Galilean-diagonalized spatio-temporal interest operators [*]

Tony Lindeberg, Amir Akbarzadeh and Ivan Laptev

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

Abstract. *This paper presents a set of image operators for detecting regions in space-time where interesting events occur. To define such regions of interest, we compute a spatio-temporal second-moment matrix from a spatio-temporal scale-space representation, and diagonalize this matrix locally, using a local Galilean transformation in space-time, optionally combined with a spatial rotation, so as to make the Galilean invariant degrees of freedom explicit. From the Galilean-diagonalized descriptor so obtained, we then formulate different types of space-time interest operators, and illustrate their properties on different types of image sequences.*

1. Introduction

For analysing the space-time structure of our environment, the ability to detect regions of interest is an important preprocessing stage for subsequent recognition. The presumably simplest approach for constructing such a mechanism is by regular frame differencing. The result of frame differencing will, however, be very sensitive to the time interval used for computing the differences. Moreover, such an operator will be sensitive to motions relative to the camera.

An interesting approach for defining regions of interest for motion patterns was taken by (Davis & Bobick 1997), who computed multiple temporal differences and constructed a motion mask, which was then represented in terms of moment descriptors to characterize the motion. This approach, however, assumes a static background as well as a stationary camera.

A general problem when interpreting spatio-temporal image data originates from the fact that motion descriptors are affected by relative motions between the objects and the camera. It is therefore essential to aim at Galilean invariant image descriptors. One approach to achieve Galilean invariance is to consider space-time receptive fields adapted to local motion directions (Lindeberg 2002). A dual approach is to stabilize the space-time pattern locally, assuming that the scene contains cues that allow for stabilization. In the spatio-temporal recognition scheme developed by (Zelnik-Manor & Irani 2001), global stabilization was used when computing the spatio-temporal derivatives. (Laptev & Lindeberg 2004b) extended this approach to recognition based on locally velocity adapted space-time filters.

The subject of this paper is to develop a set of space-time interest operators, which build upon several of the above-mentioned ideas, with emphasis on locally compensating for relative motions between the world and the observer. These operators are intended as region-of-interest operators for subsequent recognition of spatio-temporal events, in a corresponding manner as spatial interest points are used as a preprocessing stage for spatial recognition (Lowe 1999, Mikolajczyk & Schmid 2002), see also the related notion of space-time interest points in (Laptev & Lindeberg 2003). The operators to be presented will be closely related to previously developed methods for computing spatio-temporal energy (Adelson & Bergen 1985, Wildes & Bergen 2000) or curvature descriptors (Zetzsche & Barth 1991, Niyogis 1995) in space-time, with specific emphasis on achieving invariance to local Galilean transformations.

Besides the specific topic of spatio-temporal interest operators, we shall also introduce a more general notion of Galilean diagonalization, to make explicit the Galilean invariant degrees of freedom in a spatio-temporal second-moment matrix, as a complement to the more traditional notion of eigenvalue based analysis of spatio-temporal second-moment matrices (Bigün et al. 1991, Jähne 1995).

2. Spatio-temporal scale-space

Let $p = (x, y, t)^T$ denote a point in 2+1-D space-time, and let $f\colon \mathbb{R}^3 \to \mathbb{R}$ represent a spatio-temporal image. Following (Lindeberg 1997, Lindeberg 2002), consider a multi-parameter scale-space $L\colon \mathbb{R}^3 \times \mathbb{G} \to \mathbb{R}$ of f defined by convolution with a family $h\colon \mathbb{R}^3 \times \mathbb{G} \to \mathbb{R}$ of spatio-temporal scale-space kernels

$$L(\cdot;\ \Sigma) = h(\cdot;\ \Sigma) * f(\cdot)$$

parameterized by covariance matrices Σ in a semi-group $\mathbb{G}$. The covariance matrices may in turn be parameterized as

$$\Sigma = \begin{pmatrix} \lambda_1 c^2 + \lambda_2 s^2 + u^2 \lambda_t & (\lambda_2 - \lambda_1) c\, s + uv\lambda_t & u\lambda_t \\ (\lambda_2 - \lambda_1) c\, s + uv\lambda_t & \lambda_1 s^2 + \lambda_2 c^2 + v^2 \lambda_t & v\lambda_t \\ u\lambda_t & v\lambda_t & \lambda_t \end{pmatrix}$$

where $(\lambda_1, \lambda_2, c, s)$ describe the amount of spatial (possibly anisotropic) smoothing in terms of two eigenvalues and

