norm} L$ assumes extrema at the scales $\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$. Hence, the spatio-temporal extent of the Gaussian prototype can be estimated by finding the extrema of $\nabla^2_{norm} L$ over both spatial and temporal scales. In the following section, we will use this operator for estimating the extents of other spatio-temporal structures, in analogy with previous work of using the normalized Laplacian operator as a general tool for estimating the spatial extent of image structures in the spatial domain.

3.2 Scale-adapted space-time interest points

Local scale estimation using the normalized Laplace operator has shown to be very useful in the spatial domain (Lindeberg, 1998; Almansa and Lindeberg, 2000; Chomat, de Verdiere, Hall and Crowley, 2000a). In particular, Mikolajczyk and Schmid (2001) combined the

Harris interest point operator with the normalized Laplace operator and derived a scale-invariant Harris-Laplace interest point detector. The idea is to find points in scale-space that are both spatial maxima of the Harris function H^{sp} (4) and extrema over scale of the scale-normalized Laplace operator in space.

Here, we extend this idea and detect interest points that are simultaneous maxima of the spatio-temporal corner function H (8) over space and time (x, y, t) as well as extrema of the normalized spatio-temporal Laplace operator $\nabla^2_{norm} L$ (14) over scales (σ^2, τ^2). One way of detecting such points is to compute space-time maxima of H for each spatio-temporal scale level and then to select points that maximize $(\nabla^2_{norm} L)^2$ at the corresponding scale. This approach, however, requires dense sampling over the scale parameters and is therefore computationally expensive.

An alternative we follow here, is to detect interest points for a set of sparsely distributed scale values and then to track these points in the spatio-temporal scale-time-space towards the extrema of $\nabla^2_{norm} L$. We do this by iteratively updating the scale and the position of the interest points by (i) selecting the neighboring spatio-temporal scale that maximizes $(\nabla^2_{norm} L)^2$ and (ii) re-detecting the space-time location of the interest point at a new scale. Thus, instead of performing a simultaneous maximization of H and $\nabla^2_{norm} L$ over five dimensions $(x, y, t, \sigma^2, \tau^2)$, we implement the detection of local maxima by splitting the space-time dimensions (x, y, t) and scale dimensions (σ^2, τ^2) and iteratively optimizing over the subspaces until the convergence has been reached.[2] The corresponding algorithm is presented in figure 5.

The result of scale-adaptation of interest points for the spatio-temporal pattern in figure 3 is shown in figure 6. As can be seen, the chosen scales of the adapted interest points match the spatio-temporal extents of the corresponding structures in the pattern.

It should be noted, however, that the presented algorithm has been developed for processing pre-recorded video sequences. In real-time situations, when using causal scale-space representation based on recursive temporal filters (Lindeberg and Fagerström, 1996; Lindeberg, 2002), only a fixed set of discrete temporal scales is available at any moment. In that case an approximate estimate of temporal scale can still be found by choosing interest points that maximize $(\nabla^2_{norm} L)^2$ in a local neighborhood of the spatio-temporal scale-space; see also (Lindeberg, 1997) for a treatment of automatic scale selection for time-causal scale-spaces.

3.3 Experiments

In this section, we investigate the performance of the proposed scale-adapted spatio-temporal interest point detector applied to real image sequences. In the first example, we consider a sequence of a walking person with non-constant image velocities due to the oscillating motion of the legs. As can be seen in figure 7, the spatio-temporal image pattern gives rise to stable interest points. Note that the detected interest points reflect well-localized events in both space and time, corresponding to specific space-time structures of the leg. From the space-time plot in figure 7(a), we can also observe how the

[2] For the experiments presented in this paper, with image sequences of spatial resolution 160×120 pixels and temporal sampling frequency 25Hz (totally up to 200 frames per sequence), we initialized the detection of interest points using combinations of spatial scales $\sigma_l^2 = [2, 4, 8]$ and temporal scales $\sigma_l^2 = [2, 4, 8]$, while using $s = 2$ for the ratio between the integration and the local scale when computing the second-moment matrix.

1. Detect interest points $p_j = (x_j, y_j, t_j, \sigma_{l,j}^2, \tau_{l,j}^2)$, $j = 1..N$ as maxima of H (8) over space and time using sparsely selected combinations of initial spatial scales $\sigma_l^2 = \sigma_{l,1}^2, .., \sigma_{l,n}^2$ and temporal scales $\tau_l^2 = \tau_{l,1}^2, .., \tau_{l,m}^2$ as well as integration scales $\sigma_i^2 = s\sigma_l^2$ and $\tau_i^2 = s\tau_l^2$.

2. **for** each interest point p_j **do**

3. Compute $\nabla_{norm}^2 L$ at position (x_j, y_j, t_j) and combinations of neighboring scales $(\tilde{\sigma}_{i,j}^2, \tilde{\tau}_{i,j}^2)$ where $\tilde{\sigma}_{i,j}^2 = 2^\delta \sigma_{i,j}^2$, $\tilde{\tau}_{i,j}^2 = 2^\delta \tau_{i,j}^2$, and $\delta = -0.25, 0, 0.25$

5. Choose the combination of integration scales $(\tilde{\sigma}_{i,j}^2, \tilde{\tau}_{i,j}^2)$ that maximizes $(\nabla_{norm}^2 L)^2$

6. **if** $\tilde{\sigma}_{i,j}^2 \neq \sigma_{i,j}^2$ or $\tilde{\tau}_{i,j}^2 \neq \tau_{i,j}^2$

