Interest point detection and scale selection in space-time*

Ivan Laptev and Tony Lindeberg

Computational Vision and Active Perception Laboratory (CVAP)
Dept. of Numerical Analysis and Computing Science
KTH, S-100 44 Stockholm, Sweden

Abstract. Several types of interest point detectors have been proposed for spatial images. This paper investigates how this notion can be generalised to the detection of interesting events in space-time data. Moreover, we develop a mechanism for spatio-temporal scale selection and detect events at scales corresponding to their extent in both space and time.

To detect spatio-temporal events, we build on the idea of the Harris and Förstner interest point operators and detect regions in space-time where the image structures have significant local variations in both space and time. In this way, events that correspond to curved space-time structures are emphasised, while structures with locally constant motion are disregarded.

To construct this operator, we start from a multi-scale windowed second moment matrix in space-time, and combine the determinant and the trace in a similar way as for the spatial Harris operator. All space-time maxima of this operator are then adapted to characteristic scales by maximising a scale-normalised space-time Laplacian operator over both spatial scales and temporal scales. The motivation for performing temporal scale selection as a complement to previous approaches of spatial scale selection is to be able to robustly capture spatio-temporal events of different temporal extent. It is shown that the resulting approach is truly scale invariant with respect to both spatial scales and temporal scales.

The proposed concept is tested on synthetic and real image sequences. It is shown that the operator responds to distinct and stable points in space-time that often correspond to interesting events. The potential applications of the method are discussed.

1 Introduction

Analysing 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.

* 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.

Traditional approaches for motion analysis mainly involve the computation of optic flow (Barron, Fleet and Beauchemin 1994) and 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 often fail when the motion has sudden changes. Interesting solutions to this problem have been proposed by (Niyogi 1995, Fleet, Black and Jepson 1998). Feature trackers often assume the constant appearance of image patches over time and, hence, may fail when this 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 characterised by strong variations of the data in both the spatial and the temporal directions. For example, consider scenes with a person entering a room, applauding hand gestures, a car crash or a water splash. Moreover, it can be argued that changes of image velocity, i.e. accelerations of image structures are of particular interest since they may indicate the work of forces that act in the environment and change its structure.

In the spatial domain, points with a significant local variation of image intensities have been extensively investigated previously (Förstner and Gülch 1987, Harris and Stephens 1988, Schmid, Mohr and Bauckhage 2000). Such image points are frequently denoted as "interest points" and are attractive due to their high information contents. Highly successful applications of interest point detectors 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), and recognition (Lowe 1999, Hall, de Verdiere and Crowley 2000).

The purpose of this paper is to extend the notion of interest points into the spatio-temporal domain and to show that the resulting space-time features often correspond to interesting events in video. In particular we aim at the direct scheme for event detection that does not require feature tracking nor optic flow computation. As events often have characteristic extents in both space and time (Koenderink 1988, Lindeberg and Fagerström 1996, Florack 1997, Chomat, Martin and Crowley 2000b, Zelnik-Manor and Irani 2001), we investigate the behaviour of space-time interest points in spatio-temporal scale-space and adapt both the spatial and the temporal scales of the detected features to their characteristic extents in space-time. The idea of spatio-temporal interest points is illustrated in figure 1 where the result of a standard interest point detector applied to still images in a video is compared to the proposed spatio-temporal interest point detector. As can be seen, the spatio-temporal detector is more selective than the spatial one and detects specific events in the space-time cycle of a gait pattern.

