Velocity-adaptation of spatio-temporal receptive fields for direct recognition of activities: An experimental study*

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–02/04–SE

Shortened version in
Proc. ECCV'02 Workshop on Statistical Methods in Video Processing
Copenhagen, Denmark, June 2001, pp. 61–66.

Abstract

This article presents an experimental study of the influence of velocity adaptation when recognizing spatio-temporal patterns using a histogram-based statistical framework. The basic idea consists of adapting the shapes of the filter kernels to the local direction of motion, so as to allow the computation of image descriptors that are invariant to the relative motion in the image plane between the camera and the objects or events that are studied. Based on a framework of recursive spatio-temporal scale-space, we first outline how a straightforward mechanism for local velocity adaptation can be expressed. Then, for a test problem of recognizing activities, we present an experimental evaluation, which shows the advantages of using velocity-adapted spatio-temporal receptive fields, compared to directional derivatives or regular partial derivatives for which the filter kernels have not been adapted to the local image motion.

Key words: motion, spatio-temporal filtering, scale-space, recognition

1 Introduction

A recent approach for recognition consists of computing statistical descriptors of receptive field responses. In particular, histogram based schemes of derivative operators have emerged as an interesting alternative for formulating recognition schemes for static as well as time-dependent image data (Swain and Ballard, 1991; Schiele

*The support from the Swedish Research Council for Engineering Sciences (TFR), the Swedish Research Council (VR), as well as the Royal Swedish Academy of Sciences (KVA) and the Knut and Alice Wallenberg Foundation is gratefully acknowledged.

and Crowley, 2000; Chomat, de Verdiere, Hall and Crowley, 2000a; Hall, de Verdiere and Crowley, 2000; Schneiderman and Kanade, 2000; Chomat, Martin and Crowley, 2000b; Zelnik-Manor and Irani, 2001). Computing responses of local spatio-temporal receptive fields involves filtering in both space and time. This naturally rises the question of how the existing framework for spatial filtering can be extended to the spatio-temporal domain.

When analysing spatio-temporal image data, one observation that can be made is that temporal events can often be characterized by their extents over time in a similar manner as spatial structures have their characteristic scales in space. This motivates and emphasizes the need for analysing spatio-temporal data at different scales, both with respect to time and space (Witkin, 1983; Koenderink, 1984; Koenderink, 1988; Lindeberg, 1994; Lindeberg and Fagerström, 1996; Florack, 1997; Lindeberg, 1997a).

The temporal domain, however, also has a number of specific properties, which differ from spatial data, and which must be taken into account explicitly. A basic constraint on real-time processing is that the time direction is causal, and real-time algorithms may only access information from the past (Koenderink, 1988; Lindeberg and Fagerström, 1996). Another difference concerns the classes of characteristic transformations that influence the data. Whereas perspective transformations have a high influence on the image data in the spatial image domain, one of the most important sources of changes in the temporal dimension is due to motion between the observer and the patterns that are studied. This is illustrated in figure 1, where the spatio-temporal pattern of a walking person is influenced by the relative motion of the camera (figures 1(b)–(c)). If separable spatial filtering is extended to the temporal domain, we observe that filter responses are highly dependent on the relative motion between the person and the camera (figures 1(d)–(f)).

When interpreting image data, it is important to base the analysis on image representations that are invariant to the external imaging conditions. Hence, it is important to construct representations of spatio-temporal patterns that are independent of the relative motion between the patterns and the observer. Previous work has addressed this problem by first stabilizing patterns of interest in the field of view, and then computing spatio-temporal descriptors using a fixed set of filters (Zelnik-Manor and Irani, 2001); see also (Irani, Anandan and Hsu, 1995) for related stabilization approaches. Camera stabilization, however, may not always be available, for example, in situations with multiple moving objects, moving backgrounds or in cases where initial segmentation of the patterns of interest cannot be done without (preliminary) recognition.

A main aim of this work is to define and compute spatio-temporal descriptors that compensate for the relative motion between the pattern and the observer and do not rely on external camera stabilization. This is performed by local velocity adaptation of receptive fields. In Section 2 we first introduce velocity-adapted filtering using the framework of spatio-temporal scale-space. Then in section 3, the mechanism for performing local velocity adaptation is described. By integration with a histogram-based statistical framework in section 4, we then consider a test problem of recognizing activities and show how velocity adaptation results in a considerable increase in recognition performance compared to two other receptive field representations not involving velocity adaptation. Section 5 concludes the paper with a summary and discussion.

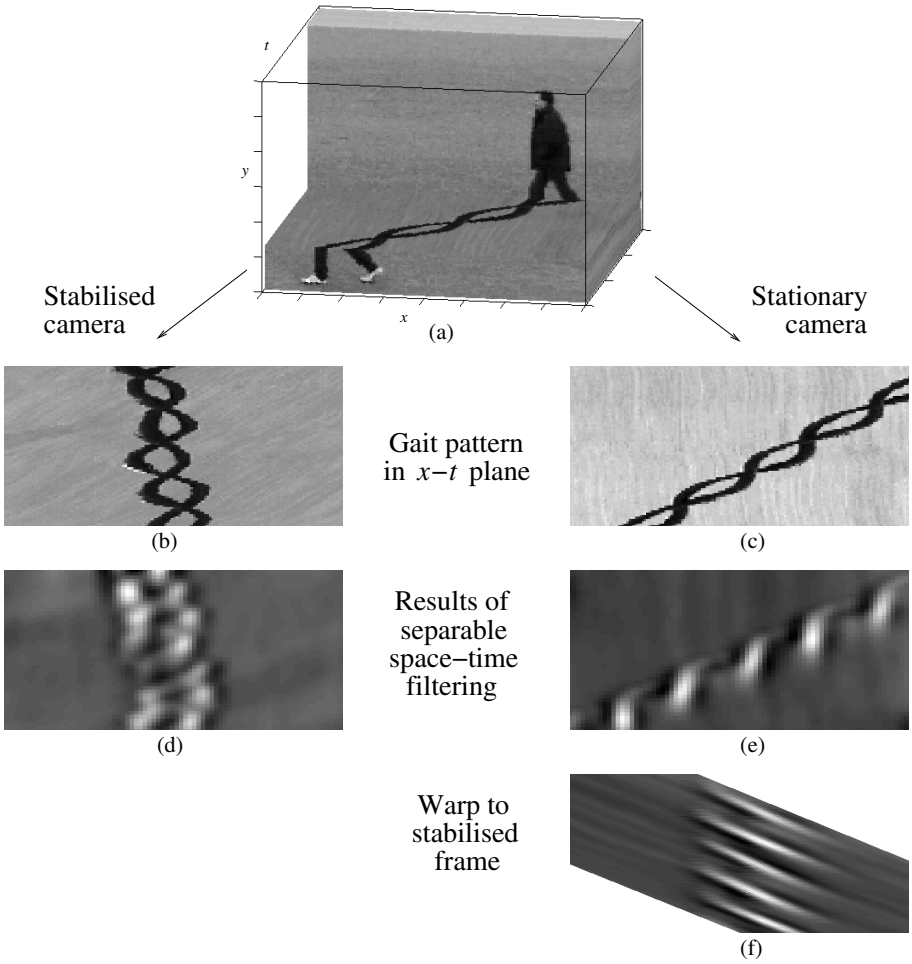

Figure 1: Spatio-temporal image of a walking person (a) depends on the relative motion between the person and the camera (b)-(c). If this motion is not taken into account, spatio-temporal filtering (here, the second order spatial derivative) results in highly different responses as illustrated in (d) and (e). Manual stabilization of the pattern in (e) shown in (f) makes the difference more explicit for comparisons with (d).

