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 (Royal Institute of Technology)
SE-100 44 Stockholm, Sweden

Technical report ISRN KTH/NA/P–04/05–SE, March 2004.
Shortened version to appear at ICPR'04, Cambridge, UK., August 2004.

Abstract

This paper presents a set of image operators for detecting regions in space-time where interesting events occur. To define such spatio-temporal interest operators, we compute a 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 and a complementary diagonalization. From the Galilean-diagonalized descriptor so obtained, we then formulate different types of space-time interest operators, and illustrate their properties on various types of real and synthetic images.

Note! This report contains a number of figures that should be viewed in colour. If you only have access to a black-and-white printout, please fetch an on-line version of this manuscript from http://www.nada.kth.se/cvap.

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

Contents

1 Introduction

For analysing the space-time structure of images from our environment, the ability to detect regions of interest is an important pre-processing stage for subsequent recognition. The presumably simplest approach for constructing such a mechanism is by regular frame differencing, i.e. computing first-order temporal derivatives followed by thresholding. The results 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 used these for constructing a motion mask, which was then represented in terms of moment descriptors, in order to characterize the motion. This approach, however, assumes a static background as well as a stationary camera.

A more local approach was developed by (Laptev & Lindeberg 2003), based on an extension of the Harris operator to spatio-temporal interest points. This operator effectively captures well localized points in space-time with strong simultaneous variations over both space and time. Due to the formulation of this operator, however, in terms of the the eigenvalues of a second-moment matrix, it can be shown that this operator is not invariant under constant velocity motion. For this reason, it is of interest to develop alternative space-time interest operators.

A general problem when interpreting spatio-temporal image data originates from the fact that motion descriptors will be affected by relative motions between the object and the camera. It is therefore essential to aim at image operators that are invariant to local Galilean transformations. One approach to achieve Galilean invariance is to consider space-time receptive fields that are 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), based on histograms of spatio-temporal receptive fields, global stabilization was used when computing spatio-temporal derivatives. (Laptev & Lindeberg 2004*b*) 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 builds upon several of the abovementioned 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 the detection of spatial interest points or the detection of spatial regions-of-interest can be used as pre-processing stages for spatial recognition (Lowe 1999, Mikolajczyk & Schmid 2002). The operators to be presented are also 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.

The paper is organized as follows: Section 2 starts with a review of spatio-temporal scale-space, and describes how a spatio-temporal second-moment matrix transforms under Galilean transformations. This material forms the theoretical background for section 3, which introduces the notion of Galilean diagonalization, which in turn constitutes the basis for defining Galilean invariant as well as Galilean-corrected spatio-temporal interest operators in section 4 and section 5. Section 6 shows experimental

results of applying these operators to different types of images, and section 7 presents extensions from grey-level to colour images as well as to local contrast normalization. Finally, section 8 concludes with a summary and discussion.

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 spatio-temporal 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

$$L(\cdot;\ \Sigma) = h(\cdot;\ \Sigma) * f(\cdot) \tag{1}$$

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

$$\Sigma = \begin{pmatrix} \lambda_1 \cos^2 \alpha + \lambda_2 \sin^2 \alpha + u^2 \lambda_t & (\lambda_2 - \lambda_1) \cos \alpha \, \sin \alpha + uv\lambda_t & u\lambda_t \\ (\lambda_2 - \lambda_1) \cos \alpha \, \sin \alpha + uv\lambda_t & \lambda_1 \sin^2 \alpha + \lambda_2 \cos \alpha^2 + v^2 \lambda_t & v\lambda_t \\ u\lambda_t & v\lambda_t & \lambda_t \end{pmatrix} \tag{2}$$

where $(\lambda_1, \lambda_2, \alpha)$ describe the amount of spatial (possibly anisotropic) smoothing in terms of two eigenvalues and their orientation α in space, 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 a spatial scale parameter $\sigma^2 = \lambda_1 = \lambda_2$ and temporal scale parameter $\tau^2 = \lambda_t$.

For simplicity, we shall here model the smoothing operation by a 3-D Gaussian kernel with covariance matrix Σ

$$h(p;\ \Sigma) = g(p;\ \Sigma) = \frac{1}{(2\pi)^{3/2} \sqrt{\det \Sigma}} e^{-p^T \Sigma^{-1} p/2}, \tag{3}$$

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.

For real-time processing, this model can be extended to time-causal smoothing kernels based on the time-causal scale-space concepts in (Koenderink 1988, Lindeberg & Fagerström 1996, Lindeberg 1997, Florack 1997, Lindeberg 2002).

Second-moment descriptor in space-time. For describing local image structures as well as for estimating local image deformations, the second moment matrix (sometimes referred to as the structure tensor) is a highly useful descriptor (Förstner & Gülch 1987, Bigün et al. 1991, Lindeberg 1994, Jähne 1995). In 2+1-D space-time, this descriptor can at any point $p = (x, y, t)^t$ be defined as

$$\mu(p;\ \Sigma) = \int_{q \in \mathbb{R}^3} (\nabla L(q))(\nabla L(q))^T \, w(p - q;\ \Sigma) \, dq, \tag{4}$$

where $\nabla L = (L_x, L_y, L_t)^T$ denotes the spatio-temporal gradient vector-time and w is a spatio-temporal window function, for simplicity modelled as a Gaussian function with covariance matrix Σ multiplied by a scaling factor γ^2

$$w(p;\ \Sigma) = g(p;\ \Sigma) = \frac{1}{2\gamma^3 \pi \sqrt{\det \Sigma}} e^{-p^T \Sigma^{-1} p/2\gamma^2}. \tag{5}$$

In terms of matrix elements, we have

$$\begin{pmatrix} \mu_{xx} & \mu_{xy} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix} = \int_{q \in \mathbb{R}^3} \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix} w(p - q;\, \Sigma)\, dq. \tag{6}$$

Transformation property under Galilean transformations. Given a spatio-temporal image sequence f, consider a Galilean transformation in space-time

$$p' = \begin{pmatrix} x' \\ y' \\ t' \end{pmatrix} = Gp = \begin{pmatrix} 1 & 0 & -u \\ 0 & 1 & -v \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ t \end{pmatrix} \tag{7}$$

and define a Galilean transformed image sequence according to $f'(p') = f(p)$. Then, define scale-space representations L and L' of f and f', respectively, according to

$$L(\cdot;\, \Sigma) = g(\cdot;\, \Sigma) * f(\cdot), \qquad L'(\cdot;\, \Sigma') = g(\cdot;\, \Sigma') * f'(\cdot). \tag{8}$$

Then, it can be shown that $L'(\cdot;\, \Sigma') = L(\cdot;\, \Sigma)$ if and only if the covariance matrices are related according to (Lindeberg 1997, Lindeberg 2002)

$$\Sigma' = G \Sigma G^T \tag{9}$$

In terms of matrix elements, this corresponds to

$$\begin{pmatrix} C'_{xx} & C'_{xy} & C'_{xt} \\ C'_{xy} & C'_{yy} & C'_{yt} \\ C'_{xt} & C'_{yt} & C'_{tt} \end{pmatrix} = \begin{pmatrix} 1 & 0 & -u \\ 0 & 1 & -v \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} C_{xx} & C_{xt} & C_{xt} \\ C_{xy} & C_{yy} & C_{yt} \\ C_{xt} & C_{yt} & C_{tt} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -u & -v & 1 \end{pmatrix}. \tag{10}$$

Next, let us define second-moment matrices μ and μ' according to

$$\mu(p;\, \Sigma) = \int_{q \in \mathbb{R}^3} (\nabla L(q))(\nabla L(q))^T\, g(p - q;\, \Sigma)\, dq, \tag{11}$$

$$\mu'(p';\, \Sigma') = \int_{q' \in \mathbb{R}^3} (\nabla L'(q'))(\nabla L'(q'))^T\, g(p' - q';\, \Sigma')\, dq'. \tag{12}$$