To detect spatio-temporal events, we build on the idea of the Harris and Förstner interest point operators (Harris and Stephens 1988, Förstner and Gülch 1987) and derive the spatio-temporal event detector in section 2. We analyse its

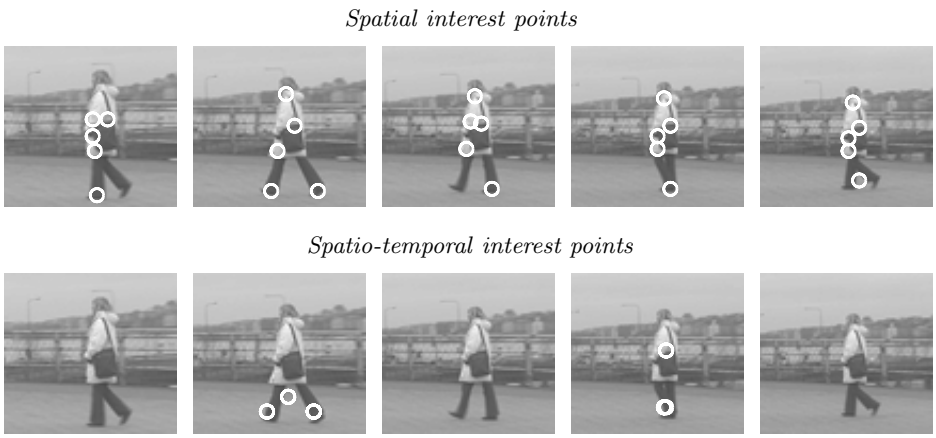


Fig. 1. Detection of spatial and spatio-temporal interest points in a video sequence. Compared to a spatial detector that selects points with high variations of image values in space, the spatio-temporal detector selects areas corresponding to distinct *events* with high variations of image values in both space and time.

behaviour on synthetic image sequences and motivate the need for automatic temporal scale selection. In section 3 we investigate a mechanism for simultaneous spatio-temporal scale selection based on the normalised spatio-temporal Laplace operator. In section 4 we propose an algorithm that adapts the detection of space-time interest points to their characteristic scales of observations by combining the theory from sections 2 and 3. The performance of the resulting detector on real image sequences is investigated in section 5. Finally, section 6 concludes the paper with the discussion of the method and its potential applications.

2 Interest point detection

2.1 Interest points in spatial domain

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

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

defined by the convolution of f^s with Gaussian kernels of variance σ_l^2

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

The idea of the Harris interest point detector is to find spatial locations where f^s 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 the variance σ_i^2 (Förstner and Gülch 1987, Bigün, Granlund and Wiklund 1991, Garding and Lindeberg 1996):

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

where $'*'$ denotes convolution operator, and L_x^s and L_y^s are Gaussian derivatives computed at the local scale σ_l^2 and defined as $L_x^s = \partial_x(g^s(\cdot;\, \sigma_l^2) * f^s(\cdot))$, $L_y^s = \partial_y(g^s(\cdot;\, \sigma_l^2) * f^s(\cdot))$. The second moment descriptor can be thought of as the covariance matrix of a two-dimensional distribution of image orientations in the local neighbourhood of a point. Hence, the eigenvalues λ_1, λ_2, $(\lambda_1 \leq \lambda_2)$ of μ^s represent characteristic variations of f^s in the both image directions while two significant 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

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

The ratio of the eigenvalues $\alpha = \lambda_2/\lambda_1$ has to be high at the positions of the interest points. From (4) it follows that for positive local maxima of H^s 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 "ideal" 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. The 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 in the top row of figure 1.

2.2 Interest points in the spatio-temporal domain

In this section, we develop an operator that responds to events in temporal image sequences with specific positions and extents in space-time. The idea of interest points in the spatial domain can be extended into the spatio-temporal domain by requiring image values in space-time to have large variations in both the spatial and the temporal directions. Points with such properties will be spatial interest points with a distinct location in time corresponding to a local spatio-temporal neighbourhoods 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

[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

variance σ_l^2 and temporal variance τ_l^2

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

where the spatio-temporal separable Gaussian kernel is defined as

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

The introduction of 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 spatial and temporal scales of observation and, hence, require separate treatment of the scale parameters σ_l^2 and τ_l^2.

Similar to the spatial domain, we consider the spatio-temporal second-moment matrix which is a 3-by-3 matrix composed of first order spatial and temporal derivatives averaged with 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 the integration scales are $\sigma_i^2 = s\sigma_l^2$ and $\tau_i^2 = s\tau_l^2$ while the first-order derivatives are defined as

$$L_\xi(\cdot;\, \sigma_l^2, \tau_l^2) = \partial_\xi (g * f).$$

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

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

To show that 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 of image gray-values in both the spatial and the temporal directions.