Re-detect interest point $\tilde{p}_j = (\tilde{x}_j, \tilde{y}_j, \tilde{t}_j, \tilde{\sigma}_{l,j}^2, \tilde{\tau}_{l,j}^2)$ using integration scales $\tilde{\sigma}_{i,j}^2 = \tilde{\sigma}_{i,j}^2$, $\tilde{\tau}_{i,j}^2 = \tilde{\tau}_{i,j}^2$, local scales $\tilde{\sigma}_{l,j}^2 = \frac{1}{s}\tilde{\sigma}_{i,j}^2$, $\tilde{\tau}_{l,j}^2 = \frac{1}{s}\tilde{\tau}_{i,j}^2$ and position $(\tilde{x}_j, \tilde{y}_j, \tilde{t}_j)$ that is closest to (x_j, y_j, t_j);
set $p_j := \tilde{p}_j$ and **goto 3**

7. **end**

Figure 5: Algorithm for scale adaption of spatio-temporal interest points.

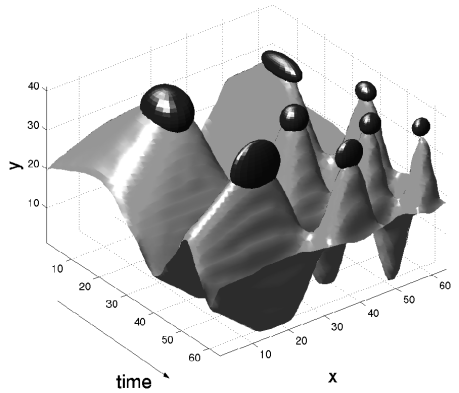
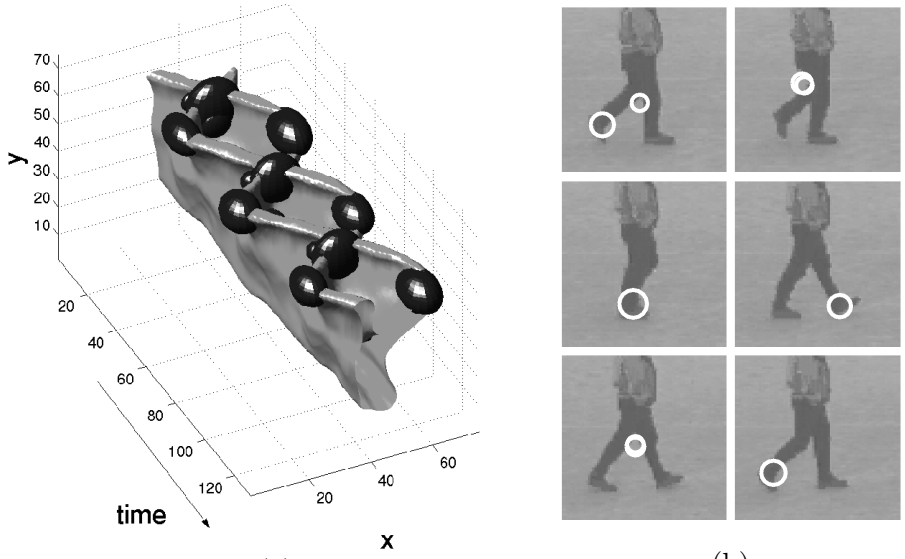


Figure 6: The result of scale-adaptation of spatio-temporal interest points computed from a space-time pattern of the form $f(x, y, t) = \text{sgn}(y - \sin(x^4) * \sin(t^4))$. The interest points are illustrated as ellipsoids showing the selected spatio-temporal scales overlayed on a surface plot of the intensity landscape.



(a) (b)

Figure 7: Results of detecting spatio-temporal interest points from the motion of the legs of a walking person. (a): 3-D plot with a thresholded level surface of a leg pattern (here shown upside down to simplify interpretation) and the detected interest points illustrated by ellipsoids; (b): spatio-temporal interest points overlayed on single frames in the original sequence.

selected spatial and temporal scales of the detected features roughly match the spatio-temporal extents of the corresponding image structures.

The top rows of figure 8 show interest points detected in an outdoor sequence with a walking person and a zooming camera. The changing values of the selected spatial scales (illustrated by the size of the circles) illustrate the invariance of the method with respect to spatial scale changes of the image structures. Note that besides events in the leg pattern, the detector finds spurious points due to the non-constant motion of the coat and the arms. Image structures with constant motion in the background, however, do not result in the response of the detector. The pure spatial interest operator[3] on the contrary gives strong responses in the static background as shown at the bottom row of figure 8

The third example explicitly illustrates how the proposed method is able to estimate the temporal extent of detected events. Figure 9 shows a person making hand-waving gestures with a high frequency on the left and a low frequency on the right. The distinct interest points are detected at the moments and at the spatial positions where the palm of a hand changes its direction of motion. Whereas the spatial scale of the detected interest points remains constant, the selected temporal scale depends on the frequency of the wave pattern. The high frequency pattern results in short events and gives rise to interest points with small temporal extent (see figure 9a). Correspondingly, hand motions with low frequency result in interest points with long temporal extent as shown in figure 9b.

[3]Here, we used the scale-adapted Harris interest point detector (Mikolajczyk and Schmid, 2001) that detects maxima of H^{sp} (4) in space and extrema of normalized Laplacian operator over scales (Lindeberg, 1998).