Then, from the general transformation property of second-moment matrices under linear transformations (Lindeberg 1994, Lindeberg & Gårding 1997), it can be shown that μ and μ' are related according to

$$\mu' = G^{-T} \mu G^{-1} \tag{13}$$

provided that the covariance matrices satisfy $\Sigma' = G \Sigma G^T$. In terms of matrix elements, we have

$$\begin{pmatrix} \mu'_{xx} & \mu'_{xt} & \mu'_{xt} \\ \mu'_{xy} & \mu'_{yy} & \mu'_{yt} \\ \mu'_{xt} & \mu'_{yt} & \mu'_{tt} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ u & v & 1 \end{pmatrix} \begin{pmatrix} \mu_{xx} & \mu_{xt} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix} \begin{pmatrix} 1 & 0 & u \\ 0 & 1 & v \\ 0 & 0 & 1 \end{pmatrix} \tag{14}$$

with

$$\mu'_{xx} = \mu_{xx} \tag{15}$$

$$\mu'_{xy} = \mu_{xy} \tag{16}$$

$$\mu'_{yy} = \mu_{yy} \tag{17}$$

$$\mu'_{xt} = u \mu_{xx} + v \mu_{xy} + \mu_{xt} \tag{18}$$

$$\mu'_{yt} = u \mu_{xy} + v \mu_{yy} + \mu_{yt} \tag{19}$$

$$\mu'_{tt} = u^2 \mu_{xx} + 2uv \mu_{xy} + v^2 \mu_{yy} + 2u \mu_{xt} + 2v \mu_{yt} + \mu_{tt}. \tag{20}$$

In other words, the purely spatial elements $(\mu_{xx}, \mu_{xy}, \mu_{yy})$ of the second moment matrix are preserved by the Galilean transformation, while the purely temporal element μ_{tt} as well as the mixed elements (μ_{xt}, μ_{yt}) are affected.

3 Galilean diagonalization

A specific convention we shall consider in this work is to determine the velocity components (u, v) in a local Galilean transformation $p' = Gp$ such that the transformed second-moment matrix μ' is block diagonal with $(\mu'_{xt}, \mu'_{yt}) = (0, 0)$:

$$\mu' = \begin{pmatrix} \mu'_{xx} & \mu'_{xt} & 0 \\ \mu'_{xy} & \mu'_{yy} & 0 \\ 0 & 0 & \mu'_{tt} \end{pmatrix} \tag{21}$$

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 $\mu' = G^{-T} \mu G^T$ 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 scene and the camera. Specifically, this form of block diagonalization implies a local normalization of local space-time structures that is invariant under superimposed Galilean transformations (see appendix A.1).

From (18) and (19) it follows that block diagonalization is obtained if (u, v) satisfy the following equations:

$$\begin{pmatrix} \mu_{xx} & \mu_{xy} \\ \mu_{xy} & \mu_{yy} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = - \begin{pmatrix} \mu_{xt} \\ \mu_{yt} \end{pmatrix} \tag{22}$$

i.e., structurally similar equations as are used for computing optic flow according to the method by (Lukas & Kanade 1981). The solution of these equations is

$$\begin{pmatrix} u \\ v \end{pmatrix} = - \frac{1}{\mu_{xx} \mu_{yy} - \mu_{xy}^2} \begin{pmatrix} \mu_{yy} \mu_{xt} - \mu_{xy} \mu_{yt} \\ -\mu_{xy} \mu_{xt} + \mu_{xx} \mu_{yt} \end{pmatrix} \tag{23}$$

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

$$\int_{x,y,t \in \mathbb{R}^3} L_x \, L_t \, g(x, y, t; \; \Sigma) \, dx \, dy \, dt = \int_{x,y,t \in \mathbb{R}^3} L_y \, L_t \, g(x, y, t; \; \Sigma) \, dx \, dy \, dt = 0 \tag{24}$$

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 $(u_\parallel, v_\parallel)$ 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 (see appendix A.2 for a proof):

$$E\left((\nabla_{space} L)(\nabla_{space} L)^T \begin{pmatrix} u \\ v \end{pmatrix}\right) = E\left(|\nabla_{space} L|^2 \begin{pmatrix} u_{\|} \\ v_{\|} \end{pmatrix}\right) = 0. \tag{25}$$

In this respect, Galilean diagonalization implies cancelling 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 two-dimensional rotation $p'' = Rp'$ in space that diagonalizes the remaining spatial second moment matrix with the elements $(\mu'_{xx}, \mu'_{xy}, \mu'_{yy})$ such that $\mu''_{xy} = 0$. Thus, we diagonalize the original second moment matrix μ into

$$\mu'' = R^{-T} G^{-T} \mu G^{-1} R^{-1} = \begin{pmatrix} \nu_1 & & \\ & \nu_2 & \\ & & \nu_3 \end{pmatrix} \tag{26}$$

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 commonly used approach of using the eigenvalues of a spatio-temporal second-moment matrix for motion analysis (Bigün et al. 1991, Jähne 1995). In terms of diagonalization, an eigenvalue analysis corresponds to transforming the space by unitary transformation, a rotation U in three dimensions, such that

$$\mu''' = U^{-T} \mu U^{-1} = \begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3 \end{pmatrix} \tag{27}$$

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.

Remarks: Note also that compared to an eigenvalue based diagonalization of a 3×3 matrix, the Galilean diagonalization is easy to compute in closed form, since we have a closed-form expression (23) for the velocity vector $(u, v)^T$ in the Galilean transformation G defined in (7), and the remaining two-dimensional rotation R in space

$$R = \begin{pmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{28}$$

is easily obtained from

$$\tan 2\phi = \frac{2\mu'_{xy}}{\mu'_{yy} - \mu'_{xx}} = \frac{2\mu_{xy}}{\mu_{yy} - \mu_{xx}}. \tag{29}$$

In many cases, and as we shall see examples of next, it is, however, not necessary to compute this spatial rotation matrix explicitly, since we can compute the sum and the

product of the diagonal elements ν_1 and ν_2 in the purely spatial part of the Galilean diagonalized second-moment matrix according to

$$\nu_1 + \nu_2 = \mu'_{xx} + \mu'_{yy} = \mu_{xx} + \mu_{yy}, \tag{30}$$

$$\nu_1 \nu_2 = \mu'_{xx} \mu'_{yy} - (\mu'_{xy})^2 = \mu_{xx} \mu_{yy} - \mu_{xy}^2. \tag{31}$$

If an explicit Galilean diagonalization is needed, it can be obtained from the following types of operations: (i) computing a rational expression for $\nu_3 = \mu'_{tt}$, (ii) computing (u, v) from a linear system of equations if the explicit transformation matrix G is needed, (iii) performing direct lookup of trigonometric functions to obtain R if needed, and (iv) solving quadratic equations to obtain ν_1 and ν_2 from (30) and (31), or alternatively performing 2×2 matrix multiplications by applying an explicit expression for R on μ. Thus, besides being more accurate than a more traditional eigenvalue decomposition, this form of normalization is also more easy to compute.

Combining Galilean diagonalization with affine normalization. In the spatial domain, an effective method for spatial normalization consists of determining a spatial affine normalization $p''_{space} = B p_{space}$ that transforms a spatial second-moment matrix

$$\mu_{space} = \begin{pmatrix} \mu_{xx} & \mu_{xy} \\ \mu_{xy} & \mu_{yy} \end{pmatrix} = \int_{(\xi,\eta) \in \mathbb{R}^2} \begin{pmatrix} L_x^2 & L_x L_y \\ L_x L_y & L_y^2 \end{pmatrix} w_{space}(x - \xi, y - \eta;\ \Sigma_{space})\, d\xi\, d\eta \tag{32}$$

into diagonal form, i.e. determining B such that

$$\mu''_{space} = B^{-T} \mu_{space} B^{-1} = cI \tag{33}$$

for some constant c where I denotes the unit matrix (Lindeberg & Gårding 1997, Mikolajczyk & Schmid 2002). Specifically, such affine normalization can be achieved with $B = \mu_{space}^{1/2}$, where $\mu_{space}^{1/2}$ denotes a solution to the equation $B^T B = \mu_{space}$. Alternatively, we can compute B from a combination of a spatial rotation and a diagonal transformation. Let first R_{space} denote a two-dimensional spatial rotation matrix of the form (28) with ϕ determined analogous to (29). Then, it follows that the rotation $p'_{space} = R_{space} p_{space}$ implies that the transformed spatial second-moment matrix μ'_{space} will be a diagonal matrix:

$$\mu'_{space} = R_{space}^{-T} \mu_{space} R_{space}^{-1} = \begin{pmatrix} \mu'_{xx} & 0 \\ 0 & \mu'_{yy} \end{pmatrix} \tag{34}$$