[*] 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.

their orientation α in space with $c = \cos\alpha$ and $s = \sin\alpha$, λ_t gives the amount of temporal smoothing, and (u, v) describes the orientation of the filter in space-time. In the special case when $\lambda_1 = \lambda_2$ and $(u, v) = (0, 0)$, this multi-parameter scale-space reduces to the scale-space obtained by space-time separable smoothing with spatial scale parameter $\sigma^2 = \lambda_1 = \lambda_2$ and temporal smoothing $\tau^2 = \lambda_t$.

For simplicity, we shall here model the smoothing operation by a 3-D Gaussian kernel with covariance matrix Σ, $h(p;\ \Sigma) = g(x;\ \Sigma) = \exp(-x^T \Sigma^{-1} x/2)/((2\pi)^{3/2}\sqrt{\det\Sigma})$, for which the space-time separable case reduces to convolution with a 2-D Gaussian $g_{2D}(x, y;\ \sigma^2) = 1/(2\pi\sigma^2)\exp(-(x^2 + y^2)/2\sigma^2)$ in space and a 1-D Gaussian $g_{1D}(t;\ \tau^2) = 1/(\sqrt{2\pi}\tau)\exp(-t^2/2\tau^2)$ over time.

Second-moment descriptor. For describing local image structures and for estimating local image deformations, the second-moment matrix (Bigün et al. 1991, Jähne 1995, Lindeberg & Gårding 1997) is a highly useful descriptor. In 2+1-D space-time, it can be defined as

$$\mu(p;\ \Sigma) = \int_{q \in \mathbb{R}^3} (\nabla L(q))(\nabla L(q))^T\, w(p - q;\ \Sigma)\, dq,$$

where $\nabla L = (L_x, L_y, L_t)^T$ denotes the spatio-temporal gradient, and w is a spatio-temporal window function, for simplicity a Gaussian function, with covariance matrix Σ,

$$\begin{pmatrix} \mu_{xx} & \mu_{xy} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix} = \int \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix} dw.$$

Galilean transformations. Given a spatio-temporal image f, consider a Galilean transformation of space-time

$$p' = \begin{pmatrix} x' \\ y' \\ t' \end{pmatrix} = Gp = \begin{pmatrix} 1 & 0 & -u \\ 0 & 1 & -v \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ t \end{pmatrix}$$

and define a transformed im1¡age according to $f'(p') = f(p)$. Define scale-space representations of f and f' according to $L(\cdot;\ \Sigma) = g(\cdot;\ \Sigma) * f(\cdot)$ and $L'(\cdot;\ \Sigma') = g(\cdot;\ \Sigma') * f'(\cdot)$. Then, it can be shown (Lindeberg 2002) that $L'(\cdot;\ \Sigma') = L(\cdot;\ \Sigma)$ if $\Sigma' = G\Sigma G^T$. Next, let us define a transformed second-moment matrix as

$$\mu'(p';\ \Sigma') = \int_{q' \in \mathbb{R}^3} (\nabla L'(q'))(\nabla L'(q'))^T\, w(p' - q';\ \Sigma')\, dq'.$$

Then, from a general transformation property of second-moment descriptors under linear transformations (Lindeberg & Gårding 1997), it can be shown that μ and μ' are related according to

$$\mu' = G^{-T} \mu G^{-1} \quad \text{provided that} \quad \Sigma' = G\Sigma G^T$$

In terms of the components of μ and μ', we have

$$\mu'_{xx} = \mu_{xx}, \quad \mu'_{xy} = \mu_{xy}, \quad \mu'_{yy} = \mu_{yy},$$
$$\mu'_{xt} = u\mu_{xx} + v\mu_{xy} + \mu_{xt}, \quad \mu'_{yt} = u\mu_{xy} + v\mu_{yy} + \mu_{yt},$$
$$\mu'_{tt} = u^2\mu_{xx} + 2uv\mu_{xy} + v^2\mu_{yy} + 2u\mu_{xt} + 2v\mu_{yt} + \mu_{tt}.$$