1.1 Related work

Velocity adaptation of spatio-temporal receptive fields follows the idea of shape adaptation in the spatial domain, which has previously been considered by (Lindeberg and Garding, 1994; Ballester and Gonzalez, 1998; Florack, Niessen and Nielsen, 1998; Weickert, 1998; Almansa and Lindeberg, 2000; Schaffalitzky and Zisserman, 2001; Mikolajczyk and Schmid, 2002). In the spatio-temporal domain, adaptive spatio-temporal filters have been studied by (Lindeberg, 1997a; Nagel and Gehrke, 1998; Lindeberg, 2002); see also (Black, 1994; Guichard, 1998). Nagel and Gehrke (Nagel and Gehrke, 1998) proposed an adaptation scheme close to ours and used it for robust estimation of optic flow.

With regard to recognition, this work relates to histogram-based methods first proposed in the spatial domain by (Swain and Ballard, 1991) using color histograms

computed from single pixel responses. Extensions to receptive field histograms were later presented by (Schiele and Crowley, 2000; Chomat et al., 2000a; Hall et al., 2000; Schneiderman and Kanade, 2000). Specifically, combinations of automatic scale selection in the spatial domain (Lindeberg, 1998) with histogram based recognition schemes have been presented by (Chomat et al., 2000a; Hall et al., 2000). In the spatio-temporal domain, histogram-based approaches have been used for the recognition of activities by (Chomat et al., 2000b; Zelnik-Manor and Irani, 2001). Here, we build upon this work and show how the performance of spatio-temporal recognition schemes can be increased by velocity adaptation.

2 Spatio-temporal scale-space

The image data we analyse is a spatio-temporal image sequence, in the continuous case modeled as a function $f\colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ or in the discrete case as $f\colon \mathbb{Z}^2 \times \mathbb{Z} \to \mathbb{Z}$. From this signal, a separable spatio-temporal scale-space L is computed by separable convolution with a set of spatial smoothing kernels $g(x, y;\ \sigma^2)$ with variances σ^2 and a set of temporal smoothing kernels $h(t;\ \tau^2)$ with variances τ^2. For continuous data, the natural choice of a spatial smoothing kernel is the Gaussian kernel (Witkin, 1983; Koenderink, 1984; Lindeberg, 1994; Florack, 1997). Regarding continuous time, we may model the temporal smoothing either by a non-causal Gaussian kernel, or as a causal Gaussian kernel on a logarithmically transformed temporal domain (Koenderink, 1988; Lindeberg and Fagerström, 1996). For discrete data, a canonical spatial scale-space concept originates from the discrete analogue of the Gaussian kernel

$$g(x, y;\ \sigma^2) = e^{-2\sigma^2} I_x(\sigma^2)\, I_y(\sigma^2) \tag{1}$$

where I_x and I_y denote the modified Bessel functions of integer order (Abramowitz and Stegun, 1964). Regarding discrete time, a natural and computationally efficient scale-space representation can be computed by coupling first-order recursive filters in cascade (Lindeberg, 1994; Lindeberg and Fagerström, 1996)

$$L_{out}(x, y, t) = \frac{1}{1+\mu}\,(L_{in}(x, y, t) + \mu L_{out}(x, y, t-1)). \tag{2}$$

The temporal mean of the corresponding filter is μ and the temporal variance is μ^2. Thus, by coupling k such filters in cascade, we obtain a filter with mean $m_k = \sum_{i=1}^{k} \mu_i$ and variance $\tau_k^2 = \sum_{i=1}^{k} \mu_i^2 + \mu_i$.

It can be shown that if for a given variance τ^2 we let $\mu_i = \tau^2/K$ become successively smaller by increasing the number of filtering steps K, then the filter kernel approaches the Poisson kernel (Lindeberg, 1997a), which corresponds to the canonical temporal scale-space concept having a continuous scale parameter on a discrete temporal domain. Another practical advantage of the recursive filtering scheme in (2) is that it enables the computation of temporal scale-space representations without need of buffering previous time frames.

2.1 Transformation properties under motion

To describe spatio-temporal smoothing step, we will henceforth use covariance matrices of filter kernels. For a separable smoothing kernel, with a spatial variance σ^2

and a temporal variance τ^2, the covariance matrix is diagonal:

$$\Sigma = \begin{pmatrix} C_{xx} & C_{xt} & C_{xt} \\ C_{xy} & C_{yy} & C_{yt} \\ C_{xt} & C_{yt} & C_{tt} \end{pmatrix} = \begin{pmatrix} \sigma^2 & & \\ & \sigma^2 & \\ & & \tau_k^2 \end{pmatrix}. \tag{3}$$

A limitation of using a separable scale-space for analysing motion patterns, however, originates from the fact that this scale-space concept is not closed under 2-D motions in the image plane. For a 2-D Galilean motion

$$\begin{pmatrix} x' \\ y' \\ t' \end{pmatrix} = \begin{pmatrix} 1 & 0 & v_x \\ 0 & 1 & v_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ t \end{pmatrix} \tag{4}$$

the covariance matrix of the smoothing kernel transforms as (Lindeberg, 1997a; Lindeberg, 2002)

$$\begin{pmatrix} C'_{xx} & C'_{xt} & C'_{xt} \\ C'_{xy} & C'_{yy} & C'_{yt} \\ C'_{xt} & C'_{yt} & C'_{tt} \end{pmatrix} = \begin{pmatrix} 1 & 0 & v_x \\ 0 & 1 & v_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} C_{xx} & C_{xt} & C_{xt} \\ C_{xy} & C_{yy} & C_{yt} \\ C_{xt} & C_{yt} & C_{tt} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ v_x & v_y & 1 \end{pmatrix} \tag{5}$$

and spatio-temporal derivatives transform according to

$$\partial_{x'} = \partial_x \qquad \partial_{y'} = \partial_y \qquad \partial_{t'} = v_x\, \partial_x + v_y\, \partial_y + \partial_t. \tag{6}$$

Hence, if we consider separable smoothing kernels only and if we do not take the transformation property of spatio-temporal derivatives into explicit account, it will not be possible to perfectly match the spatio-temporal scale-space representations for different amounts of motion.