Next, if we perform a diagonal transformation $p''_{space} = D_{space} p'_{space}$ with

$$D_{space} = \begin{pmatrix} \sqrt[4]{\frac{\mu'_{xx}}{\mu'_{yy}}} & 0 \\ 0 & \sqrt[4]{\frac{\mu'_{yy}}{\mu'_{xx}}} \end{pmatrix} \tag{35}$$

the transformed second-moment matrix μ'' will be of the form:

$$\mu''_{space} = D_{space}^{-T} \mu'_{space} D_{space}^{-1} = \begin{pmatrix} \mu''_{xx} & 0 \\ 0 & \mu''_{yy} \end{pmatrix} = \begin{pmatrix} \sqrt{\mu'_{xx} \mu'_{yy}} & 0 \\ 0 & \sqrt{\mu'_{xx} \mu'_{yy}} \end{pmatrix} \tag{36}$$

i.e. suitable for affine normalization according to (33). By applying a corresponding form of spatial normalization to the Galilean diagonalized spatio-temporal second-moment matrix (26), i.e., performing a spatio-temporal transformation $p''' = Dp''$ with

$$D = \begin{pmatrix} \sqrt[4]{\frac{\nu_1}{\nu_2}} & & \\ & \sqrt[4]{\frac{\nu_2}{\nu_1}} & \\ & & 1 \end{pmatrix} \tag{37}$$

we obtain

$$\mu''' = D^{-1} R_{space}^{-T} G^{-T} \mu G^{-1} R_{space}^{-1} D^{-1} = \begin{pmatrix} \sqrt{\nu_1 \nu_2} & & \\ & \sqrt{\nu_1 \nu_2} & \\ & & \nu_3 \end{pmatrix} \tag{38}$$

In other words, Galilean diagonalization can be easily combined with affine normalization, and we can interpret the combined transformation (38) as a canonical generalization of affine normalization from space to space-time.[1]

4 Galilean invariant and Galilean corrected 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:

- 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.
- 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 a major part of the relative motion to the camera.[2] If we aim at affine invariance in addition to Galilean invariance, then also an affine Gaussian scale-space concept should be considered over the spatial domain (Lindeberg 1994).

[1] Note, however, that in order to normalize for scale variations, a complementary scale selection step will be needed.

[2] In comparison with the related notions of affine shape-adaptation in space (Lindeberg & Gårding 1997, Mikolajczyk & Schmid 2002) or velocity adaptation in space-time (Lindeberg 1997, Nagel & Gehrke 1998, Lindeberg 2002, Laptev & Lindeberg 2004*b*), 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-corrected motion descriptors. By differentiating the Galilean transformation (7), it follows that the first-order derivative operators transform according to

$$\partial_{x'} = \partial_x, \tag{39}$$

$$\partial_{y'} = \partial_y, \tag{40}$$

$$\partial_{t'} = \partial_t + u\partial_x + v\partial_y. \tag{41}$$

In other words, the spatial derivative operators are unaffected by Galilean transformations, while the temporal derivative transforms according to the image velocity. Given that a second-moment descriptor has been computed at any image point and that this descriptor been Galilean diagonalized with velocity vector $(u, v)^T$ according to (23), a general approach to motion compensation is therefore to express all temporal derivatives in a Galilean transformed frame with the same image velocity. If the derivative expression are computed from scale-space concept that is closed under this Galilean transformation, it follows that these derivatives will be truly invariant. Otherwise, the effect of this Galilean correction is to perform a partial compensation for the influence of the Galilean transformation.

For example, if we apply this approach to the purely temporal element μ_{tt} in the second-moment matrix, alternatively if we use the transformation property (20), we obtain

$$\mu'_{tt} = u^2\mu_{xx} + 2uv\mu_{xy} + v^2\mu_{yy} + u\mu_{xt} + 2v\mu_{yt} + \mu_{tt} \tag{42}$$

which after insertion of the explicit expression for $(u, v)^T$ can be simplified to

$$\mu'_{tt} = \mu_{tt} - \frac{\mu_{xx}\mu_{yt}^2 + \mu_{yy}\mu_{xt}^2 - 2\mu_{xy}\mu_{xt}\mu_{yt}}{\mu_{xx}\mu_{yy} - \mu_{xy}^2}. \tag{43}$$

This form of Galilean correction of temporal derivatives is, however, not restricted to the elements of a second-moment matrix, and applies to any temporal derivative expression, with applications to spatio-temporal feature extraction and recognition.

Remark: Note that also in the absence of a second-moment matrix, a corresponding determination of a unique Galilean transformation for an image point can be performed based on the spatio-temporal Hessian matrix, by determining a velocity vector (u, v) such that the transformed mixed second-order derivatives satisfy $L_{x't'} = L_{y't'} = 0$. This property follows from the fact that under linear transformations the Hessian matrix transforms in a similar way as the second-moment matrix. Therefore, the idea of Galilean diagonalization of the second-moment matrix applies to Galilean diagonalization of the spatio-temporal Hessian matrix as well. A local velocity estimate for Galilean correction obtained from a pointwise Hessian matrix can, however, be expected to be less robust than a velocity estimate computed from a regional second-moment matrix.

5 Spatio-temporal interest operators

In the following, we shall apply the abovementioned notion of Galilean correction for defining spatio-temporal interest operators. A first approach we shall follow is to use

$$I_1 = \nu_3 = \mu'_{tt} \tag{44}$$

as a basic measure for computing candidate regions of interest. If the space-time image is 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 space-time filters) the value of this descriptor will be zero. Hence, I_1 can be regarded as a measure of how the local space-time image structure deviates from that of a pure translation model. 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 elements of μ at every point according to

$$I_1 = \mu_{tt} - \frac{\mu_{xx}\mu_{yt}^2 + \mu_{yy}\mu_{xt}^2 - 2\mu_{xy}\mu_{xt}\mu_{yt}}{\mu_{xx}\mu_{yy} - \mu_{xy}^2} \tag{45}$$

The operator I_1 will respond to rather wide classes of space-time events. If one is interested in more restrictive space-time interest operators, we can, for example, consider two extensions of the Harris operator (Harris & Stephens 1988) to space-time. Given a spatial second moment matrix μ_{2D} with eigenvalues (λ_1, λ_2), the traditional Harris operator is defined as

$$H = \lambda_1\lambda_2 - C(\lambda_1 + \lambda_2)^2 = \det \mu_{2D} - C(\operatorname{trace} \mu_{2D})^2 \tag{46}$$

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

$$I_2 = (\nu_1 + \nu_2)\nu_3 - C_2(\nu_1 + \nu_2 + \nu_3)^2, \tag{47}$$

which using $\nu_1 + \nu_2 = \mu_{xx} + \mu_{yy}$ can also be written

$$I_2 = (\mu_{xx} + \mu_{yy})\mu'_{tt} - C_2(\mu_{xx} + \mu_{yy} + \mu'_{tt})^2 \tag{48}$$

By treating all diagonal elements individually, we can define the following modification[3] of the operator in (Laptev & Lindeberg 2003)

$$I_3 = \nu_1\nu_2\nu_3 - C_3(\nu_1 + \nu_2 + \nu_3)^3 \tag{49}$$

which using $\nu_1\nu_2 = \mu_{xx}\mu_{yy} - \mu_{xy}^2$ can be expressed as

$$I_3 = (\mu_{xx}\mu_{yy} - \mu_{xy}^2)\mu'_{tt} - C_3(\mu_{xx} + \mu_{yy} + \mu'_{tt})^3 \tag{50}$$

[3]With λ_1, λ_2 and λ_3 denoting the eigenvalues of a spatio-temporal second-moment matrix, in (Laptev & Lindeberg 2003) a space-time interest operator H is defined as $H = \lambda_1\lambda_2\lambda_3 - C_3(\lambda_1 + \lambda_2 + \lambda_3)^3$. While this operator has been demonstrated to give intuitively reasonable space-time interest points corresponding to high spatial and temporal variations in the image structures, due to the fact that this operator is defined in terms of the eigenvalues of the second-moment matrix it follows that this operator is not Galilean invariant. By redefining H into $I_3 = \nu_1\nu_2\nu_3 - C_3(\nu_1 + \nu_2 + \nu_3)^3$, however, we obtain a Galilean invariant operator provided that either of the following mechanisms are included: (i) considering the entire family of spatio-temporal smoothing kernels, or (ii) performing velocity adaptation.