3. Galilean diagonalization

Our goal is to define spatio-temporal image descriptors that are stable under relative motions between the camera and the scene. In this aim towards Galilean invariance, we shall follow a specific convention of determining the velocity components (u, v) in a local Galilean transformation G, such that the transformed second moment matrix μ' assumes a block diagonal form with $(\mu'_{xt}, \mu'_{yt}) = (0, 0)$:

$$\mu' = G^{-T} \mu G^{-1} = \begin{pmatrix} \mu'_{xx} & \mu'_{xy} & 0 \\ \mu'_{xy} & \mu'_{yy} & 0 \\ 0 & 0 & \mu'_{tt} \end{pmatrix}$$

This form of block diagonalization of a spatio-temporal second-moment matrices can be seen as a canonical way of extracting a unique representative of the family of second-moment matrices $\mu' = G^{-T}\mu G^{-1}$ that will be obtained if we for a given spatio-temporal pattern consider the whole group of Galilean transformations G of space-time that represents all possible relative motions with constant velocity between the camera and the scene. Specifically, this form of block diagonalization implies a local normalization of local space-time structures that is invariant under superimposed Galilean transformations (Lindeberg et al. 2004, Appendix A.1). It follows from the transformation property of μ, that block diagonalization is achieved if (u, v) satisfies

$$\begin{pmatrix} \mu_{xx} & \mu_{xy} \\ \mu_{xy} & \mu_{yy} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = -\begin{pmatrix} \mu_{xt} \\ \mu_{yt} \end{pmatrix}$$

i.e., structurally similar equations as for computing optic flow according to (Lukas & Kanade 1981). Hence, if the local space-time structures represent a pure translational model, the result of Galilean diagonalization will be a stationary pattern. The same form of normalization, however, also applies to spatio-temporal events that cannot be modelled by a pure translational model. In the latter case, the result of this normalization will be a local spatio-temporal pattern that satisfies

$$\int_{x,y,t \in \mathbb{R}^3} L_x\, L_t\, g(x, y, t;\ \Sigma)\, dx\, dy\, dt = 0,$$
$$\int_{x,y,t \in \mathbb{R}^3} L_y\, L_t\, g(x, y, t;\ \Sigma)\, dx\, dy\, dt = 0.$$

In other words, after Galilean diagonalization the elements L_x, L_y and L_t in the local spatio-temporal pattern will be scattered according to a non-biased distribution, such that the spatial and temporal derivatives are locally uncorrelated with respect to (here) a Gaussian window function. In situations when the constant brightness assumption is satisfied, there is an interpretation of this property in terms of the weighted average of local normal flow vectors $(u_{\|}, v_{\|})$ being zero, using the product of the window function and the magnitude of the spatial gradient vector $\nabla_{space} L = (L_x, L_y)^T$ as weight (Lindeberg et al. 2004, Appendix A.2):

$$E\left((\nabla_{space} L)(\nabla_{space} L)^T \begin{pmatrix} u \\ v \end{pmatrix} \right) =$$

$$= E\left(|\nabla_{space} L|^2 \begin{pmatrix} u_{\|} \\ v_{\|} \end{pmatrix} \right) = 0.$$

In this respect, Galilean diagonalization implies cancelling of the average velocity also for spatio-temporal events that cannot be locally modelled by a Galilean transformation.

Given that we have block diagonalized μ', we can continue with a 2-D rotation R_{space} that diagonalizes the remaining spatial components such that $\mu''_{xy} = 0$. In other words, given any second moment matrix μ, we can determine a Galilean transformation G in space-time and a rotation R_{space} in space, such that

$$\mu'' = R_{space}^{-T} G^{-T} \mu \, G^{-1} R_{space}^{-1} = \begin{pmatrix} \nu_1 & & \\ & \nu_2 & \\ & & \nu_3 \end{pmatrix}$$

where (ν_1, ν_2, ν_3) are the diagonal elements. There is a close structural similarity between such a Galilean/rotational diagonalization and the more common approach of using the eigenvalues of a spatio-temporal second moment matrix for motion analysis (Bigün et al. 1991, Jähne 1995). An eigenvalue-based analysis of μ corresponds to determining a rotation matrix U such that

$$\mu''' = U^{-T} \mu \, U^{-1} = \begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3 \end{pmatrix}$$

There is, however, no physical correspondence to a rotation in 2+1-D space-time. For a second-moment matrix defined over a 3-D space (x, y, z), an eigenvalue analysis has a clear physical interpretation, since it corresponds to determining a 3-D rotation in space, such that μ will be a diagonal matrix with the eigenvalues as entries. If similar algebraic manipulations are applied to a second-moment matrix over space-time, however, there is no physical analogue. For this reason, we propose that a Galilean/rotational transformation is a more natural concept for diagonalizing a spatio-temporal second-moment matrix. This type of matrix diagonalization is also easy to compute in closed form.