(Koenderink 1988, Lindeberg and Fagerström 1996, Florack 1997). In this paper, however, we simplify the investigation and assume that the data is available for a sufficiently long period of time and the image sequence can be convolved with a Gaussian kernel in both space and time.

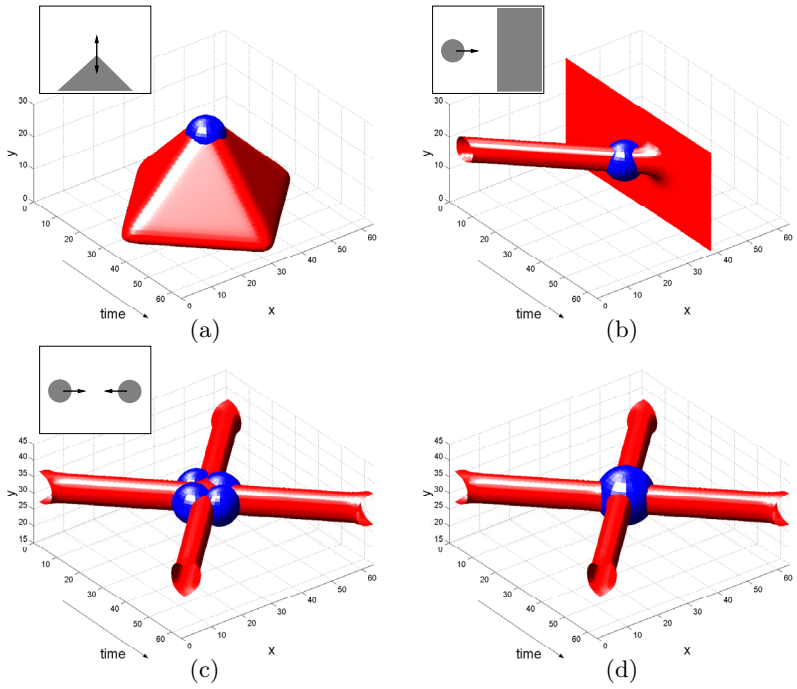


Fig. 2. Results of detecting spatio-temporal interest points on synthetic image sequences. (a): Moving corner; (b) A merge of a ball and a wall; (c): Collision of two balls with interest points detected at scales $\sigma_l^2 = 8$ and $\tau_l^2 = 8$; (d): the same as in (c) but with interest points detected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

In particular, if we set the maximal value of α, β to 23 as in the spatial domain, the value of k to be used in H (8) will then be $k = 0.005$. Thus, spatio-temporal interest points of f can be found by detecting local positive spatio-temporal maxima in H.

2.3 Experimental results on 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 presented by ellipsoids with positions corresponding to the space-time location of interest points 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 illustrates 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 image structures with constant motion do not give rise to the detection of interest points. Other typical types of events 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 emphasise 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 the threshold surface $y = \sin(x^4) * \sin(t^4)$ (see figure 3). As can be seen, the result of detecting the strongest interest points highly depends on the scale parameters σ_l^2 and τ_l^2. We 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 events is related to the value of the spatial scale parameter σ_l^2.

From the presented examples it follows that a correct selection of temporal and spatial scales is crucial when capturing the events with different spatial and temporal extents. Moreover, estimation of the spatio-temporal extents of events can be interesting for their further interpretation. In the next section, we propose a mechanism for simultaneous estimation of spatio-temporal scales. This mechanism is combined with the interest point detector in section 4.

3 Scale selection in space-time

During recent years, the problem of automatic scale selection has been addressed in several different ways, based on the maximisation of normalised derivative expressions over scale, or the behaviour of entropy measures or error measures over scales (see the companion paper by 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 the spatial and the temporal scales that are characteristic for an event with a particular spatio-temporal location.