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. Thus, we can expect the operator I_3 to be most selective and I_1 to be the least selective operator of these three.

Remarks: With regard to invariance of operator responses, a possible drawback of defining interest point operators in space-time in a fully analogous way as done in the Harris operator, i.e., in terms of a product of diagonal elements minus a sum of diagonal elements raised to a suitable power so as to make the expression homogeneous, is that the operator response will be dependent on the actual units by which the spatial dimensions are measured. To obtain invariance to, alternatively to compensate for such effects, one could also consider the following alternative definitions, which will share qualitatively similar properties as I_2 and I_3:

$$\bar{I}_2 = (\nu_1 + \nu_2)\nu_3, \tag{51}$$

$$\bar{\bar{I}}_2 = (\nu_1 + \nu_2)\nu_3 - (C_{2,space}(\nu_1 + \nu_2) + C_{2,time}\nu_3)^2, \tag{52}$$

$$\bar{I}_3 = \nu_1\nu_2\nu_3, \tag{53}$$

$$\bar{\bar{I}}_3 = \nu_1\nu_2\nu_3 - (C_{3,space}(\nu_1 + \nu_2) + C_{3,time}\nu_3)^3. \tag{54}$$

$$\bar{\bar{\bar{I}}}_3 = \nu_1\nu_2\nu_3 - (C_{3,space}\sqrt{\nu_1\nu_2} + C_{3,time}\nu_3)^3. \tag{55}$$

With these modified definitions, it follows that the maps of $\bar{I}_2$ and $\bar{I}_3$ will transform by constant scaling factors under uniform rescalings of the spatial or the temporal domains. Therefore, spatio-temporal maxima of these operators will be preserved under any uniform scaling of either space, time or both dimensions. If one instead aims at building explicit thresholding on ranges of parameter values into the operator, the parameters $C_{2,space}$, $C_{2,time}$, $C_{3,space}$ and $C_{3,time}$ in $\bar{\bar{I}}_2$ and $\bar{\bar{I}}_3$ can be adjusted so as to perform thresholding on the magnitudes of the spatial and temporal contributions to the operator response.

With regard to combined affine and Galilean invariance, it follows from (38) that the operators I_3 and $\bar{\bar{\bar{I}}}_3$ will be both affine and Galilean invariant (for $\bar{\bar{\bar{I}}}_3$,the constants $C_{3,space}$ and $C_{3,time}$ should be appropriately adjusted), i.e. invariant to both affine transformations in space and Galilean transformations in space-time.

6 Experiments

Figures 1–3 show a few snapshots of computing I_1, I_2 and I_3 for different types of spatio-temporal image patterns, for simplicity computed by space-time separable filtering. [4] For comparison, we also show maps of the corresponding entities without

[4]For simplicity, we have in these experiments used a scale-space constructed from space-time separable spatio-temporal smoothing kernels, resulting in Galilean-corrected as opposed to truly Galilean invariant image descriptors. In a companion paper (Laptev & Lindeberg 2004*a*), we explore the combination of non-separable velocity-adapted filters with interest points derived from I_3.

Galilean-correction i.e.,

$$\tilde{I}_1 = \mu_{tt} \tag{56}$$

$$\tilde{I}_2 = (\mu_{xx} + \mu_{yy})\mu_{tt} - C_2(\mu_{xx} + \mu_{yy} + \mu_{tt})^2 \tag{57}$$

$$\tilde{I}_3 = \det \mu - C_3(\mathrm{trace}\, \mu)^3 \tag{58}$$

as well as sample frames from the original image sequence f and its spatio-temporal scale-space representation L.

Figure 1 shows result from a synthetic experiment with two circular blobs that are moving against a static textured background. In the left columns, the blobs differ in size, while having the same image velocity $u = 2.5$ pixels/frame. In the right columns, the blobs have the same size while differing in image velocity (low velocity 0.5 pixels/frame, high velocity 3 pixels/frame). The results in the leftmost column have been computed at a fine spatial scale ($\sigma = 1$, $\tau = 0.5$), and the middle left column shows corresponding results at a coarser spatial scale ($\sigma = 6$, $\tau = 0.5$).[5] In the middle right column, the results have been computed at a fine temporal scale ($\sigma = 3$, $\tau = 0.5$), and in the right column the results have been computed at a coarse temporal scale ($\sigma = 3$, $\tau = 6$) As can be seen from the results, all operators give a strong response in regions where a local translational model is not valid. By comparison the results in the leftmost and the middle left column, we can moreover note that finer spatial scales give more emphasis to small size image structures, and coarse spatial scales gives more emphasis to image structures with large spatial extent. Similarly, by comparing the results in the middle right and the rightmost columns, we can observe that fine temporal scales give more emphasis to objects that move with high image velocities and that coarser temporal scales give more emphasis to slowly moving objects.

Figure 2 shows the result of computing corresponding descriptors for real images of (i) a walking person with approximately stabilized camera, (ii) a jumping person with the camera slowly following the person, (iii) two walking persons with camera stabilized on right person, (v) walking person with camera stabilized on person. All sequences have been taken with a handheld camera. Figure 3 shows more results with (i-ii) pedestrian lights turning green while the camera is shaking, (iii-iv) a traffic scene with nearby cars and cars far away registered at a fine spatial scale and coarse temporal scale ($\sigma = 0.4, \tau = 16.0$) as well as at coarse spatial scale and fine temporal scale ($\sigma = 3, \tau = 0.5$). 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 $\tilde{I}_1 = \mu_{tt}$, 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.

[5] In all experiments, we have set the integration scales proportional to the local scale, i.e., $\sigma_i = \gamma\sigma$ and $\tau_i = \gamma\tau$, with $\gamma = 2$.