4. Spatio-temporal interest operators

The notion of Galilean diagonalization can be used for defining spatio-temporal image descriptors that are either fully invariant or approximately invariant under Galilean transformations. Operators within the first class will be referred to as Galilean invariant, while operators within the latter class will be referred to as Galilean corrected.

The context we consider is that the spatio-temporal second moment matrix is computed at every point p in space-time for a set of scale parameters Σ. Two main approaches can be considered: (i) Consider the full family of spatio-temporal scale-space kernels, parameterized over both the amount of spatial smoothing, the amount of temporal smoothing, and the orientation of the filter in space-time. (ii) Restrict the analysis to space-time separable scale-space kernels only. A motivation for using the first approach is that the spatio-temporal scale-space will be truly closed under Galilean transformations only if the full family of covariance matrices is considered. Thus, this alternative has advantages in terms of robustness and accuracy, while the second alternative will be more efficient on a serial architecture. In the first case, (ν_1, ν_2, ν_3) will be truly Galilean invariant, while in the second case the effect of the Galilean diagonalization is to compensate for the motion relative to the camera. In comparison with the related notions of affine shape-adaptation in space (Lindeberg & Gårding 1997, Mikolajczyk & Schmid 2002) or velocity adaptation in space-time (Nagel & Gehrke 1998, Lindeberg 2002, Laptev & Lindeberg 2004b), we can interpret the combination of Galilean diagonalization with space-time separable scale-space as an estimate of the first step in an iterative velocity adaptation procedure.

Galilean diagonalized motion descriptors. A first approach we shall follow is to use $I_1 = \nu_3 = \mu'_{tt}$ as a basic measure for computing candidate regions of interest. If the space-time structures are locally constant over time, or if the local space-time structure corresponds to a translation with constant velocity, then in the ideal case (of using velocity adapted filters) the value of this descriptor will be zero. Hence, I_1 can be regarded as a measure of how much the local space-time image structure deviates from a pure translation. Note that compared to a more traditional stabilization scheme, there is no need for warping the space-time image according to a local motion estimate. Instead, we use the closed-form expression for (u, v) for evaluating I_1 from the components of μ at every point according to

$$I_1 = \mu'_{tt} = \mu_{tt} - \frac{\mu_{xx}\mu_{yt}^2 + \mu_{yy}\mu_{xt}^2 - 2\mu_{xy}\mu_{xt}\mu_{yt}}{\mu_{xx}\mu_{yy} - \mu_{xy}^2}$$

The operator I_1 will respond to rather wide classes of space-time events. If we are interested in more restrictive space-time interest operators, we can, for example, consider two