2.2 Scale-space with velocity adaptation

A natural way of defining a scale-space that is closed under Galilean motion in the image plane, is by considering a scale-space representation that is parameterized by the full family of (positive definite) covariance matrices (Lindeberg, 1997a; Florack, 1997; Lindeberg, 2002). In terms of implementation, there are two basic ways of computing such a scale-space — either by transforming the smoothing kernels themselves, or by transforming the input image prior to smoothing (see figure 2). In this work, the latter approach is taken, and for reasons of simplicity and computational efficiency, we restrict the set of image velocities to integer multiples of the pixel size. Thus, in combination with spatial smoothing

$$L^{(0)}(x, y, t;\, \sigma^2) = g(x, y;\, \sigma^2) * f(x, y, t), \tag{7}$$

a set of velocity-adapted time-recursive smoothing steps is computed according to

$$L^{(k+1)}(x, y, t;\, \sigma^2) = \frac{1}{1 + \mu_k} \left(L^{(k)}(x, y, t;\, \sigma^2) + \mu_k L^{(k+1)}(x - v_x, y - v_y, t - 1;\, \sigma^2)\right), \tag{8}$$

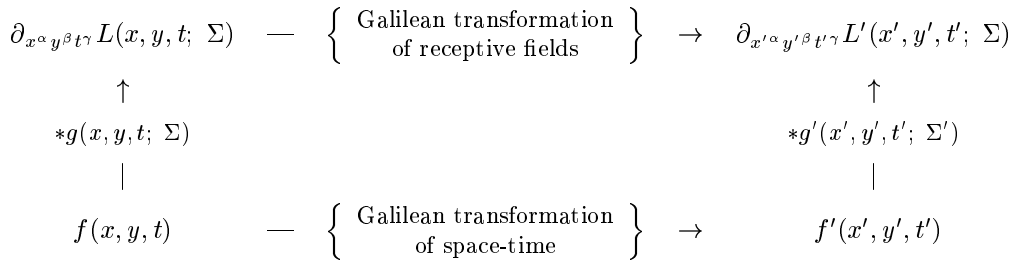

Figure 2: A pre-requisite for perfect matching of spatio-temporal receptive field responses for different amounts of motion is that the image representation is closed under motions in the image domain. The aim of the velocity-adaptation mechanism is to allow for such closedness, and to permit the construction of a velocity invariant recognition scheme.

where k represents the level of temporal smoothing corresponding to the convolution with temporal kernels with variances τ_k^2. The scale-space concept we make use of, will hence be parameterized by a spatial scale σ^2, a temporal scale τ^2 and a set of discrete image velocities $(v_x, v_y)^T$.

The result of applying such velocity-adapted filters to spatio-temporal image data is illustrated in figure 3. Here a synthetic pattern with one spatial and one temporal dimension has been filtered using different values of velocity parameter v. As can be seen, depending on the value of v, the filtering is able to emphasize either the moving pattern (fig. 3(b)) or the stationary background (fig. 3(c)).

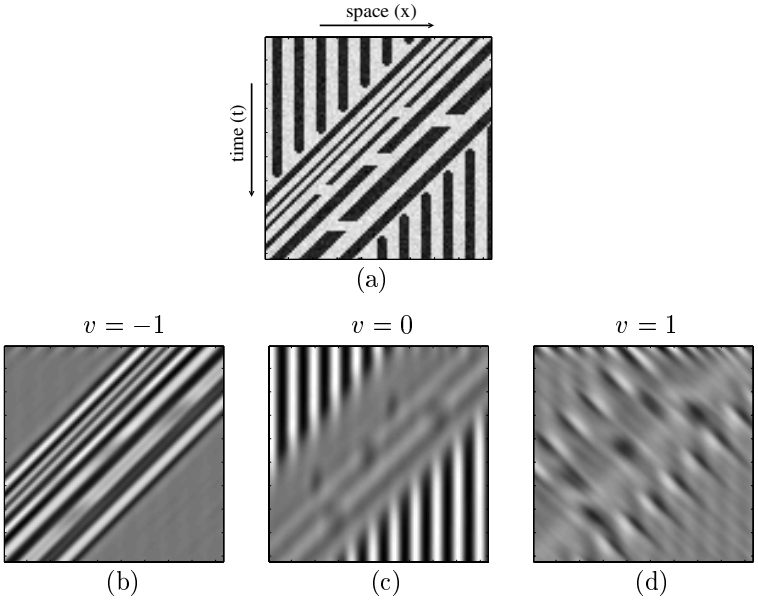

Figure 3: The effect of global velocity adaptation for a synthetic spatio-temporal pattern in (a). (b)-(d): convolution of (a) with spatio-temporal second-order derivative operators with $s_x = 32$, $s_t = 32$ and velocity parameters $v = -1, 0, 1$, respectively. Note, that depending on the velocity parameter, global velocity adaptation emphasizes either the moving pattern (b) or the stationary pattern (c).

3 A mechanism for local velocity adaptation

If we want to interpret events independently of their relative motion to the camera, one approach is to adapt the receptive fields *globally* with respect to the velocity of the events in the field of view. This approach also corresponds to camera stabilization followed by non-adapted filtering. As illustrated in figure 3(b), the result of filtering with globally adapted receptive fields with $v = -1$ indeed enhances the structure of the moving pattern. However, the stationary pattern is suppressed and it follows that global velocity adaptation is not able to handle multiple motions. Moreover, global velocity adaptation is likely to fail if the external velocity information is incorrect (figure 3(d)).

To address these problems, we propose to make use of *local* velocity adaptation of receptive fields. The main idea is to obtain information about motion in the local neighborhood and to use this information for velocity adaptation of receptive fields in the same neighborhood.

Before proceeding to specific schemes for local velocity adaptation in space-time, however, let us observe that there are two main approaches for handling multiple image velocities. One approach is to consider the entire ensemble of receptive fields over image motions as the representation, while the other is to select receptive field outputs corresponding to a single motion estimate. From basic arguments, the first approach can be expected to be more robust in critical situations (compare with biological vision systems), while the second approach followed in this work could be expected to be more accurate and also computationally more efficient on a serial architecture.