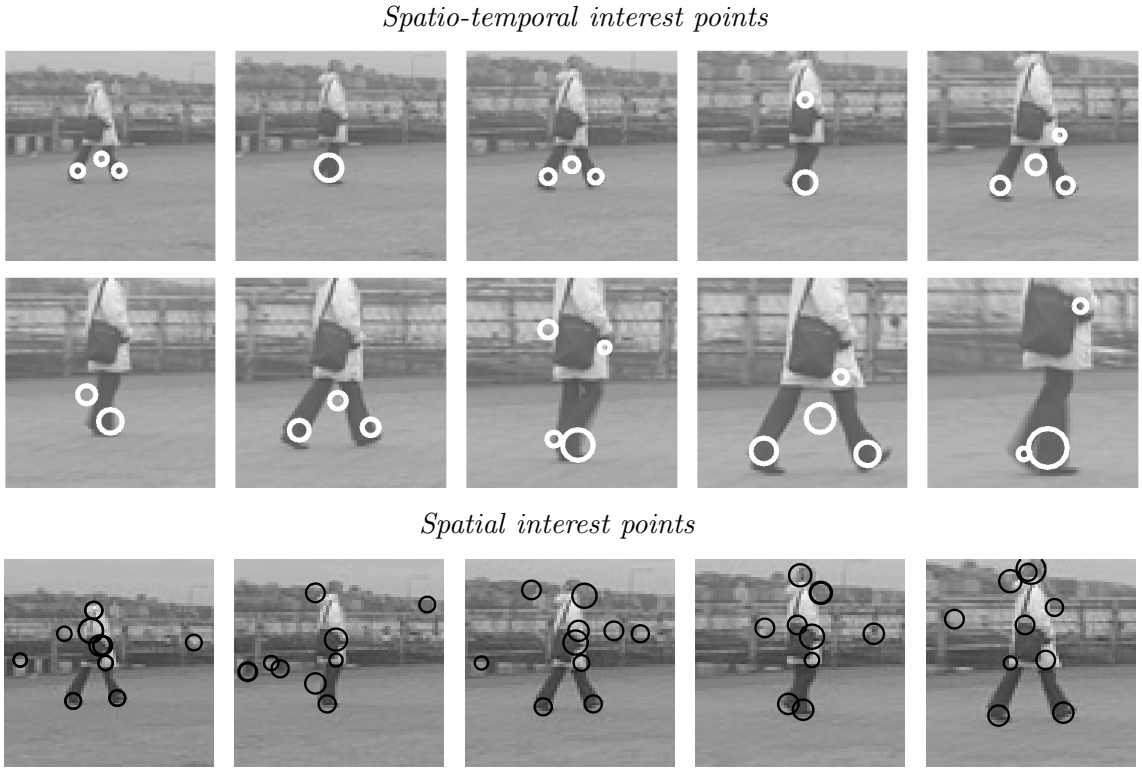


Figure 8: Top: Results of spatio-temporal interest point detection for a zoom-in sequence of a walking person. The spatial scale of the detected points (corresponding to the size of circles) matches the increasing spatial extent of the image structures and verifies the invariance of the interest points with respect to changes in spatial scale. Bottom: Pure spatial interest point detector (here, Harris-Laplace) selects both moving and stationary points and is less restrictive.

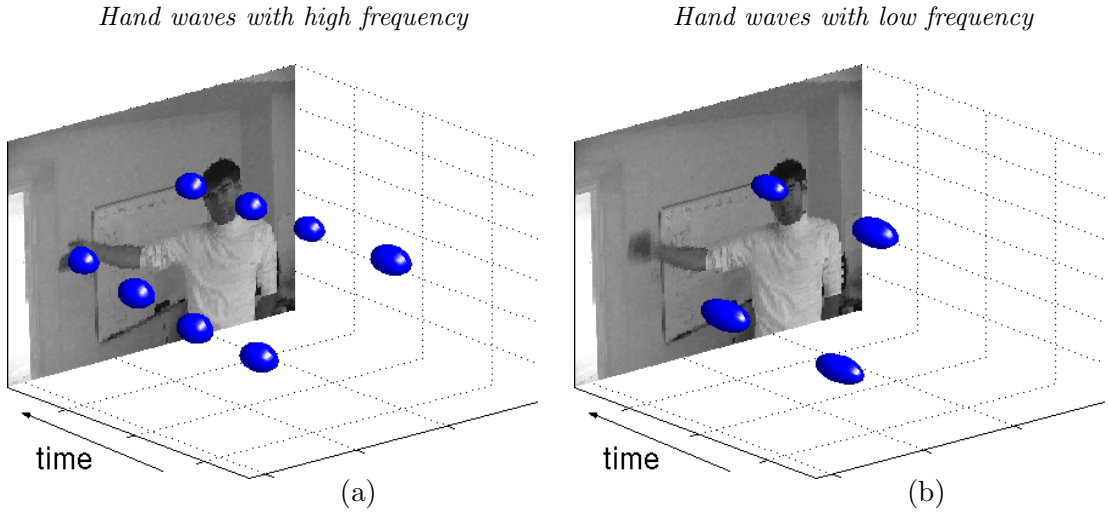


Figure 9: Result of interest point detection for a sequence with waving hand gestures: (a) Interest points for hand movements with high frequency; (b) Interest points for hand movements with low frequency.

4 Classification of events

The detected interest points have significant variations of image values in a local spatio-temporal neighborhood. To differentiate events from each other and from noise, one approach is to compare local neighborhoods and assign points with similar neighborhoods to the same class of events. A similar approach has proven to be highly successful in the spatial domain for the task of image representation (Malik, Belongie, Shi and Leung, 1999) indexing (Schmid and Mohr, 1997) and recognition (Hall et al., 2000; Weber, Welling and Perona, 2000; Leung and Malik, 2001). In the spatio-temporal domain, local descriptors have been previously used by (Chomat et al., 2000b) and others.