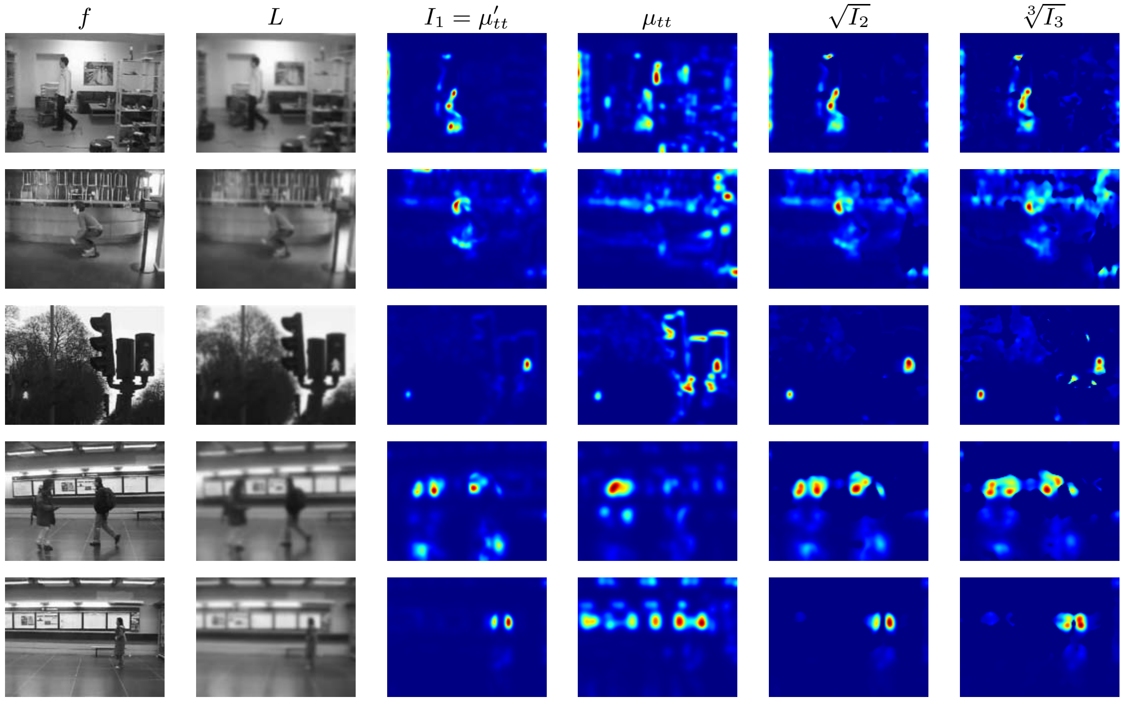

Figure 1: *Maps of the Galilean-diagonalized interest operators I_1, I_2 and I_3, computed from space-time separable spatio-temporal scale-space representations L of different image sequences f. From top to bottom: (i) walking person, (ii) jumping person, (iii) pedestrian lights turning green, (iv) two walking persons with camera stabilized on right person, (v) walking person with camera stabilized on the person.*

extensions of the Harris operator (Harris & Stephens 1988) to space-time that will be described below. Given a spatial second moment matrix μ_{2D} with eigenvalues (λ_1, λ_2), the traditional Harris operator is defined as

$$H = \lambda_1 \lambda_2 - C(\lambda_1 + \lambda_2)^2 = \det \mu_{2D} - C(\operatorname{trace} \mu_{2D})^2$$

where C is usually chosen as $C = 0.04$, and values of H below zero are thresholded away. For images on a 2-D spatial domain, this operator will give high responses if both the eigenvalues of μ_{2D} are high, and the image thus contains significant variations along both of the two dimensions.

We can build upon this idea for defining two space-time operators of different forms, either by treating the spatial dimensions together or separately. By treating the spatial diagonal elements together, it is natural to let $\lambda_1 = \nu_1 + \nu_2$ and $\lambda_2 = \nu_3$, and we can define an operator of the form

$$\begin{aligned} I_2 &= (\nu_1 + \nu_2)\nu_3 - C_2(\nu_1 + \nu_2 + \nu_3)^2, \\ &= (\mu_{xx} + \mu_{yy})\mu'_{tt} - C_2(\mu_{xx} + \mu_{yy} + \mu'_{tt})^2 \end{aligned}$$

By treating all the diagonal elements individually, we can define the following modification of the spatio-temporal interest operator in (Laptev & Lindeberg 2003)

$$\begin{aligned} I_3 &= \nu_1 \nu_2 \nu_3 - C_3(\nu_1 + \nu_2 + \nu_3)^3 \\ &= (\mu_{xx}\mu_{yy} - \mu_{xy}^2)\mu'_{tt} - C_3(\mu_{xx} + \mu_{yy} + \mu'_{tt})^3 \end{aligned}$$

In both cases, C_2 and C_3 are parameters to be determined. Initially, we use $C_2 = 0.04$ and $C_3 = 0.005$ in analogy with (Harris & Stephens 1988, Laptev & Lindeberg 2003). The requirement for I_1 to respond is that there are significant variations in the image structures over the temporal dimension beyond those that can be described by a local translation model. For I_2 to respond, it is necessary that there are strong image variations over at least one spatial dimension in addition to the temporal dimension. For I_3 to respond, there must be significant variations over both of the two spatial dimensions in addition to the temporal dimension.