The mechanism we will use for accomplishing local velocity adaptation is inspired by related work on automatic scale selection (Lindeberg, 1998) extended to a multi-parameter scale-space (Lindeberg, 1997a) as well as by motion energy approaches for computing optic flow (Adelson and Bergen, 1985; Heeger, 1988). Given a set of image velocities, the normalized Laplacian response is computed for each image velocity in a motion compensated frame (8) in the spatio-temporal scale-space. Then, for each scale, a motion estimate is computed from the velocity $(v_x, v_y)^T$ that maximizes the normalized derivative response

$$(\hat{v}_x, \hat{v}_y)^T(x, y, t)^{(k)} = \mathrm{argmax}_{v_x, v_y} \left(\nabla^2_{norm} L^{(k)}(x, y, t;\ \sigma^2, v_x, v_y) \right)^2, \qquad (9)$$

where $\nabla^2_{norm} = \sigma^2(\partial_{xx} + \partial_{yy})$ is a scale-normalized Laplacian operator in space. This approach is equivalent to the application of a set of velocity-adapted Laplacian operators (figure 4) at each spatio-temporal scale, and selecting the motion estimate from the spatio-temporal filter parameters that gives the maximum response. While one could also consider the use of optic flow estimation schemes for computing the velocity estimates (Nagel and Gehrke, 1998), a main reason why we here consider maximization of normalized receptive field responses over image velocities is that a similar mechanism, when extended to maximization over spatial scales and temporal scales, can also be used for performing simultaneous automatic selection of spatial scales and temporal scales (Lindeberg, 1997b; Lindeberg, 1998).

Figure 5 illustrates the results of local velocity adaptation for a synthetic spatio-temporal pattern (fig. 5(a)) and its Galilean transformation (fig. 5(d)). From the

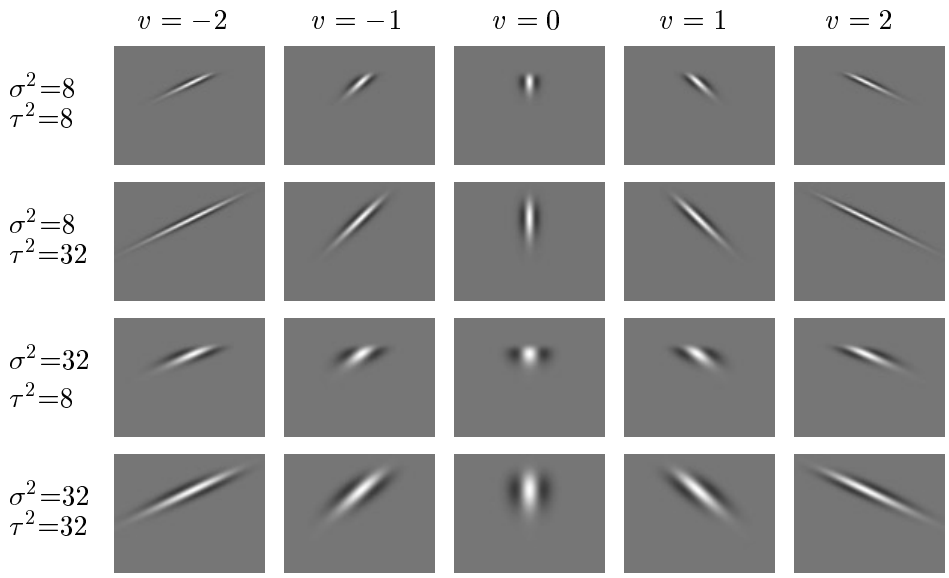

Figure 4: Spatio-temporal filters L_{xx} computed from a velocity-adapted spatio-temporal scale-space for a 1+1–D image pattern, for different values of the velocity parameter v, the spatial scale σ^2 and the temporal scale τ^2.

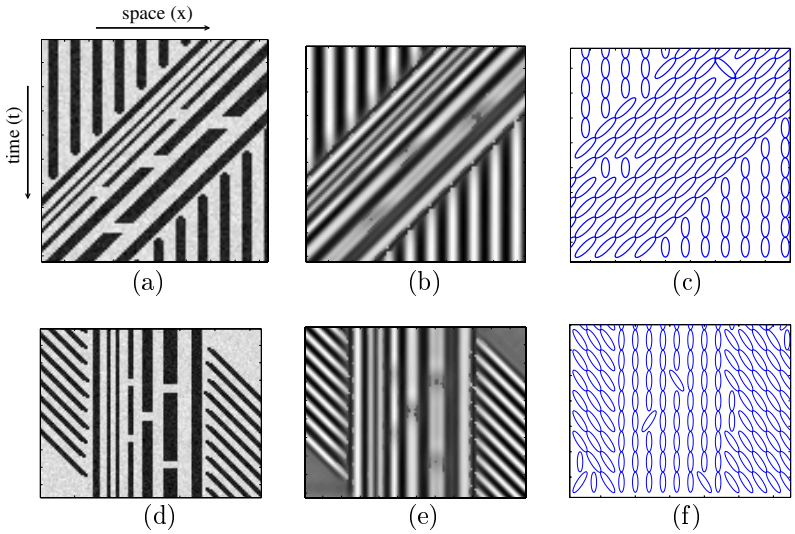

Figure 5: Results of filtering original patterns in (a) and (d) using the proposed *local velocity adaptation* are illustrated in (b) and (e) respectively. The orientation of the ellipses in (c) and (f) show the chosen velocity at each point of the pattern. Note that filtering with local velocity adaptation preserves the details of the moving and stationary pattern. The similarity of the filter responses in (b) and (e) also illustrates the independence of the filtering results with respect to the amount of camera motion.

responses of velocity-adapted receptive fields and from the ellipses displaying the selected orientation of filters in space-time, it is apparent that the proposed filtering scheme adapts to the local motion and enhances structures both in the moving pattern and in the static background. Moreover, by comparing results in figures 5(e) and 5(f), we can visually confirm the invariance of locally adapted receptive field responses with respect to the Galilean transformation of the pattern or, equivalently, to the relative motion between the pattern and the camera.

Application of the local velocity adaptation to a sequence with a walking person is illustrated in figure 6. Note, that filtering here has been done in three dimensions while for the purpose of demonstration, the results are shown only for one $x-t$-slice of a spatio-temporal cube (see fig. 1). As for the synthetic pattern above, we observe successful adaptation of filter kernels to the motion structure of a gait pattern (figures 6(c)-(d)). The results in figures 6(e)-(g) also demonstrate approximative invariance of filter responses with respect to camera motion. The desired effect of the proposed local velocity adaptation is especially evident when these results are compared to the results of separable filtering as shown in figures 1(e)-(g).