To describe a spatio-temporal neighborhood, we consider normalized spatio-temporal Gaussian derivatives defined as

$$L_{x^m y^n t^k} = \sigma^{m+n} \tau^k (\partial_{x^m y^n t^k} g) * f, \tag{15}$$

computed at the scales used for detecting the corresponding interest points. The normalization with respect to scale parameters guarantees the invariance of the derivative responses with respect to image scalings in both the spatial domain and the temporal domain. Using derivatives, we define event descriptors from the third order local jet[4] (Koenderink and van Doorn, 1987) computed at spatio-temporal scales determined from the detection scales of the corresponding interest points

$$j = (L_x, L_y, L_t, L_{xx}, ..., L_{ttt})\Big|_{\sigma^2=\tilde{\sigma}_i^2, \tau^2=\tilde{\tau}_i^2} \tag{16}$$

To compare two events, we compute the Mahalanobis distance between their descriptors as

$$d^2(j_1, j_2) = (j_1 - j_2)\Sigma^{-1}(j_1 - j_2)^T, \tag{17}$$

where Σ is a covariance matrix corresponding to the typical distribution of interest points in training data.

To detect similar events in the data, we apply k-means clustering (Duda, Hart and Stork, 2001) in the space of point descriptors and detect groups of points with similar spatio-temporal neighborhoods. Thus clustering of spatio-temporal neighborhoods is similar to the idea of textons (Malik et al., 1999) used to describe image texture as well as to detect object parts for spatial recognition (Weber et al., 2000). Given training sequences with periodic motion, we can expect repeating events to give rise to populated clusters. On the contrary, sporadic interest points can be expected to be sparsely distributed over the descriptor space giving rise to weakly populated clusters. To test this idea we applied k-means clustering with $k = 15$ to the sequence of a walking person in the upper row of figure 11. We found out that the four most densely populated clusters $c1, ..., c4$ indeed corresponded to stable interest points of the gait pattern. Local spatio-temporal neighborhoods of these points are shown in figure 10, where we can confirm the similarity of patterns inside the clusters and their difference between clusters.

To represent characteristic repetitive events in video, we compute cluster means $m_i = \frac{1}{n_i} \sum_{k=1}^{n_i} j_k$ for each significant cluster c_i consisting of n_i points. Then, in order to classify

[4]Note that our representation is currently not invariant with respect to planar image rotations. Such invariance could be added by considering steerable derivatives or rotationally invariant operators in space.

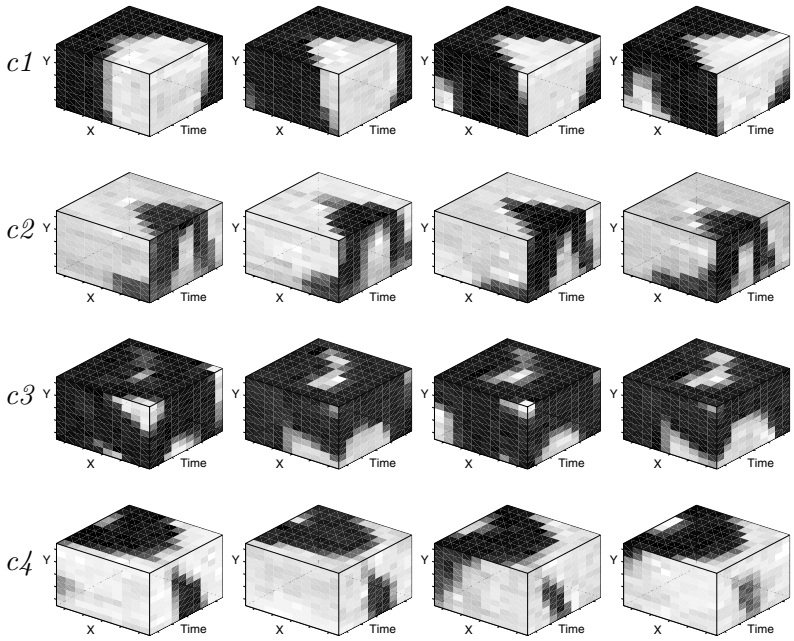


Figure 10: Local spatio-temporal neighborhoods of interest points corresponding to the first four most populated clusters obtained from a sequence of walking person.

an event on an unseen sequence, we assign the detected point to the cluster c_i that minimizes the distance $d(m_i, j_0)$ (17) between the jet of the interest point j_0 and the cluster mean m_i. If the distance is above a threshold, the point is classified as background. An application of this classification scheme is demonstrated in the second row of figure 11. As can be seen, most of the points corresponding to the gait pattern are correctly classified, while the other interest points are discarded. Observe that the person in the second sequence of figure 11 undergoes significant size changes in the image. Due to the scale-invariance of the interest points as well as their jet responses, the size transformations do not effect neither the result of event detection nor the result of classification.