5. Experiments

Figure 1 shows snapshots of computing I_1, I_2 and I_3 for different types of spatio-temporal image patterns. For comparison, we also show μ_{tt} without Galilean-correction, as well as sample frames from the original image sequence f and its spatio-temporal scale-space representation L, for simplicity computed by space-time separable filtering.

The rows show from top to bottom: (i) a walking person with approximately stabilized camera, (ii) a person jumping with the camera slowly following the person, (iii) pedestrian lights turning green, (iv) two people walking in different directions with the camera stabilized on the right person, (v) a walking person with the camera stabilized on the person. All sequences have been taken with a handheld camera.

As can be seen from the results, there is a substantial difference between the output from the Galilean diagonalized $I_1 = \mu'_{tt}$ and the corresponding non-diagonalized entry μ_{tt}, with I_1 being much more specific to motion events in the scene. For the pedestrian light scene, a small camera motion results in responses of μ_{tt} at object edges, while μ'_{tt} gives relatively stronger responses to the lights switching to green. In the case of two persons walking in different directions, $I_1 = \mu'_{tt}$ gives responses of similar magnitude for the two persons, while for μ_{tt} the response of one person dominates. In the case of a walking person against a moving background (the camera following the person), the built-in Galilean correction in $I_1 = \mu_{tt}$ effectively suppresses a major part of the background motion compared to μ_{tt}. In comparison with I_1, the operators I_2 and I_3 give somewhat stronger responses at edges and corners, respectively.

To quantitatively evaluate the stability of these descriptors under relative motions, we subjected a set of image sequences to synthetic Galilean transformations $u \in \{1, 2, 3\}$, and computed the following correlation error measure

$$E(M) = C(M_f, M_{G_u f}) = \frac{\sum_{p \leftrightarrow p'} (M_f(p) - M_{G_u f}(p'))^2}{\sqrt{\sum_p M_f(p)^2}\sqrt{\sum_{p'} M_{G_u f}(p')^2}}$$

between the maps M_f and $M_{G_u f}$ of these descriptors computed from the original image sequence f as well as its corresponding Galilean transformed image sequence $G_u f$ at corresponding points $p \leftrightarrow p'$ in space-time (see table 1). As can be seen, the Galilean-corrected spatio-temporal interest operators I_1, I_2 and I_3 give a better approximation to Galilean invariance than the corresponding non-corrected entities $\tilde{I}_1$, $\tilde{I}_2$ and $\tilde{I}_3$.

$C(M_f, M_{G_u f})$	I_1	$\tilde{I}_1$	I_2	$\tilde{I}_2$	I_3	$\tilde{I}_3$
$u = 1$	0.03	0.07	0.03	0.05	0.06	0.51
$u = 2$	0.11	0.31	0.10	0.11	0.19	1.17
$u = 3$	0.21	0.77	0.20	0.18	0.36	2.13

$C(M_f, M_{G_u f})$	I_1	$\tilde{I}_1$	I_2	$\tilde{I}_2$	I_3	$\tilde{I}_3$
$u = 1$	0.08	0.30	0.06	0.27	0.08	0.48
$u = 2$	0.27	1.44	0.26	0.71	0.31	1.22
$u = 3$	0.44	1.63	0.36	0.89	0.40	1.49

Table 1: *Correlation error measures between interest operators responses under Galilean transformations for two image sequences.*

Then, we formed ratios between these measures of deviations from Galilean invariance for (I_1, I_2, I_3) and their corresponding non-diagonalized descriptors $(\tilde{I}_1, \tilde{I}_2, \tilde{I}_3)$; the geometric average and the geometric standard deviations for seven image sequences are given in table 2.

For this data set, the use of Galilean diagonalization reduced the correlation errors with factors typically in the range between 2 and 5, depending on the image contents and the type of descriptor. As can be seen from the results, the ratio between the error measures for Galilean-corrected as opposed to corresponding uncorrected entities is largest for small image velocities and decreases with increasing velocity, indicating that in combination with a space-time separable smoothing kernels, the relative compensatory effect of Galilean-diagonalization is largest for small image velocities and decreases with increasing image velocity.

velocity	$E(\tilde{I_1})/E(I_1)$	$E(\tilde{I_2})/E(I_2)$	$E(\tilde{I_3})/E(I_3)$
$u = 1$	3.2 (2.2)	4.5 (3.4)	4.6 (2.2)
$u = 2$	3.2 (1.6)	2.5 (2.3)	3.8 (1.9)
$u = 3$	2.6 (1.4)	1.7 (2.0)	3.9 (1.6)
all u	3.0 (1.7)	2.7 (2.5)	4.1 (1.9)