For the purpose of analysis we study a prototype event represented by a spatio-temporal Gaussian blob $f = g(x, y, t; \sigma_0^2, \tau_0^2)$ 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 $L(x, y, t; \sigma^2, \tau^2) = g(x, y, t; \sigma_0^2 + \sigma^2, \tau_0^2 + \tau^2)$.

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

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

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

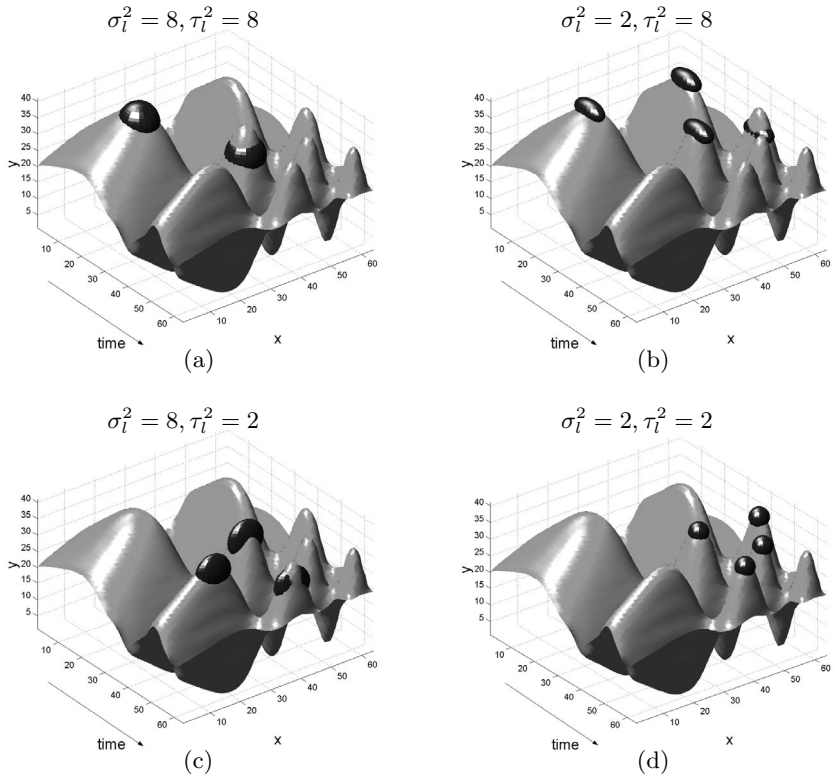


Fig. 3. Results of interest point detection at different spatial and temporal scales for a synthetic sequence with impulses having varying extents in space and time. The extents of the detected events roughly corresponds to the scale parameters σ_l^2 and τ_l^2 used in the computation of H.

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 centre of the blob

$$(L_{xx,norm})'_{\sigma^2} = -\frac{a\sigma^2 - 2\sigma^2 + a\sigma_0^2}{\sqrt{(2\pi)^3(\sigma_0^2 + \sigma^2)^6(\tau_0^2 + \tau^2)}}\, \sigma^{2(a-1)}\tau^{2b} \tag{10}$$

$$(L_{xx,norm})'_{\tau^2} = -\frac{2b\tau_0^2 + 2b\tau^2 - \tau^2}{\sqrt{2^5\pi^3(\sigma_0^2 + \sigma^2)^4(\tau_0^2 + \tau^2)^3}}\, \tau^{2(b-1)}\sigma^{2a}. \tag{11}$$