5 Application to video interpretation

In this section, we illustrate how a sparse representation of video sequences by classified spatio-temporal interest points can be used for video interpretation. We consider the problem of detecting walking people and estimating their poses when viewed from the side in outdoor scenes. Such a task is complicated, since the variations in appearance of people together with the variations in the background may lead to ambiguous interpretations. Human motion is a strong cue that has been used to resolve this ambiguity in a number of previous works. Some of the works rely on pure spatial image features while using sophisticated body models and tracking schemes to constrain the interpretation (Baumberg and Hogg, 1996; Bregler and Malik, 1998; Sidenbladh, Black and Fleet, 2000). Other approaches use spatio-temporal image cues such as optical flow (Black, Yacoob, Jepson and

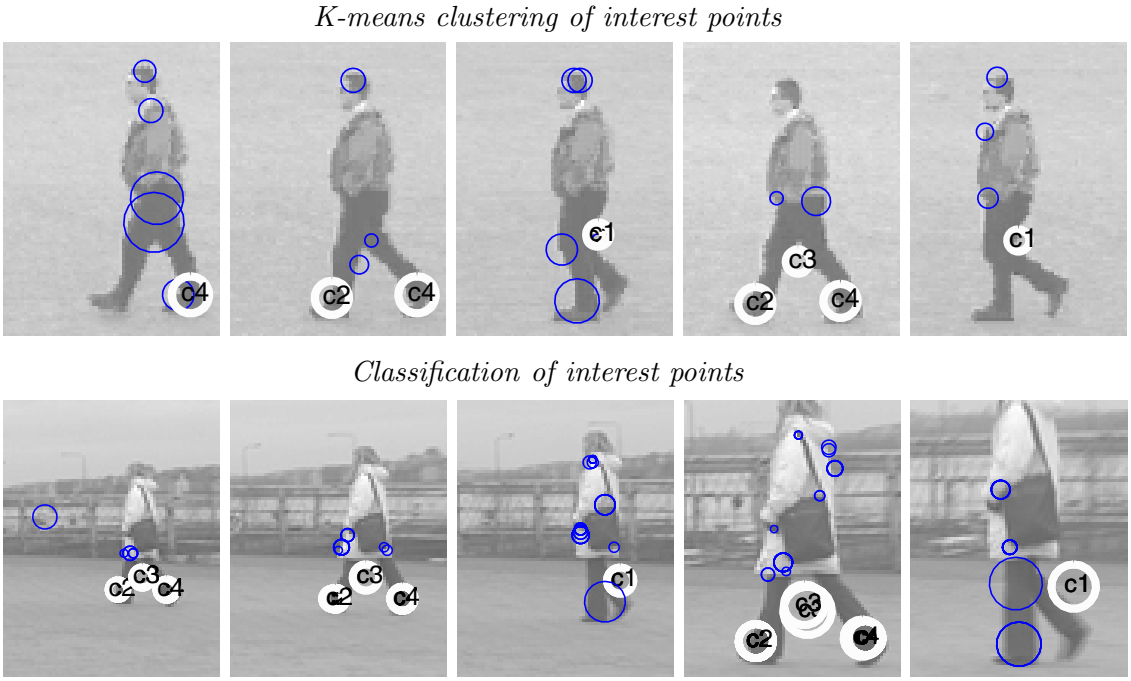


Figure 11: Interest points detected for sequences of walking persons. First row: the result of clustering spatio-temporal interest points in training data. The labelled points correspond to the four most populated clusters; Second row: the result of classifying interest points with respect to the clusters found in the first sequence.

Fleet, 1997) or motion templates (Baumberg and Hogg, 1996). The work of Niyogi and Adelson (1994) concerns the structure of the spatio-temporal gait pattern and is closer to ours.

The idea of the following approach is to represent both the model and the data using local and discriminative spatio-temporal features and to match the model by matching its features to the correspondent features of the data inside a spatio-temporal window (see figure 12).

5.1 Walking model

To obtain a model of a walking person, we consider the upper sequence in figure 11 and manually select a time interval (t_0, t_0+T) corresponding to the period T of the gait pattern. Then, given n features $f_i^m = (x_i^m, y_i^m, t_i^m, \sigma_i^m, \tau_i^m, c_i^m)$, $i = 1, ..., n$ (m stands for model) defined by the positions (x_i^m, y_i^m, t_i^m), scales (σ_i^m, τ_i^m) and classes c_i^m of interest points detected in the selected time interval, i.e. $t_i^m \in (t_0, t_0+T)$, we define the walking model by a set of periodically repeating features $M = \{f_i + (0, 0, kT, 0, 0, 0, 0) | i = 1, ..., n, k \in \mathcal{Z}\}$. Furthermore, to account for variations of the position and the size of a person in the image, we introduce a state for the model determined by the vector $X = (x, y, \theta, s, \xi, v_x, v_y, v_s)$. The components of X describe the position of the person in the image (x, y), his size s, the frequency of the gait ξ, the phase of the gait cycle θ at the current time moment as well as the temporal variations (v_x, v_y, v_s) of (x, y, s); v_x and v_y describe the velocity in

the image domain, while v_s describes how fast size changes occur. Given the state X, the parameters of each model feature $f \in M$ transform according to

$$\begin{aligned} \tilde{x}^m &= x + sx^m + \xi v_x(t^m + \theta) + s\xi x^m v_s(t^m + \theta) \\ \tilde{y}^m &= y + sy^m + \xi v_y(t^m + \theta) + s\xi y^m v_s(t^m + \theta) \\ \tilde{t}^m &= \xi(t^m + \theta) \\ \tilde{\sigma}^m &= s\sigma^m + v_s s\sigma^m(t^m + \theta) \\ \tilde{\tau}^m &= \xi\tau^m \\ \tilde{c}^m &= c^m \end{aligned} \tag{18}$$

It follows that this type of scheme is able to handle translations and uniform rescalings in the image domain as well as uniform rescalings in the temporal domain. Hence, it allows for matching of patterns with different image velocities as well as with different frequencies over time.