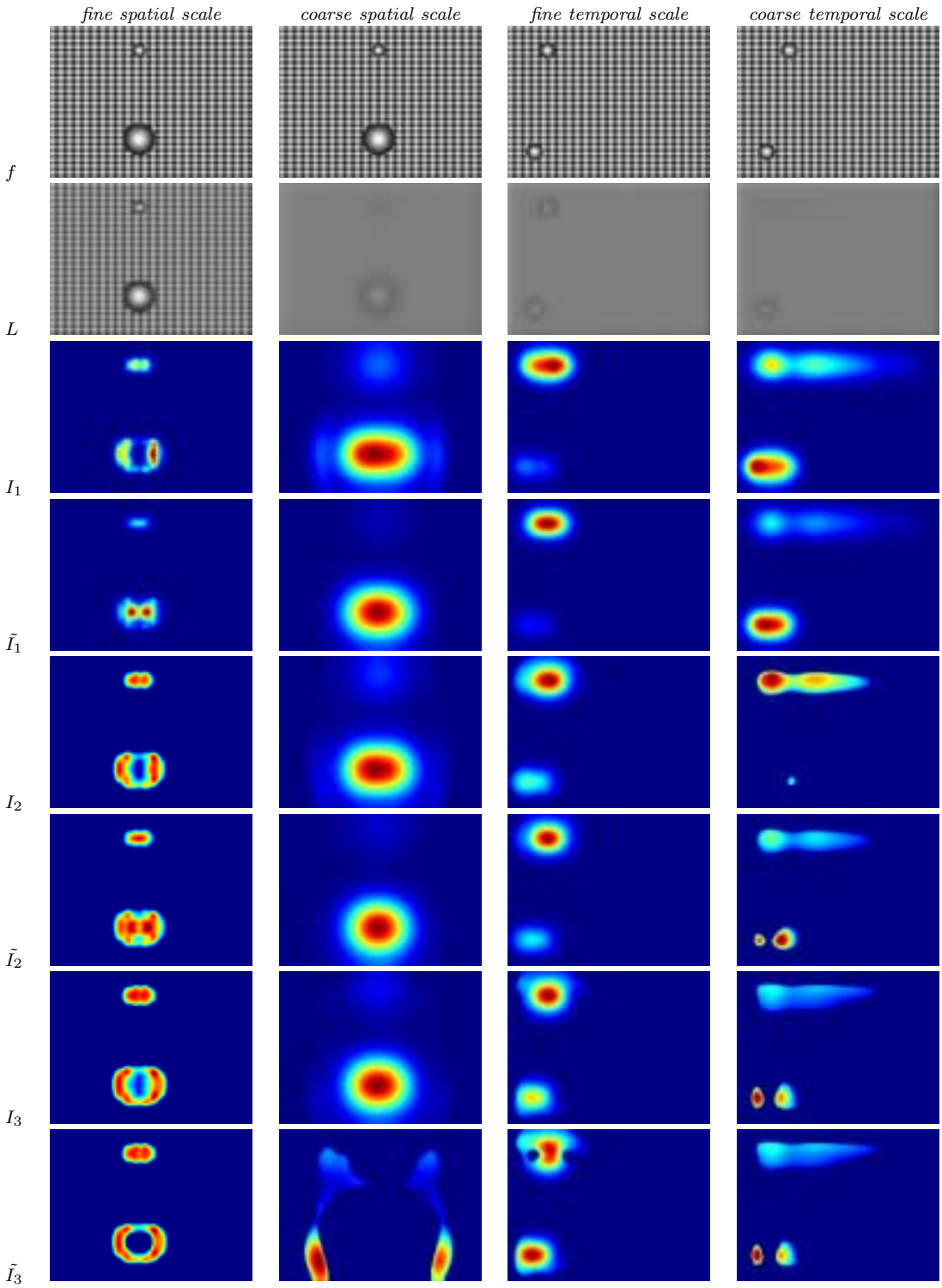

Figure 1: *Maps of the Galilean-corrected interest operators* I_1, I_2 *and* I_3 *as well as corresponding non-corrected descriptors* $\tilde{I}_1$, $\tilde{I}_2$ *and* $\tilde{I}_3$ *computed from space-time separable spatio-temporal scale-space representations* L *of different synthetic image sequences* f *with moving circular blobs against a static textured background. (i-ii) blobs of different sizes* ($sigma = 2.5, \sigma = 6$) *moving with same image velocity (2.5 pixels/frame): (i) fine spatial scale* ($\sigma = 1, \tau = 0.5$), *(ii) coarse spatial scale* ($\sigma = 6, \tau = 0.5$). *(iii-iv) blobs of same size* ($\sigma = 3$) *moving with different velocities (0.5 and 3 pixels/frame): (iii) fine temporal scale* ($\sigma = 3, \tau = 0.5$), *(iv) coarse temporal scale* ($\sigma = 3, \tau = 6$). *Image size:* 160×120 *pixels. (Note! This figure should be viewed in colour. Moreover, notice that the colour scales are different in all images.)*

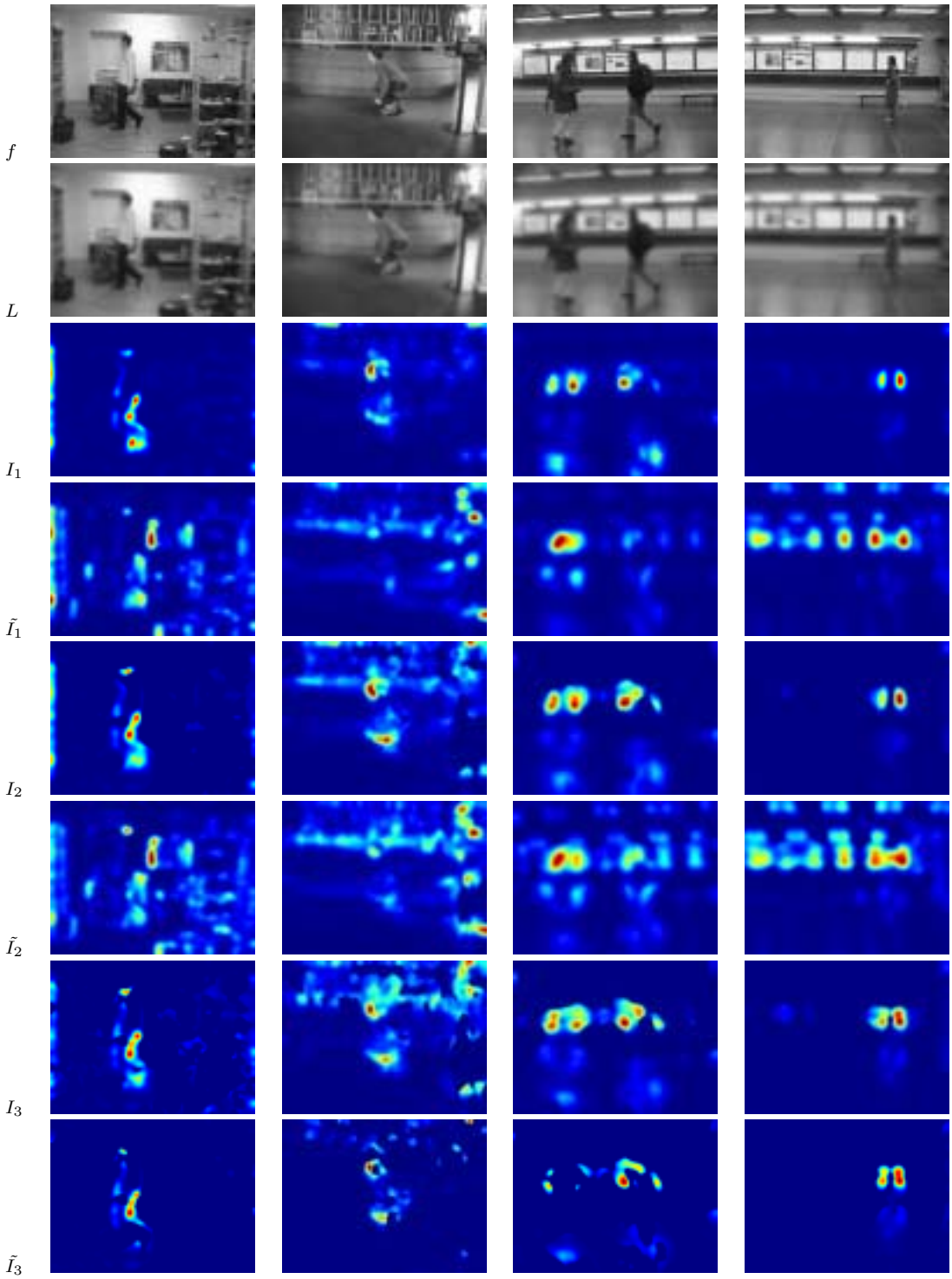

Figure 2: *Maps of the Galilean-corrected interest operators I_1, I_2 and I_3 as well as corresponding non-corrected descriptors $\tilde{I}_1$, $\tilde{I}_2$ and $\tilde{I}_3$ computed from space-time separable spatio-temporal scale-space representations L of different image sequences f. From left to right: (i) a walking person with approximately stabilized camera, (ii) a jumping person with the camera slowly following the person, (iii) two walking persons with camera stabilized on right person, (v) walking person with camera stabilized on person. In columns (i)–(ii) the scale parameters were $(\sigma = 1.0, \tau = 0.5)$, while in columns (iii)–(iv) the scale parameters were $(\sigma = 1.5, \tau = 0.5)$. Image size: 160×120 pixels. (Note! This figure should be viewed in colour.)*