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

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

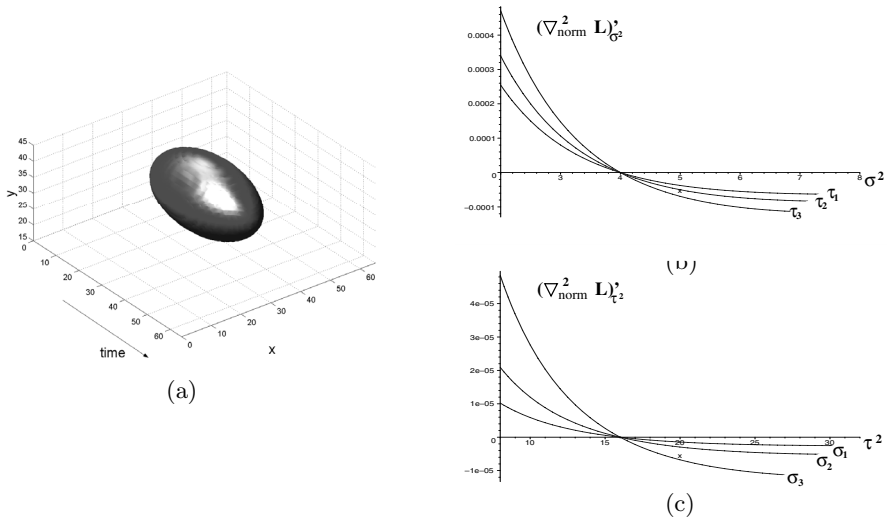


Fig. 4. (a): Spatio-temporal Gaussian blob with spatial variance $\sigma_0^2 = 4$ and temporal variance $\tau_0^2 = 16$; (b)-(c) derivatives of $\nabla^2_{norm} L$ with respect to scales. The zero-crossings of $(\nabla^2_{norm} L)'_{\sigma^2}$ and $(\nabla^2_{norm} L)'_{\tau^2}$ indicate extrema of $\nabla^2_{norm}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

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

$$(L_{tt,norm})'_{\sigma^2} = -\frac{c\sigma^2 - \sigma^2 + c\sigma_0^2}{\sqrt{(2\pi)^3(\sigma_0^2 + \sigma^2)^4(\tau_0^2 + \tau^2)^3}}\, \sigma^{2(c-1)}\tau^{2d} \tag{12}$$

$$(L_{tt,norm})'_{\tau^2} = -\frac{2d\tau_0^2 + 2d\tau^2 - 3\tau^2}{\sqrt{2^5\pi^3(\sigma_0^2 + \sigma^2)^2(\tau_0^2 + \tau^2)^5}}\, \tau^{2(d-1)}\sigma^{2c} \tag{13}$$

leads to the expressions

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

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

The derived normalisation of derivatives in (9) guarantees that all of them assume space-time-scale extrema at the centre of the blob f and at scales corresponding to the spatial and the temporal extents of f, i.e. $\sigma = \sigma_0$ and $\tau = \tau_0$. The sum of these derivatives defines the normalised spatio-temporal Laplace operator

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

Figures 4b-c show derivatives of this operator with respect to the scale parameters evaluated at the centre 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 blob can be estimated by finding the extrema of $\nabla^2_{norm} L$ over both spatial and temporal scales.

4 Scale-adapted space-time interest points

Local scale estimation using the normalised 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 normalised 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^s (4) and extrema over scale of the scale-normalised 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 normalised 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 maximise $(\nabla^2_{norm} L)^2$ 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 neighbouring spatio-temporal scale that maximises $(\nabla^2_{norm} L)^2$ and (ii) re-detecting

---

1. Detect interest points $p_j = (x_j, y_j, t_j, \sigma^2_{l,j}, \tau^2_{l,j})$, $j = 1..N$ as maxima of H (8) over space and time using combinations of initial spatial scales $\sigma_l^2 = \sigma^2_{l,1}, .., \sigma^2_{l,n}$ and temporal scales $\tau_l^2 = \tau^2_{l,1}, .., \tau^2_{l,m}$ 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 $\nabla^2_{norm} L$ at position (x_j, y_j, t_j) and combinations of neighbouring scales $(\tilde{\sigma}^2_{i,j}, \tilde{\tau}^2_{i,j})$ where $\tilde{\sigma}^2_{i,j} = 2^\delta \sigma^2_{i,j}$, $\tilde{\tau}^2_{i,j} = 2^\delta \tau^2_{i,j}$, and $\delta = -0.25, 0, 0.25$