To estimate the boundary of the person, we extract silhouettes $S = \{x^s, y^s, \theta^s | \theta^s = 1, ..., T\}$ on the model sequence (see figure 11) one for each frame corresponding to the discrete value of the phase parameter θ. The silhouette is used here only for visualization purpose and allows us to approximate the boundary of the person in the current frame using the model state X and a set of points $\{(x^s, y^s, \theta^s) \in S | \theta^s = \theta\}$ transformed according to $\tilde{x}^s = sx^s + x$, $\tilde{y}^s = sy^s + y$.

5.2 Model matching

Given a model state X, a current time t_0, a length of the time window t_w, and a set of data features $D = \{f^d = (x^d, y^d, t^d, \sigma^d, \tau^d, c^d) | t^d \in (t_0, t_0 - t_w)\}$ detected from the the recent time window of the data sequence, the match between the model and the data is defined by a weighted sum of distances h between the model features f_i^m and the data features f_j^d

$$\mathcal{H}(\tilde{M}(X), D, t_0) = \sum_i^n h(\tilde{f}_i^m, f_j^d) e^{-(\tilde{t}_i^m - t_0)^2/\xi}, \tag{19}$$

where $\tilde{M}(X)$ is a set of n model features in the time window $(t_0, t_0 - t_w)$ transformed according to (18), i.e. $\tilde{M} = \{\tilde{f}^m | t^m \in (t_0, t_0 - t_w)\}$, $f_j^d \in D$ is a data feature minimizing the distance h for a given f_i^m and ξ is the variance of the exponential weighting function that gives more importance to recent features.

The distance h between two features of the same class is defined as a Euclidean distance between two points in space-time, where the spatial and the temporal dimensions are weighted with respect to a parameter ν as well as by the extents of the features in space-time

$$h^2(f^m, f^d) = (1 - \nu)\frac{(x^m - x^d)^2 + (y^m - y^d)^2}{(\sigma^m)^2} + \nu\frac{(t^m - t^d)^2}{(\tau^m)^2}. \tag{20}$$

Here, the distance between features of different classes is regarded as infinite. Alternatively, one could measure the feature distance by taking into account their descriptors and distances from several of the nearest cluster means.

To find the best match between the model and the data, we search for the model state $\tilde{X}$ that minimizes $\mathcal{H}$ in (19)

$$\tilde{X} = \text{argmin}_X \, \mathcal{H}(\tilde{M}(X), D, t_0) \tag{21}$$

using a standard Gauss-Newton optimization method. The result of such an optimization for a sequence with data features in figure 12(a) is illustrated in figure 12(b). Here, the match between the model and the data features was searched over a time window corresponding to three periods of the gait pattern or approximately 2 seconds of video. As can be seen from figure 12(c), the overlaps between the model features and the data features confirm the match between the model and the data. Moreover, the model silhouette transformed according to $\tilde{X}$ matches with the contours of the person in the current frame and confirms a reasonable estimate of the model parameters.

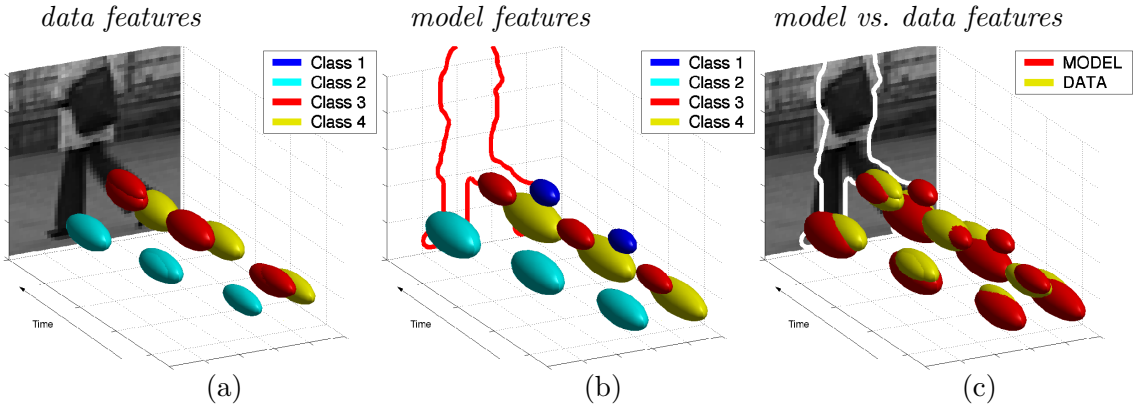


Figure 12: Matching of spatio-temporal data features with model features: (a) Features detected from the data sequence over a time interval corresponding to three periods of the gait cycle; (b) Model features minimizing the distance to the features in (a); (c) Model features and data features overlaid. The estimated silhouette overlayed on the current frame confirms the correctness of the method.

5.3 Results

Figure 13 presents results of the described approach applied to two outdoor sequences. The first sequence illustrates the invariance of the method with respect to size variations of the person in the image plane. The second sequence shows the successful detection and pose estimation of a person despite the presence of a complex non-stationary background and occlusions. Note that these results have been obtained by re-initializing model parameters before optimization at each frame. Hence, the approach is highly stable and could be improved further by tracking the model parameters $\tilde{X}$ over time.

The need for careful initialization and/or simple background are frequent obstacles in previous approaches for human motion analysis. The success of our method is due to the low ambiguity and simplicity of the matching scheme originating from the distinct and stable nature of the spatio-temporal features. In this respect, we want to propose direct detection of spatio-temporal events as an interesting alternative when representing and interpreting video data.