Table 2: *Ratios between Galilean correlation errors for Galilean-diagonalized vs. corresponding non-diagonalized descriptors computed from a space-time separable spatio-temporal scale-space.*

6. Extension to colour cues

With minor modifications, we can apply corresponding ideas to colour images, to make use of the additional information available in colour channels if there is poor contrast in the pure grey-level information. Based on a derivation in (Lindeberg et al. 2004, Section 7.1), we

1. compute second moment matrices $\mu^{(i)}$ for all individual colour channels,
2. sum up the elements in these in order to form:

$$A = \sum_i A^{(i)} = \begin{pmatrix} \mu_{xx}^{(i)} & \mu_{xy}^{(i)} \\ \mu_{xy}^{(i)} & \mu_{yy}^{(i)} \end{pmatrix},$$

$$b = \sum_i b^{(i)} = \begin{pmatrix} \mu_{xt}^{(i)} \\ \mu_{yt}^{(i)} \end{pmatrix},$$

3. compute a joint velocity estimate $u = (u_x, u_y)^T$ according to $u = -A^{-1}b$,
4. for each colour channel insert this estimate into the expression for $(\mu'_{tt})^{(i)}$, analogous to μ'_{tt},
5. sum up these entities over all colour channels to define the following analogue of the purely temporal diagonal element: $\nu_3 = \sum_i (\mu'_{tt})^{(i)}$,
6. compute analogues to the spatial diagonal elements ν_1 and ν_2 from $\nu_1 + \nu_2 = \operatorname{trace} A$ and $\nu_1 \nu_2 = \det A$,
7. define I_1, I_2 and I_3 from ν_1, ν_2 and ν_3 in analogy with the previously stated equations.

Figure 2 shows examples of computing spatio-temporal interest operators in this way. As can be seen, the use of complementary colour cues may give more prominent regions of interest if there is poor contrast in the pure grey-level cues.

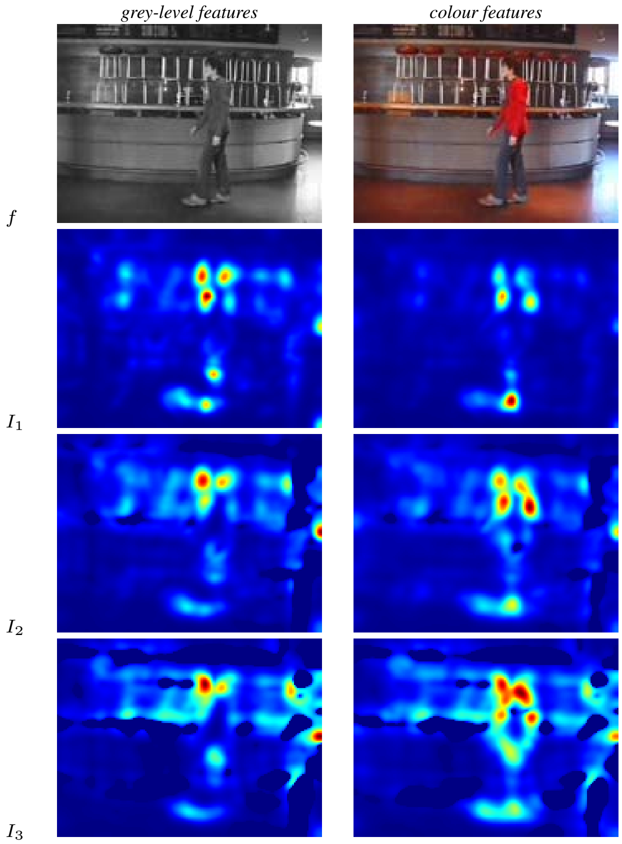

Figure 2: *The result of computing grey-level based as well as colour-based spatio-temporal interest operator responses for an image sequence with a walking person against a cluttered background, in which there is sometimes poor grey-level contrast between the moving object and the background.*

7. Summary and discussion

We have presented a theory for how Galilean diagonalization can be used for reducing the influence of local relative motions on spatio-temporal image descriptors, and used this theory for defining a set of spatio-temporal interest operators. In combination with velocity-adapted scale-space filtering, these image descriptors are truly Galilean invariant. Combined with space-time separable filtering, they allow for a substantial reduction of the influence of Galilean motions. In this respect, these operators allow for more robust regions of interest under relative motions of the camera.