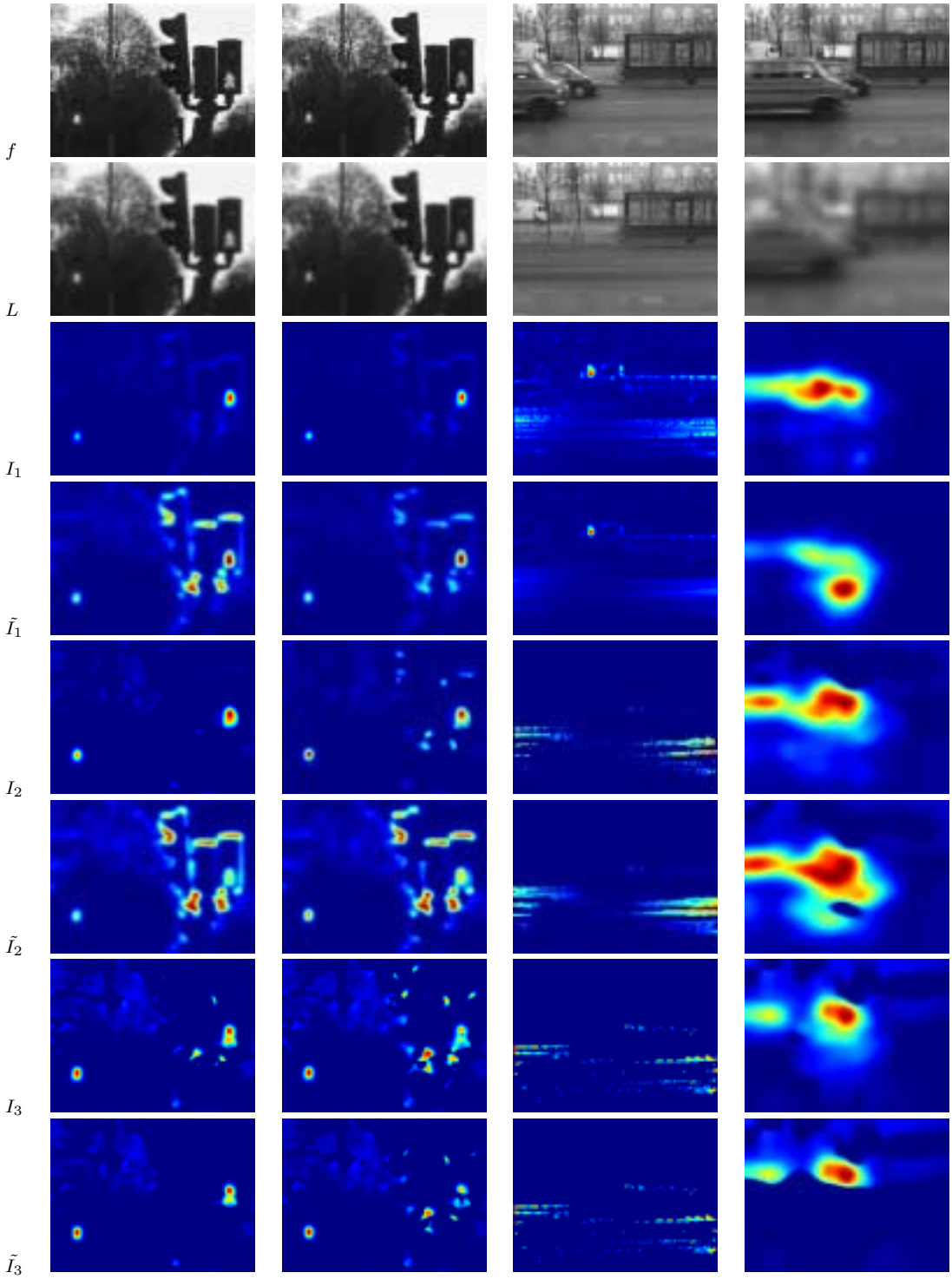

Figure 3: *Maps of the Galilean-corrected interest operators I_1, I_2 and I_3 as well as corresponding non-corrected descriptors $\tilde{I}_1$, $\tilde{I}_2$ and $\tilde{I}_3$ computed from space-time separable spatio-temporal scale-space representations L of different image sequences f. From left to right: (i-ii) traffic lights turning green while camera moves, (iii-iv) a traffic scene with nearby cars and cars far away registered at a fine spatial scale and coarse temporal scale ($\sigma = 0.4, \tau = 16.0$) as well as at coarse spatial scale and fine temporal scale ($\sigma = 3, \tau = 0.5$). Image size: 160×120 pixels. (Note! This figure should be viewed in colour.)*

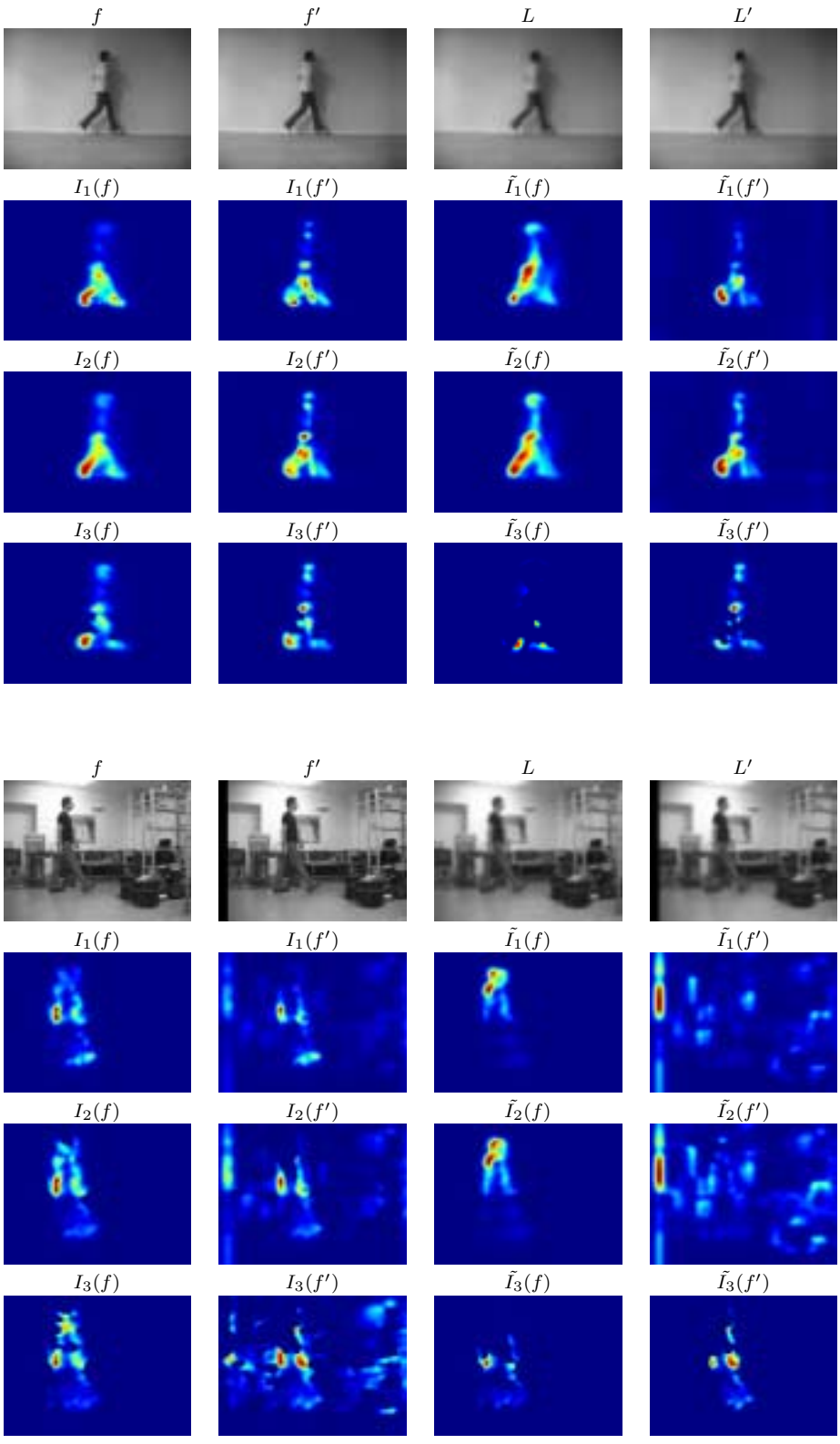

Figure 4: *The result of computing maps of the Galilean-corrected spatio-temporal interest operators I_1, I_2 and I_3 as well as corresponding non-diagonalized descriptors for two real image sequences subjected to a synthetic Galilean transformation with $u = 2.5$ pixels/frame. As can be seen from a visual comparison, the Galilean-corrected entities in the left columns give a better approximation to Galilean invariance than the corresponding non-diagonalized entities in the right columns. (Note! This figure should be viewed in colour.)*