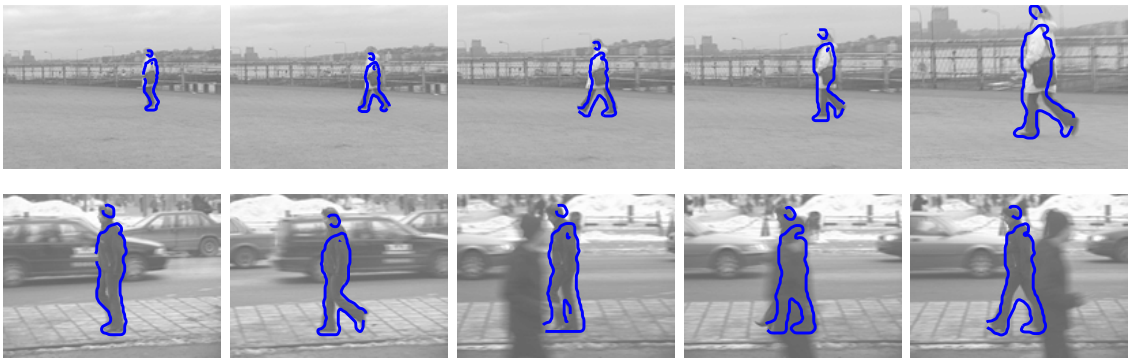


Figure 13: The result of matching a spatio-temporal walking model to sequences of outdoor scenes.

6 Summary

We have described an interest point detector that finds local image features in space-time characterized by a high variation of the image values in space and non-constant motion over time. From the presented examples, it follows that many of the detected points indeed correspond to meaningful events. Moreover, we propose local maximization of the normalized spatio-temporal Laplacian operator as a general tool for scale selection in space-time. Using this mechanism, we estimated characteristic spatio-temporal extents of detected events and computed their scale-invariant spatio-temporal descriptors.

Using scale-adapted descriptors in terms of N-jets we then addressed the problem of event classification and illustrated how classified spatio-temporal interest points constitute distinct and stable descriptors of events in video, which can be used for video representation and interpretation. In particular, we have shown how a video representation by spatio-temporal interest points enables detection and pose estimation of walking people in the presence of occlusions and highly cluttered and dynamic background. Note that this result was obtained using a standard optimization method without careful manual initialization or tracking.

In future work, we plan to extend application of interest points to the field of motion-based recognition (Schüldt, Laptev and Caputo, 2004). Moreover, as the current scheme of event detection is not invariant under Galilean transformations, future work should investigate the possibilities of including such invariance and making the approach independent of the relative camera motion (Laptev and Lindeberg, 2002; Laptev and Lindeberg, 2004). Another extension should consider the invariance of spatio-temporal descriptors with respect to the direction of motion, changes in image contrast and rotations. Finally, other types of space-time interest operators will be considered and investigated (Lindeberg, Akbarzadeh and Laptev, 2004).

7 Acknowledgments

We thank Anastasiya Syromyatnikova, Josephine Sullivan and Carsten Rother for their help in obtaining video data for the experiments.

References

Almansa, A. and Lindeberg, T. (2000). Fingerprint enhancement by shape adaptation of scale-space operators with automatic scale-selection, *IEEE Transactions on Image Processing* **9**(12): 2027–2042.

Barron, J., Fleet, D. and Beauchemin, S. (1994). Performance of optical flow techniques, *International Journal of Computer Vision* **12**(1): 43–77.

Baumberg, A. M. and Hogg, D. (1996). Generating spatiotemporal models from examples, *Image and Vision Computing* **14**(8): 525–532.

Bigün, J., Granlund, G. and Wiklund, J. (1991). Multidimensional orientation estimation with applications to texture analysis and optical flow, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(8): 775–790.

Black, M. and Jepson, A. (1998). Eigentracking: Robust matching and tracking of articulated objects using view-based representation, *International Journal of Computer Vision* **26**(1): 63–84.

Black, M., Yacoob, Y., Jepson, A. and Fleet, D. (1997). Learning parameterized models of image motion, *Proc. Computer Vision and Pattern Recognition*, pp. 561–567.

Blake, A. and Isard, M. (1998). Condensation – conditional density propagation for visual tracking, *International Journal of Computer Vision* **29**(1): 5–28.

Bregler, C. and Malik, J. (1998). Tracking people with twists and exponential maps, *Proc. Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. 8–15.

Bretzner, L. and Lindeberg, T. (1998). Feature tracking with automatic selection of spatial scales, *Computer Vision and Image Understanding* **71**(3): 385–392.

Chomat, O., de Verdiere, V., Hall, D. and Crowley, J. (2000a). Local scale selection for Gaussian based description techniques, *Proc. Sixth European Conference on Computer Vision*, Vol. 1842 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. I:117–133.

Chomat, O., Martin, J. and Crowley, J. (2000b). A probabilistic sensor for the perception and recognition of activities, *Proc. Sixth European Conference on Computer Vision*, Vol. 1842 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. I:487–503.

Duda, R., Hart, P. and Stork, D. (2001). *Pattern Classification*, Wiley.

Fergus, R., Perona, P. and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning, *Proc. Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. II:264–271.

Fleet, D., Black, M. and Jepson, A. (1998). Motion feature detection using steerable flow fields, *Proc. Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. 274–281.

Florack, L. M. J. (1997). *Image Structure*, Kluwer Academic Publishers, Dordrecht, Netherlands.