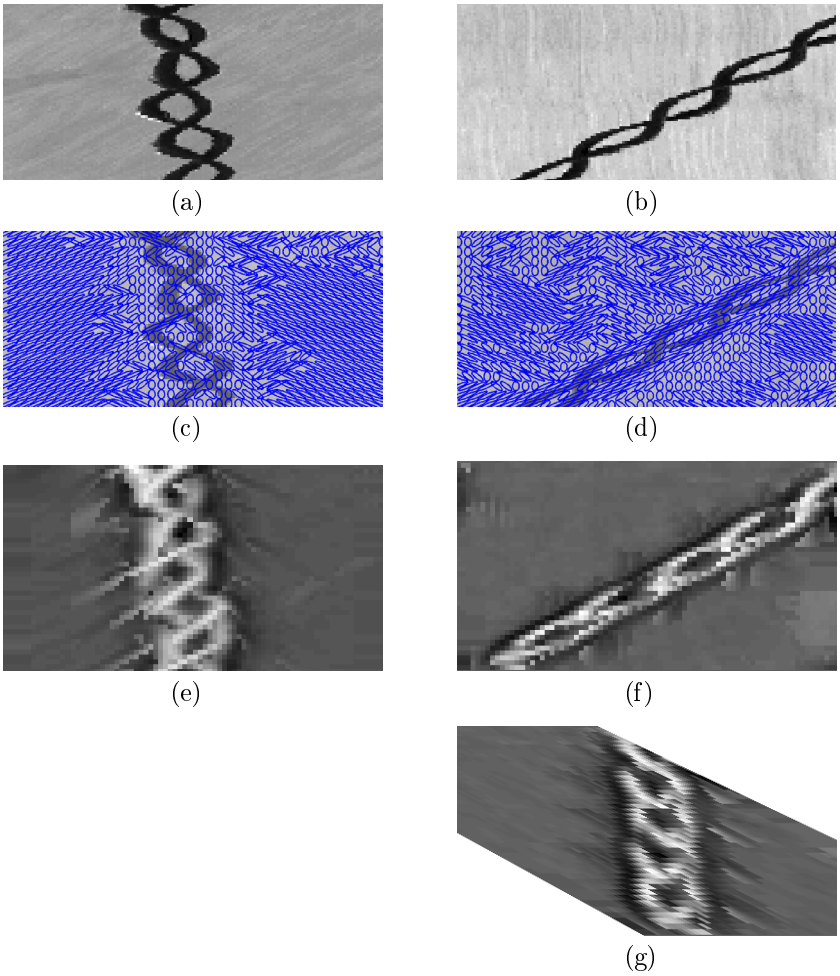

Figure 6: Spatio-temporal filtering with local velocity adaptation applied to a gait pattern recorded with a stabilized camera (a) and a stationary camera (b) (see fig. 1 for comparison). (c)-(d): velocity adapted shape of filter kernels; (e)-(f): results of filtering with a second-order derivative operator; (g): warped version of (f) showing high similarity with (e).

3.1 Comparison with steerable filters

When computing spatio-temporal derivatives, we perform velocity-adaptation of both, the shapes of smoothing kernels and the derivatives according to (5) and (6). An alternative approach that is more efficient but less accurate consists of separable smoothing step followed by adaptation of the derivatives only. Such a scheme is closely related to steerable filters (Freeman and Adelson, 1991) for computing higher-order spatial derivatives in a rotationally invariant way. To differentiate these two approaches, we will refer to them as *velocity-adapted filtering* and *velocity-steered filtering*.

To compare these two alternatives and to illustrate the importance of shape adaptation of filter kernels, we will here compare results of filtering applied to a synthetic prototype of a moving spatio-temporal impulse. The original signal is shown in figure 7(a) in two spatial and one temporal dimensions. Figure 7(b) illustrates the result of computing a partial spatio-temporal derivative ∂_{xxt} using velocity-adapted filtering. With positive and negative filter values represented by different colors, we can visually confirm the correctness of the resulting shape. On the contrary, computation of the same derivative using velocity-steered filtering (fig. 7(c)) results in a different and incorrect shape. A similar result is obtained when filtering is performed without adaptation of neither the smoothing kernels nor the derivatives (fig. 7(d)).

In the next section we apply these filtering schemes to a recognition task and give their quantitative comparison as well as emphasize the importance of velocity-adapted filtering in practice.

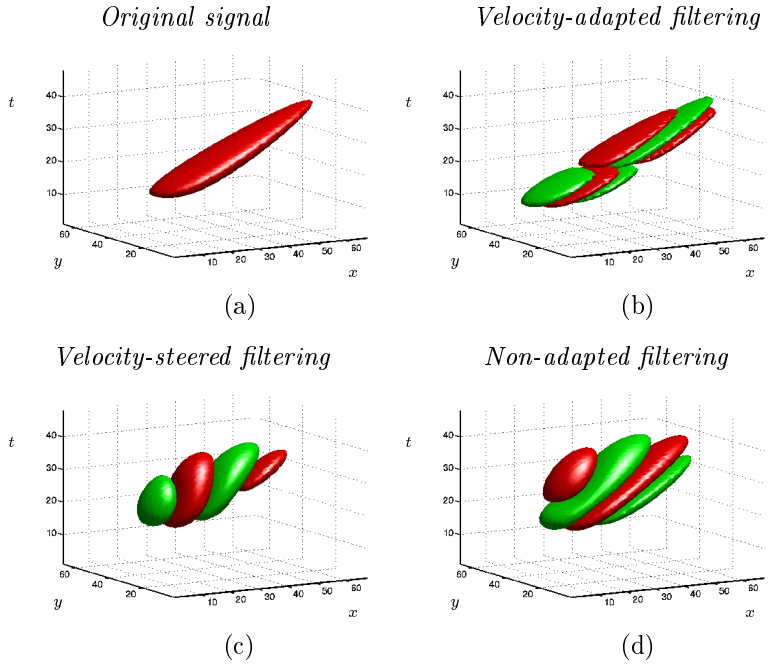

Figure 7: (a): Prototype spatio-temporal blob signal with velocity $v_x = 2$. (b)-(d): Responses to the ∂_{xxt}-derivative operator when using (b): velocity-adapted filters; (c): velocity-steered filters; (d): non-adapted filters. A correct shape of the filter response is obtained only for the case of velocity-adapted filtering.

4 Histogram based recognition

Following (Schiele and Crowley, 2000; Chomat et al., 2000b; Chomat et al., 2000a; Zelnik-Manor and Irani, 2001), let us represent image patterns by histograms of receptive field responses. For this purpose, we use mixed spatio-temporal derivative operators up to order four and collect histograms of these at different spatial and temporal scales. For simplicity, we restrict ourselves to 1-D histograms for each type of filter response. To achieve independence with respect to the direction of motion (left/right or up/down) and the sign of the spatial grey-level variations, we simplify the problem by only considering the absolute values of the filter responses. Moreover, to emphasize the parts of the histograms that correspond to stronger spatio-temporal responses, we also weight the accumulated histograms $H(i)$ by a function $f(i) = i^2$ resulting in $h(i) = i^2 H(i)$.

4.1 Experimental setup

As a test problem we have chosen a data set with image sequences containing people performing actions of type *walking* $W_1 ... W_4$ and *exercise* $E_1 ... E_4$ as illustrated in figure 8. Some of the sequences were taken with a stationary camera, while the others were recorded with a manually stabilized camera. Each of these 4 sec. long sequences were subsampled to a spatio-temporal resolution of $80 \times 60 \times 50$ pixels and convolved with a set of spatio-temporal smoothing kernels for all combinations of seven velocities $v_x = -3...3$, five spatial scales $\sigma^2 = \{2, 4, 8, 16, 32\}$ and five temporal scales $\tau^2 = \{2, 4, 8, 12, 16\}$.