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

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

Re-detect interest point $\tilde{p}_j = (\tilde{x}_j, \tilde{y}_j, \tilde{t}_j, \tilde{\sigma}^2_{l,j}, \tilde{\tau}^2_{l,j})$ using integration scales $\tilde{\sigma}^2_{i,j} = \tilde{\sigma}^2_{i,j}$, $\tilde{\tau}^2_{i,j} = \tilde{\tau}^2_{i,j}$, local scales $\tilde{\sigma}^2_{l,j} = \frac{1}{s}\tilde{\sigma}^2_{i,j}$, $\tilde{\tau}^2_{l,j} = \frac{1}{s}\tilde{\tau}^2_{i,j}$ 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**

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

the space-time location of the interest point at a new scale. 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.

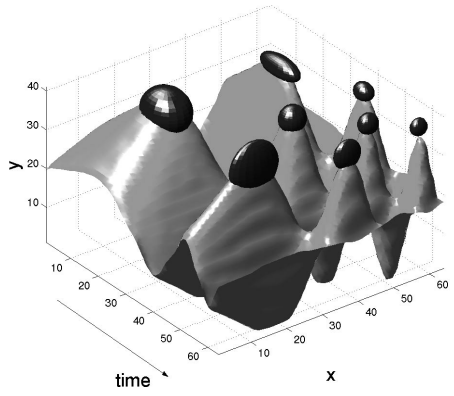
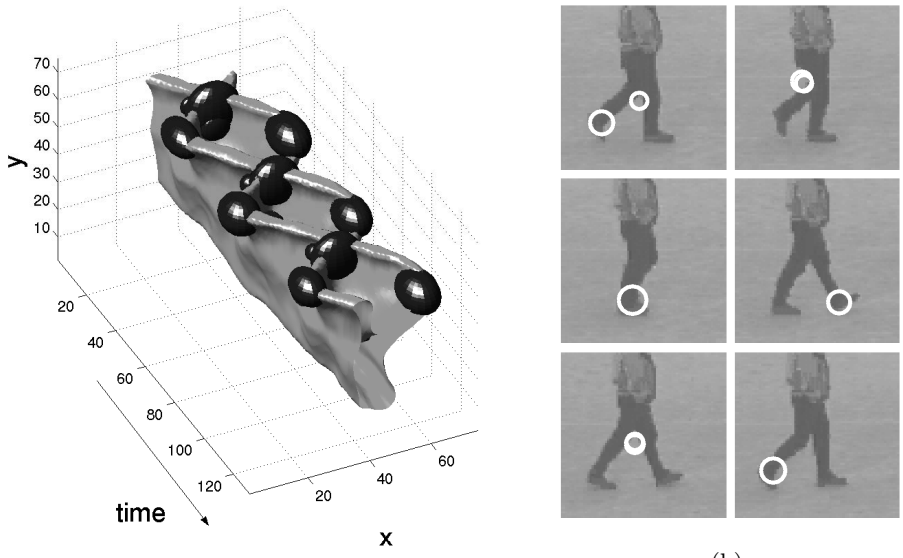


Fig. 6. The result of scale-adaptation of spatio-temporal interest points computed from a space-time pattern of the form $y = \sin(x^4) * \sin(t^4)$. The interest points are illustrated as ellipsoids showing the selected spatio-temporal scales overlayed on a surface plot of the intensity landscape.

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), 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 maximise $(\nabla^2_{norm} L)^2$ in a local neighbourhood of the spatio-temporal scale-space.

5 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 pattern gives rise to stable interest points. Note that the detected interest points reflect well-localised events in both space and time, corresponding to space-time structures such as the starting and the stopping feet. From the space-time plot in figure 7(a) we can also observe how the selected spatial and temporal scales of the detected features roughly match the spatio-temporal extents of the corresponding image structures.



(a) (b)

Fig. 7. Results of detecting spatio-temporal interest points for the motion of the legs of a walking person. (a): 3-D plot with a threshold surface of a leg pattern (up side down) and detected interest points; (b): interest points overlayed on single frames of a sequence.