Förstner, W. A. and Gülch, E. (1987). A fast operator for detection and precise location of distinct points, corners and centers of circular features, *Proc. Intercommission Workshop of the Int. Soc. for Photogrammetry and Remote Sensing*, Interlaken, Switzerland.

Hall, D., de Verdiere, V. and Crowley, J. (2000). Object recognition using coloured receptive fields, *Proc. Sixth European Conference on Computer Vision*, Vol. 1842 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. I:164–177.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector, *Alvey Vision Conference*, pp. 147–152.

Hoey, J. and Little, J. (2000). Representation and recognition of complex human motion, *Proc. Computer Vision and Pattern Recognition*, Hilton Head, SC, pp. I:752–759.

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

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

Koenderink, J. J. and van Doorn, A. J. (1992). Generic neighborhood operators, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**(6): 597–605.

Laptev, I. and Lindeberg, T. (2002). Velocity-adaptation of spatio-temporal receptive fields for direct recognition of activities: An experimental study, *in* D. Suter (ed.), *Proc. ECCV'02 Workshop on Statistical Methods in Video Processing*, Copenhagen, Denmark, pp. 61–66.

Laptev, I. and Lindeberg, T. (2004). Velocity adaptation of space-time interest points, *in preparation*.

Leung, T. and Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons, *International Journal of Computer Vision* **43**(1): 29–44.

Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*, Kluwer Academic Publishers, Boston.

Lindeberg, T. (1997). On automatic selection of temporal scales in time-causal scale-space, *AFPAC'97: Algebraic Frames for the Perception-Action Cycle*, Vol. 1315 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, pp. 94–113.

Lindeberg, T. (1998). Feature detection with automatic scale selection, *International Journal of Computer Vision* **30**(2): 77–116.

Lindeberg, T. (2002). Time-recursive velocity-adapted spatio-temporal scale-space filters, *Proc. Seventh European Conference on Computer Vision*, Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, pp. I:52–67.

Lindeberg, T., Akbarzadeh, A. and Laptev, I. (2004). Galilean-corrected spatio-temporal interest operators, *in preparation*.

Lindeberg, T. and Bretzner, L. (2003). Real-time scale selection in hybrid multi-scale representations, *in* L. Griffin and M. Lillholm (eds), *Scale-Space'03*, Vol. 2695 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, pp. 148–163.

Lindeberg, T. and Fagerström, D. (1996). Scale-space with causal time direction, *Proc. Fourth European Conference on Computer Vision*, Vol. 1064 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Cambridge, UK, pp. I:229–240.

Lindeberg, T. and Garding, J. (1997). Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure, *Image and Vision Computing* **15**(6): 415–434.

Lowe, D. (1999). Object recognition from local scale-invariant features, *Proc. Seventh International Conference on Computer Vision*, Corfu, Greece, pp. 1150–1157.

Malik, J., Belongie, S., Shi, J. and Leung, T. (1999). Textons, contours and regions: Cue integration in image segmentation, *Proc. Seventh International Conference on Computer Vision*, Corfu, Greece, pp. 918–925.

Mikolajczyk, K. and Schmid, C. (2001). Indexing based on scale invariant interest points, *Proc. Eighth International Conference on Computer Vision*, Vancouver, Canada, pp. I:525–531.

Mikolajczyk, K. and Schmid, C. (2002). An affine invariant interest point detector, *Proc. Seventh European Conference on Computer Vision*, Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, pp. I:128–142.

Niyogi, S. A. (1995). Detecting kinetic occlusion, *Proc. Fifth International Conference on Computer Vision*, Cambridge, MA, pp. 1044–1049.

Niyogi, S. and Adelson, H. (1994). Analyzing and recognizing walking figures in XYT, *Proc. Computer Vision and Pattern Recognition*, pp. 469–474.

Schmid, C. and Mohr, R. (1997). Local grayvalue invariants for image retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(5): 530–535.

Schmid, C., Mohr, R. and Bauckhage, C. (2000). Evaluation of interest point detectors, *International Journal of Computer Vision* **37**(2): 151–172.

Schüldt, C., Laptev, I. and Caputo, B. (2004). Recognizing human actions: a local SVM approach, *in preparation*.

Sidenbladh, H., Black, M. and Fleet, D. (2000). Stochastic tracking of 3D human figures using 2D image motion, *Proc. Sixth European Conference on Computer Vision*, Vol. 1843 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. II:702–718.

Smith, S. and Brady, J. (1995). ASSET-2: Real-time motion segmentation and shape tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**(8): 814–820.

Tell, D. and Carlsson, S. (2002). Combining topology and appearance for wide baseline matching, *Proc. Seventh European Conference on Computer Vision*, Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, pp. I:68–83.

Tuytelaars, T. and Van Gool, L. (2000). Wide baseline stereo matching based on local, affinely invariant regions, *British Machine Vision Conference*, pp. 412–425.

Wallraven, C., Caputo, B. and Graf, A. (2003). Recognition with local features: the kernel recipe, *Proc. Ninth International Conference on Computer Vision*, Nice, France, pp. 257–264.

Weber, M., Welling, M. and Perona, P. (2000). Unsupervised learning of models for visual object class recognition, *Proc. Sixth European Conference on Computer Vision*, Vol. 1842 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. I:18–32.

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

Zelnik-Manor, L. and Irani, M. (2001). Event-based analysis of video, *Proc. Computer Vision and Pattern Recognition*, Kauai Marriott, Hawaii, pp. II:123–130.