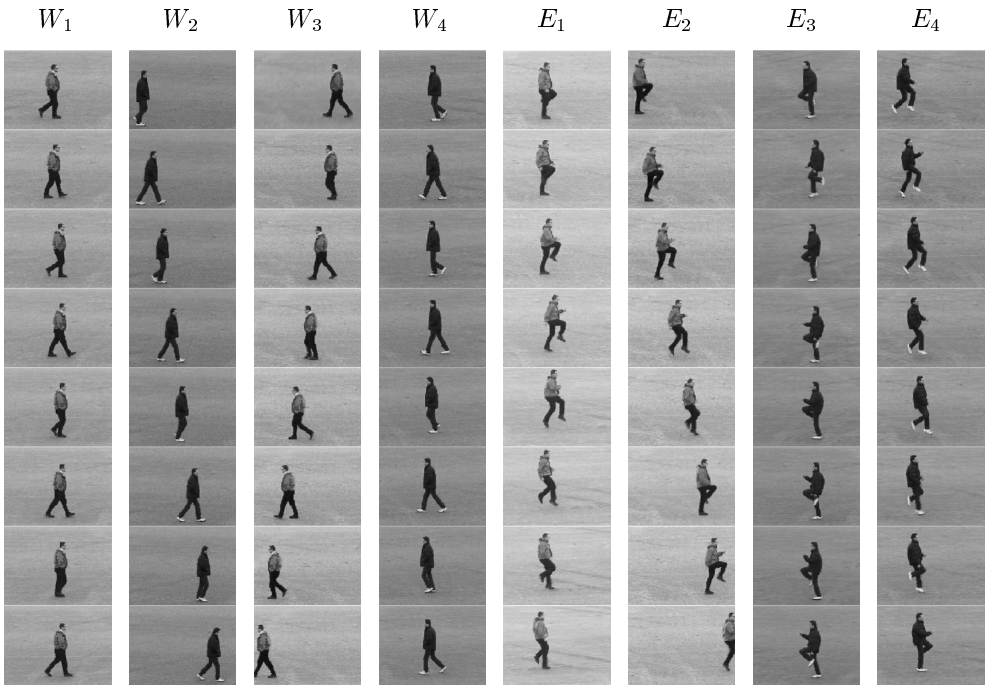

Figure 8: Test sequences of people walking W_1–W_4 and people performing an exercise E_1–E_4. Whereas the sequences W_1,W_4,E_1,E_3 were taken with a manually stabilized camera, the other four sequences were recorded using a stationary camera.

For each spatial scale σ_i, velocity adaptation was performed according to (9) at scale level σ_{i+1}. Since in our examples the relative camera motion was mostly horizontal, we maximized (9) over v_x only. The result of this adaptation for the sequences W_2 and E_1 is illustrated in fig. 9.

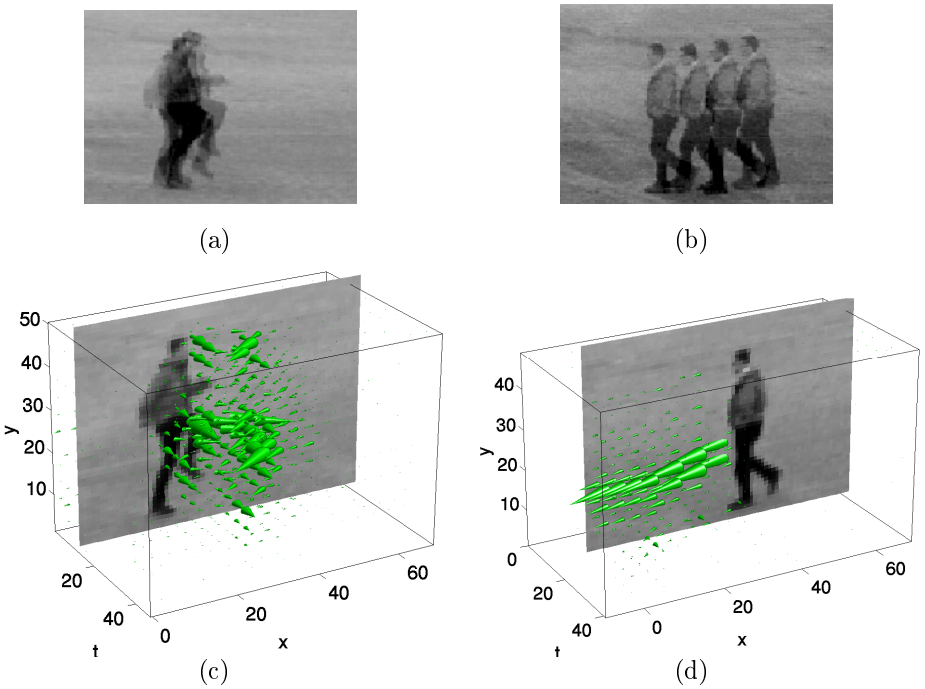

Figure 9: Results of local velocity adaptation for image sequences recorded with a manually stabilized camera (a), and with a stationary camera (b). Directions of cones in (c)-(d) correspond to the velocity chosen by the proposed adaptation algorithm. The size of the cones corresponds the value of the squared Laplacian $\left((\partial_{xx} + \partial_{yy})L(x, y, t; \sigma, \tau)\right)^2$ at the selected velocities.

To represent the patterns, we accumulated histograms of derivative responses for each combination of scales and each type of derivatives. For the purpose of evaluation, separate histograms were accumulated over (i) velocity-adapted derivative responses; (ii) velocity-steered directional derivative responses and (iii) non-adapted partial derivative responses computed at velocity $v = 0$.

4.2 Discriminability of histograms

Figure 10 illustrates the means and the variances of the histograms computed separately for both of the classes. As can be seen from figures 10(a)-(c), velocity-adaptation of receptive fields results in discriminative class histograms and low variation of histograms computed for the same class of activities. On the contrary, the high variations in the histograms in figures 10(d)-(f) and figures 10(g)-(i) clearly indicate that activities are much harder to recognize when using velocity-steered or non-adapted receptive fields.

Whereas figure 10 presents histograms for three types of derivatives L_{xxt}, L_{xyt}

and L_{yyt} at scales $\sigma^2 = 4$, $\tau^2 = 4$ only, we have observed a similar behavior for other derivatives at most of the other scales considered.