Quantitative evaluation of approximation of Galilean invariance. To evaluate the stability of these descriptors under relative motions, we subjected a set of image sequences to synthetic Galilean transformations $u \in \{1, 2, 3\}$. Thus, given an image sequence f, we computed a Galilean transformed image sequence $f' = G_u f$, and computed spatio-temporal scale-space representations of L and L', respectively. Then, we computed maps M_f and $M_{G_u f}$ of each spatio-temporal interest operator from the original as well as the Galilean transformed image sequences. As can be seen from the illustration in figure 4, 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 $\tilde{I}_3$. To form a quantitative measure on the difference in terms of deviations from Galilean invariance, we computed the following correlation error measure

$$E(M) = C(M_f, M_{G_u f}) = \frac{\sum_{p \leftrightarrow p'} (M_f(p) - M_{G_u f}(p'))^2}{\sqrt{\sum_p M_f(p)^2} \sqrt{\sum_{p'} M_{G_u f}(p')^2}} \tag{59}$$

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 for a few examples.

$C(M_f, M_{G_u f})$	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 synthetic Galilean transformations for two sample 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 representation. The values are geometric averages computed over a set of seven image sequences. (The geometric standard deviation is given within parentheses.) As can be seen, the ratios are largest for small image velocities and decrease with increasing velocity, indicating that in combination with a space-time separable scale-space, the relative compensatory effect of Galilean-diagonalization is largest for small image velocities and decreases with increasing image velocity. Moreover, the compensatory effect is usually larger for a cluttered background.*

7 Extensions

7.1 Spatio-temporal interest operators for colour images

With minor modifications, the ideas behind these interest operators can also be extended to colour images, in order to make use of the additional information available in colour channels in situations when there is poor contrast in the grey-level information. In order to develop a corresponding scheme for colour cues, let us make use of the close analogue between the equations for determining the Galilean-diagonalization of the second-moment matrix (22) and the solution of the equations for optic flow according to (Lukas & Kanade 1981).

Given an RGB colour image, let us first transform this image into three colour channels $C^{(1)}C^{(2)}C^{(3)}$, which separate the intensity information $C^{(1)}$ from two chromatic channels $C^{(2)}C^{(3)}$ according to[6]

$$\begin{pmatrix} C^{(1)} \\ C^{(2)} \\ C^{(3)} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & -1 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \tag{60}$$

Next, in analogy with the least-squares formulation of the Lukas and Kanade method, let us differentiate the constant brightness assumption for each individual channel, $C^{(i)} = C_0^{(i)}$,

$$u_x \partial_x C^{(i)} + u_y \partial_y C^{(i)} + \partial_t C^{(i)} = 0 \tag{61}$$

and let us integrate the square of this relation using a window function w, and sum up these expressions over all channels i in order to state the following least-squares problem for determining $u = (u_x, u_y)$ at any point $p = (x, y, t)^T$:

$$\min_{u=(u_x,u_y)} \sum_{i \in \{1,2,3\}} \int_{q \in \mathbb{R}^3} \Big(u_x \partial_x C^{(i)}(p-q) + u_y \partial_y C^{(i)}(p-q) + \partial_t C^{(i)}(p-q) \Big)^2 w(q)\, dq \tag{62}$$

[6]This colour space, which can be seen as a slight variation of the more common linear colour transformation ($C_1 = R-G, C_2 = G-B$) has been previously been used in connection with Gaussian derivative operators by (Hall 2001) and has the qualitative effect of approximating red-green and blue-yellow colour opponent channels, as opposed to red-green and green-blue colour opponent channels which result from the more commonly used $R-G$- and $G-B$-transformation.

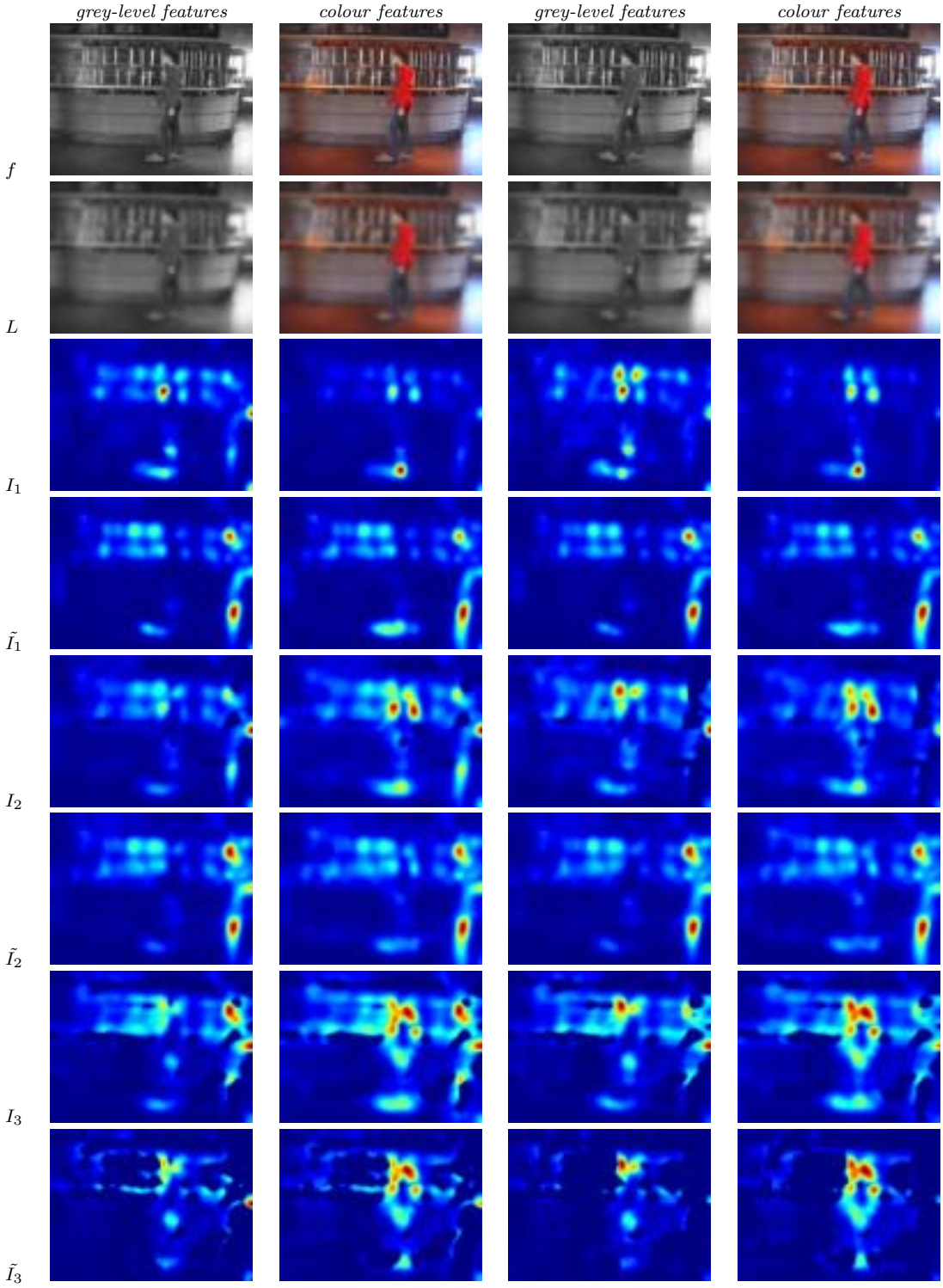

Figure 5: *The result of computing grey-level based as well as colour-based spatio-temporal interest operators for an image sequence with a walking person against a cluttered background, in which there is sometimes poor grey-level contrast between the moving object and the background. As can be seen, the use of colour-based spatio-temporal interest operators may give stronger responses for the moving objects. (Note! This figure should be viewed in colour.)*