Figure 8 illustrates 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 beside events in the leg pattern, the detector finds spurious points due to the non-constant motion of a coat and arms. However, image structures with constant motion in the background do not result in the response of the detector.

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 high frequency on the left and low frequency on the right. The distinct interest points are detected at the moments and at 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). On the contrary, hand waves with low frequency result in interest points with long temporal extent as shown in figure 9b.

Figure 10 illustrates a football sequence with a player heading the ball. The sequence has multiple motions due to camera zooming and motion of objects in the scene and is probably hard to analyse using standard methods for motion estimation and tracking. However, the strongest output of the proposed detector

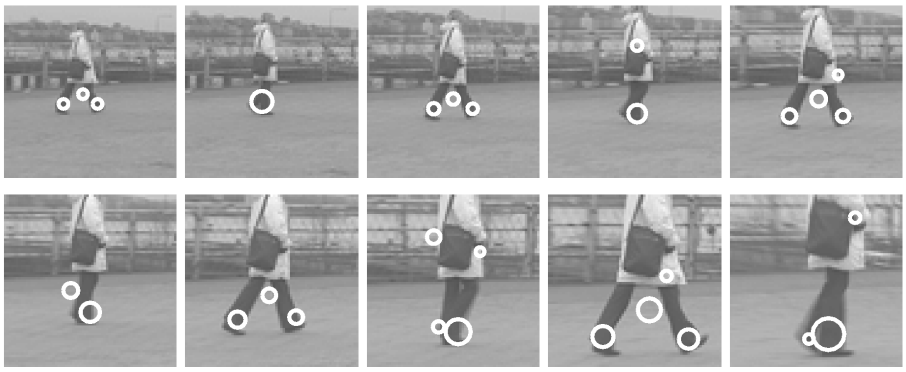


Fig. 8. Results of interest point detection for a zoom-in sequence of a walking person. The spatial scale of the detected points (corresponding to the size of circles) matches the increasing spatial extent of image structures and verifies the invariance of the interest points with respect to changes in spatial scale.

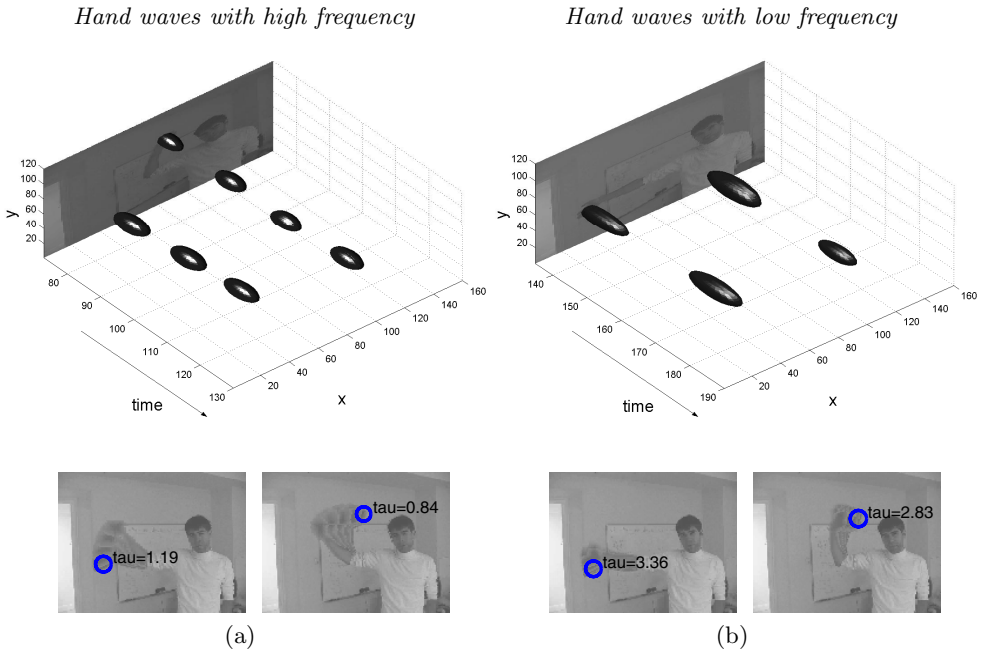


Fig. 9. Result of interest point detection for a sequence with waving hand gestures. (a) Interest points for hand gestures with high frequency; (b) Interest points for hand gestures with low frequency.