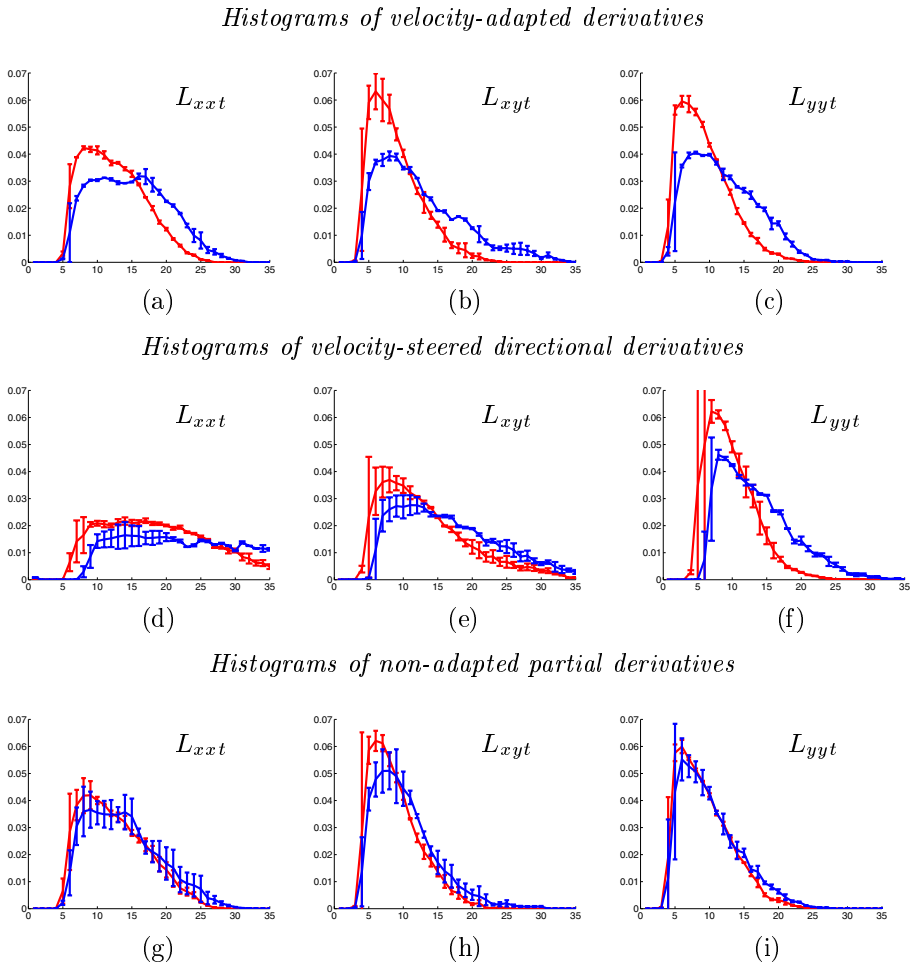

Figure 10: Means and variances of histograms for the activities "walking" (red) and "exercise" (blue). (a)-(c): histograms of *velocity-adapted* derivatives L_{xxt}, L_{xyt}, L_{yyt}; (d)-(f): histograms of velocity-steered directional derivatives L_{xxt}, L_{xyt}, L_{yyt}; (g)-(i): histograms of non-adapted partial derivatives L_{xxt}, L_{xyt}, L_{yyt}. As can be seen, the velocity-adapted filter responses give considerably better possibility to discriminate the motion patterns compared to velocity-steered or non-adapted filters.

4.3 Discriminability measure

To quantify these results, let us measure the distance between pairs of histograms (h_1,h_2) defined according to the χ^2-divergence measure

$$D(h_1, h_2) = \sum_i \frac{(h_1(i) - h_2(i))^2}{h_1(i) + h_2(i)}, \tag{10}$$

where i is the index to the histogram bin. To evaluate the distance between a pair of sequences, we accumulate differences of histograms over different spatial

and temporal scales as well as over different types of receptive fields according to $d(h_1, h_2) = \sum_{l,\sigma,\tau} D(h_1, h_2)$, where l denotes the type of the spatio-temporal filters, σ^2 the spatial scale and τ^2 the temporal scale.

To measure the degree of discrimination between different actions, we compare the distances between pairs of sequences that belong to the same class d_{same} with distances between sequences of different classes d_{diff}. Then, to quantify the average performance of the velocity adaptation algorithm, we compute the mean distances $\bar{d}_{same}$, $\bar{d}_{diff}$ for all valid pairs of examples and define a *distance ratio* according to $r = \bar{d}_{same}/\bar{d}_{diff}$. Hence, low values of r indicate good discriminability, while r close to one corresponds to a performance no better than chance.

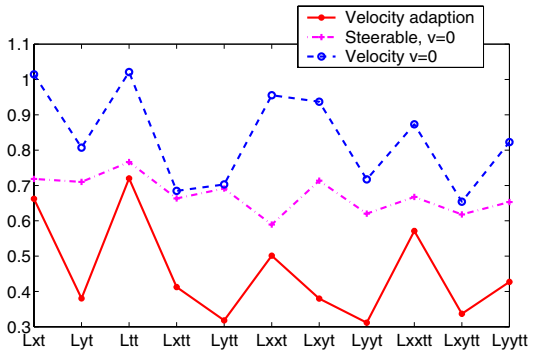

Figure 11: Distance ratios computed for different types of derivatives and for velocity-adapted (solid lines), velocity-steered (point-dashed lines) and non-adapted (dashed lines) filter responses. As can be seen, local velocity adaptation results in lower values of the distance ratio and therefore better recognition performance compared to steered or non-adapted filter responses.

Figure 11 shows distance ratios computed separately for different types of receptive fields. The lower values of the curve corresponding to velocity-adaptation clearly indicate the better recognition performance obtained by using velocity-adapted filters compared to velocity-steered or non-adapted filters. Computing distance ratios over all types of derivatives and scales used, results in the following distance ratios: $r_{adapt} = 0.64$ when using velocity-adapted filters, $r_{steered} = 0.81$ using velocity-steered filters, and $r_{non-adapt} = 0.92$ using non-adapted filters.

4.4 Dependency on scales

When analysing discrimination performance for different types of derivatives and different scales, we have observed an interesting dependency of the distance ratio on the spatial and the temporal scales. Figures 12(a)-(b) show how the distance ratio has a clear minimum over scales at $\sigma^2 = 2$, $\tau^2 = 8$ indicating that these scales give rise to the best discrimination for patterns considered here. In particular, it can be noted that $\tau^2 = 8$ approximately corresponds to the temporal extent of one gait cycle in our examples.

Computation of distance ratios for the selected scale values results in $r_{adapt} = 0.41$ when using velocity-adapted filters, $r_{steered} = 0.71$ using velocity-steered filters

and $r_{non-adapt} = 0.79$ using non-adapted filters. The existence of such preferred scales motivates approaches for automatic selection of both spatial (Lindeberg, 1998) and temporal (Lindeberg, 1997b) scales.