Besides the application to spatio-temporal interest operators considered here, the notion of Galilean diagonalization is of wider applicability and should be regarded as an interesting conceptual tool also in the following contexts: (i) as an alternative to local eigenvalue analysis of space-time image structures, (ii) when extracting spatio-temporal features, and (iii) when performing local normalization of space-time structures for subsequent spatio-temporal recognition.

An integration of Galilei-diagonalization with velocity adaptation for detecting Galilean invariant spatio-temporal interest points is presented in (Laptev & Lindeberg 2004a), including a more extensive evaluation. For real-time processing, the scale-space model used here can be extended to the time-causal scale-space concepts in (Koenderink 1988, Lindeberg & Fagerström 1996, Lindeberg 1997, Florack 1997, ter Haar Romeny et al. 2001, Lindeberg 2002).

References

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

Bigün, J., Granlund, G. H. & Wiklund, J. (1991). Multidimensional orientation estimation with applications to texture analysis and optical flow, *IEEE-PAMI* **13**(8): 775–790.

Davis, J. & Bobick, A. (1997). The representation and recognition of action using temporal templates, *CVPR'97*, 928–934.

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

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

Jähne, B. (1995). *Digital Image Processing*, Springer-Verlag.

Koenderink, J. J. (1984), 'The structure of images', *Biol. Cyb.* **50**, 363–370.

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

Laptev, I. & Lindeberg, T. (2003). Interest point detection and scale selection in space-time, *Scale-Space'03*, Springer LNCS 2695, 372–387.

Laptev, I. & Lindeberg, T. (2004a). Velocity adaptation of space-time interest points, *ICPR'04*, Cambridge, U.K.

Laptev, I. & Lindeberg, T. (2004b). Velocity-adapted spatio-temporal receptive fields for direct recognition of activities, *IVC* **22**(2): 105–116.

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

Lindeberg, T. (1997). Linear spatio-temporal scale-space, *Scale-Space'97*, Springer LNCS 1252, 113–127. Extended version in Technical Report ISRN KTH NA/P–01/22–SE.

Lindeberg, T. (2002). Time-recursive velocity-adapted spatio-temporal scale-space filters, *ECCV'02*, Springer LNCS 2350, 52–67.

Lindeberg, T., Akbarzadeh, A. & Laptev, I. (2004). Galilean-diagonalized spatio-temporal interest operators, *Technical Report ISRN KTH/NA/P--04/05--SE*, KTH, Stockholm, Sweden.

Lindeberg, T. & Fagerström, D. (1996). Scale-space with causal time direction, *ECCV'96*, Springer LNCS 1064, 229–240.

Lindeberg, T. & Gårding, J. (1997). Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D structure, *Image and Vision Computing* **15**: 415–434.

Lowe, D. (1999). Object recognition from local scale-invariant features, *ICCV'99*, 1150–1157.

Lukas, B. D. & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision, *Image Understanding Workshop*.

Mikolajczyk, K. & Schmid, C. (2002). An affine invariant interest point detector, *ECCV'02*, Springer LNCS 2350, I:128–142.

Nagel, H. & Gehrke, A. (1998). Spatiotemporal adaptive filtering for estimation and segmentation of optical flow fields, *ECCV'98*, Springer LNCS 1407, II:86–102.

Niyogis, S. A. (1995). Detecting kinetic occlusions, *ICCV'95*, 1044–1049.

ter Haar Romeny, B., Florack, L. & Nielsen, M. (2001), Scale-time kernels and models, *Scale-Space'01*, Springer LNCS 2106, 255–263.

Wildes, R. & Bergen, J. (2000). Qualitative spatio-temporal analysis using an oriented energy representation, *ECCV'00*, Springer LNCS 1843, II:768–784.

Zelnik-Manor, L. & Irani, M. (2001). Event-based analysis of video, *CVPR'01*, II:123–130.

Zetzsche, C. & Barth, E. (1991). Direct detection of flow discontinuities by 3-D curvature operators, *Patt. Recogn. Lett.* **12**(12): 771–779.