(the interest point with the highest maxima of H in (8)) corresponds to the position and the moment of the most significant event in the sequence, i.e. the heading of the ball.

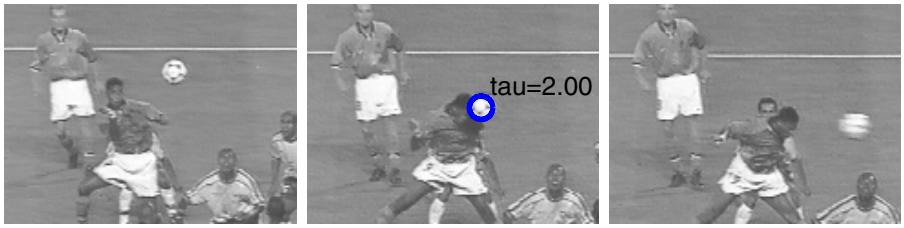


Fig. 10. Detection of the strongest interest point in a football sequence with a player heading the ball.

6 Summary and discussion

We have proposed an interest point detector that finds events in space-time with high variations of the image values in space and non-constant motion in time. From the experimental results that have been presented in previous sections, it can be seen that many of the points detected in this way correspond to space-time structures that we would intuitively regard as meaningful events. For example, for the sequences with walking people (figures 7-8) we obtain responses at the beginning and the end of the gait cycle and at spatial locations corresponding to distinct body parts.

As temporal events exist over finite periods of time, the notion of temporal scale is incorporated in the detector and the method for automatic scale selection is used for estimating temporal as well as spatial extents of detected events.

The current implementation of this interest point detector is based on separable space-time filters and is therefore not invariant to Galilean transformations over time, e.g. caused by relative motions of the camera. To aim at Galilean invariance, one could either perform local stabilisation as done by Zelnik-Manor and Irani (2001) or consider (possibly ensembles of) spatio-temporal receptive fields that have been adapted to local directions in space-time (Laptev and Lindeberg 2002).

Regarding potential applications of the presented techniques, one area of interest concerns sparse representation of video data. A representation in terms of interest points could be used for matching between image sequences, or for matching an articulated model over time to a given video sequence. Furthermore, the spatio-temporal interest points could be attributed with higher order spatio-temporal derivatives or other types of image descriptors evaluated at positions, moments and scales estimated by the proposed detector. Combinations of classified spatio-temporal interest points could then be used for describing and analysing image sequences based on similar techniques as have been proposed for interest points in the spatial domain.

Figure 11 shows an example of using such an approach for matching a template gait pattern derived from one walking person to the gait pattern of another person using classified spatio-temporal interest points that have been grouped based on similar spatio-temporal receptive field responses and K-means cluster-

Fig. 11. Result of alignment of the model sequence (see figure 7) to the data sequence using classified spatio-temporal interest points. The details of the method are leaved out due to space limitations and will be presented elsewhere.

ing. Notably this result was obtained in relation to a complex cluttered background with multiple motions and using neither manual initialisation nor explicit tracking.

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.

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). Eigen tracking: Robust matching and tracking of articulated objects using view-based representation, *International Journal of Computer Vision* **26**(1): 63–84. .

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

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. 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*, Dublin, Ireland, pp. I:487–503.

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.

Garding, J. and Lindeberg, T. (1996). Direct computation of shape cues using scale-adapted spatial derivative operators, *International Journal of Computer Vision* **17**(2): 163–191.

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. 164–177.

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

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.

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. and Bretzner, L. (2003). Real-time scale selection in hybrid multi-scale representations, *Proc. Scale-Space'03*, LNCS, Springer Verlag, these proceedings.

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. .

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

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.

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.

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.

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.