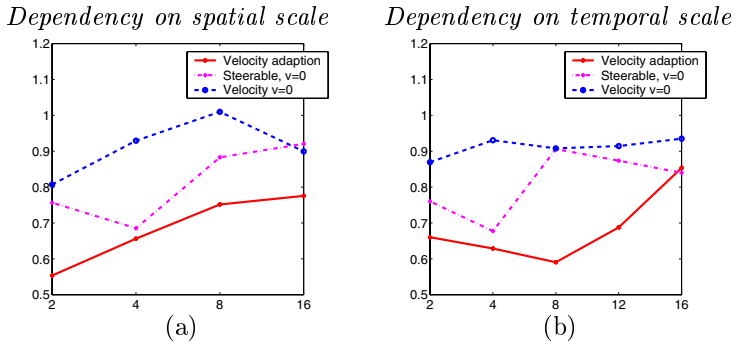

Figure 12: Evolution of the distance ratio r over spatial scales (a) and temporal scales (b). Minima over scales indicate scale values with the highest discrimination ability.

5 Summary and discussion

We have addressed the problem of representing and recognizing events in video in situations where the relative motion between the camera and the observed events is unknown. Experiments on a test problem of recognizing activities show that the use of a velocity adaptation scheme results in a clear improvement in the recognition performance compared to using either (steerable) directional derivatives or regular partial derivatives computed from a non-adapted spatio-temporal filtering step. Whereas for the treated set of examples, recognition could also have been accomplished by using a camera stabilization approach, a major aim here has been to consider a filtering scheme that can be extended to recognition in complex scenes, where reliable camera stabilization may not be possible, i.e. scenes with complex non-static backgrounds or multiple events of interest. Full-fledged recognition in such situations, however, requires more sophisticated statistical methods for recognition than the present histogram-based scheme. We plan to investigate such extensions in future work.

Less restricted to this specific visual task, the results of our investigation also indicate how, when dealing with filter-based representations of spatio-temporal image data, velocity adaptation appears as an essential complement to more traditional approaches of using separable filtering in space-time. For the purpose of performing a clean experimental investigation, we have in this work made use of an explicit velocity-adapted spatio-temporal filtering for each image velocity. While such an implementation has interesting qualitative similarities to biological vision systems (where there are two main classes of receptive fields in space-time — separable filters and non-separable ones (DeAngelis, Ohzawa and Freeman, 1995)), there is a need for developing more sophisticated multi-velocity filtering schemes for efficient implementations in practice.

Finally, future work should also address the problem of selecting appropriate scales in both the spatial and the temporal domains. The preliminary results in section 4.4 indicate the potential of performing joint scale selection in space-time for increasing the recognition performance.

References

Abramowitz, M. and Stegun, I. A. (eds) (1964). *Handbook of Mathematical Functions*, Applied Mathematics Series, 55 edn, National Bureau of Standards.

Adelson, E. and Bergen, J. (1985). Spatiotemporal energy models for the perception of motion, *J. of the Optical Society of America* **A 2**: 284–299.

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.

Ballester, C. and Gonzalez, M. (1998). Affine invariant texture segmentation and shape from texture by variational methods, *J. of Mathematical Imaging and Vision* **9**: 141–171.

Black, M. (1994). Recursive non-linear estimation of discontinuous flow fields, *Proc. Third European Conference on Computer Vision*, Vol. 801 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Stockholm, Sweden, pp. A:138–145.

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.

DeAngelis, G. C., Ohzawa, I. and Freeman, R. D. (1995). Receptive field dynamics in the central visual pathways, *Trends in Neuroscience* **18**(10): 451–457.

Florack, L. M. J. (1997). *Image Structure*, Series in Mathematical Imaging and Vision, Kluwer Academic Publishers, Dordrecht, Netherlands.

Florack, L., Niessen, W. and Nielsen, M. (1998). The intrinsic structure of optic flow incorporating measurement duality, *Int. J. of Computer Vision* **27**(3): 263–286.

Freeman, W. T. and Adelson, E. H. (1991). The design and use of steerable filters, *IEEE Trans. Pattern Analysis and Machine Intell.* **13**(9): 891–906.

Guichard, F. (1998). A morphological, affine, and galilean invariant scale-space for movies, *IEEE Trans. Image Processing* **7**(3): 444–456.

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.

Heeger, D. (1988). Optical flow using spatiotemporal filters, *Int. J. of Computer Vision* **1**: 279–302.

Irani, M., Anandan, P. and Hsu, S. (1995). Mosaic based representations of video sequences and their applications, *Proc. Fifth International Conference on Computer Vision*, Cambridge, MA, pp. 605 –611.

Koenderink, J. J. (1984). The structure of images, *Biological Cybernetics* **50**: 363–370.

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

Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*, The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Dordrecht, Netherlands.

Lindeberg, T. (1997a). Linear spatio-temporal scale-space, *in* B. M. ter Haar Romeny, L. M. J. Florack, J. J. Koenderink and M. A. Viergever (eds), *Scale-Space Theory in Computer Vision: Proc. First Int. Conf. Scale-Space'97*, Vol. 1252 of *Lecture Notes in Computer Science*, Springer Verlag, New York, Utrecht, The Netherlands, pp. 113–127. Extended version available as Technical Report ISRN KTH/NA/P–01/22–SE from KTH (http://www.nada.kth.se/cvap/abstracts/cvap257.html).

Lindeberg, T. (1997b). On automatic selection of temporal scales in time-casual scale-space, *in* G. Sommer and J. J. Koenderink (eds), *Proc. AFPAC'97: Algebraic Frames for the Perception-Action Cycle*, Vol. 1315 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Kiel, Germany, pp. 94–113.

Lindeberg, T. (1998). Feature detection with automatic scale selection, *Int. J. 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. and Fagerström, D. (1996). Scale-space with causal time direction, *Proc. 4th European Conf. on Computer Vision*, Vol. 1064, Springer Verlag, Berlin, Cambridge, UK, pp. 229–240.

Lindeberg, T. and Garding, J. (1994). Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D structure, *Proc. Third European Conference on Computer Vision*, Stockholm, Sweden, pp. A:389–400.

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.

Nagel, H. and Gehrke, A. (1998). Spatiotemporal adaptive filtering for estimation and segmentation of optical flow fields, *Proc. Fifth European Conference on Computer Vision*, Freiburg, Germany, pp. II:86–102.

Schaffalitzky, F. and Zisserman, A. (2001). Viewpoint invariant texture matching and wide baseline stereo, *Proc. 8th Int. Conf. on Computer Vision*, Vancouver, Canada, pp. II:636–643.

Schiele, B. and Crowley, J. (2000). Recognition without correspondence using multidimensional receptive field histograms, *International Journal of Computer Vision* **36**(1): 31–50.

Schneiderman, H. and Kanade, T. (2000). A statistical method for 3D object detection applied to faces and cars, *Proc. Computer Vision and Pattern Recognition*, Vol. I, Hilton Head, SC, pp. 746–751.

Swain, M. and Ballard, D. (1991). Color indexing, *International Journal of Computer Vision* **7**(1): 11–32.

Weickert, J. (1998). *Anisotropic Diffusion in Image Processing*, Teubner-Verlag, Stuttgart, Germany.

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.