By dropping the arguments of $C^{(i)}(p-q)$ and $w(q)$, and with

$$A = \sum_{i \in \{1,2,3\}} \int_{q \in \mathbb{R}^3} \begin{pmatrix} (C_x^{(i)})^2 & C_x^{(i)} C_y^{(i)} \\ C_x^{(i)} C_y^{(i)} & (C_y^{(i)})^2 \end{pmatrix} w \, dq, \tag{63}$$

$$b = \sum_{i \in \{1,2,3\}} \int_{q \in \mathbb{R}^3} \begin{pmatrix} C_x^{(i)} C_t^{(i)} \\ C_y^{(i)} C_t^{(i)} \end{pmatrix} w \, dq, \tag{64}$$

$$c = \sum_{i \in \{1,2,3\}} \int_{q \in \mathbb{R}^3} (C_t^{(i)})^2 w \, dq, \tag{65}$$

this least squares problem can be written

$$\min_u u^T A u + 2 b^T u + c \qquad \text{with the solution} \qquad u = -A^{-1} b \tag{66}$$

In the special case when there is only one colour channel, these are the equations for optic flow according to the method by (Lukas & Kanade 1981). In the case when we have multiple colour channels, we proceed as follows for expressing colour analogues to the previously expressed spatio-temporal interest operators.

1. Compute second moment matrices $\mu^{(i)}$ for all individual motion channels.

2. Sum up the elements in these in order to form:

$$A = \sum_i A^{(i)} = \begin{pmatrix} \mu_{xx}^{(i)} & \mu_{xy}^{(i)} \\ \mu_{xy}^{(i)} & \mu_{yy}^{(i)} \end{pmatrix}, \quad \text{and} \quad b = \sum_i b^{(i)} = \begin{pmatrix} \mu_{xt}^{(i)} \\ \mu_{yt}^{(i)} \end{pmatrix}. \tag{67}$$

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

$$(\mu'_{tt})^{(i)} = u_x^2 \mu_{xx}^{(i)} + 2 u_x u_y \mu_{xy}^{(i)} + u_y^2 \mu_{yy}^{(i)} + 2 u_x \mu_{xt}^{(i)} + 2 u_y \mu_{yt}^{(i)} + \mu_{tt}^{(i)}. \tag{68}$$

5. Sum up these entities over all colour channels to define the following analogue of the purely temporal diagonal element:

$$\nu_3 = \sum_i (\mu'_{tt})^{(i)} \tag{69}$$

6. Compute analogues to the spatial diagonal elements ν_1 and ν_2 from

$$\nu_1 + \nu_2 = \operatorname{trace} A, \quad \text{and} \quad \nu_1 \nu_2 = \det A. \tag{70}$$

7. Define I_1, I_2 and I_3 from ν_1, ν_2 and ν_3 in analogy with the previously stated equations (44), (47) and (49).

Figure 5 shows a few examples of computing spatio-temporal operators in this way for an image sequence with a cluttered background, for which the contrast is sometimes low between the moving object in the background. As can be seen, the use of complementary colour cues may give more prominent regions of interest in situations when there is poor contrast in terms of grey-level information only.

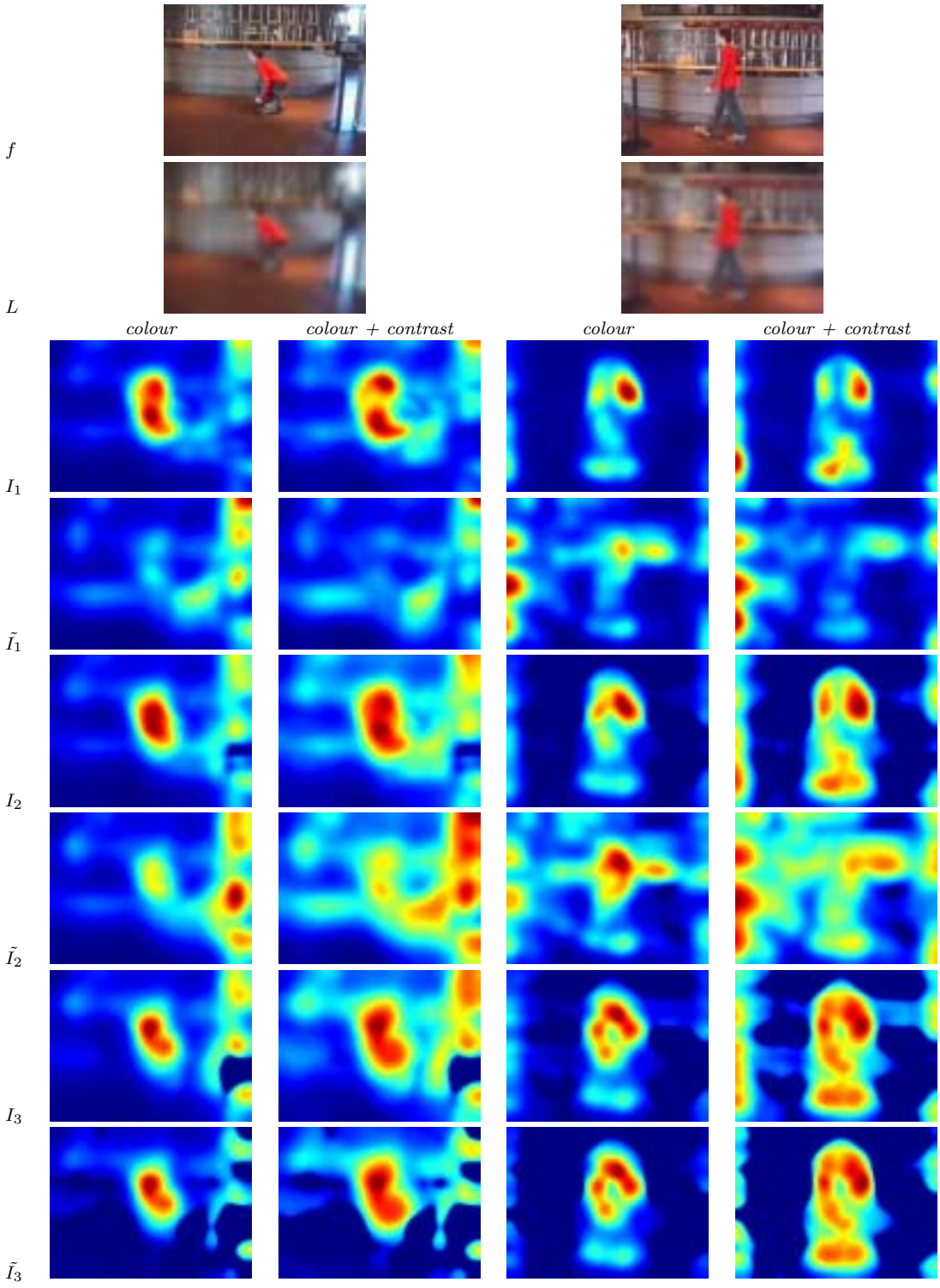

Figure 6: *Colour-based spatio-temporal interest operators computed for two image sequences with and without using complementary contrast normalization. (Note! This figure should be viewed in colour.)*

7.2 Contrast normalization

One limitation of the previously stated definitions of spatio-temporal interest operators is that the magnitude of the response is influenced by the local image contrast. Specifically, regions with high spatial contrast may result in relatively higher values of the magnitudes of I_k. To compensate for local intensity variations, a straightforward approach is to define contrast normalized interest operators according to

$$J_k = \frac{(I_k)^{1/k}}{(c_1 + \nu_1 + \nu_2)^{\alpha}} \tag{71}$$

where $\alpha \in [0, 1]$ is a constant and the purpose of the exponentiation operator $(I_k)^{1/k}$ is to compensate for the different dimensionality of I_1, I_2 and I_3 in terms of the diagonal elements ν_1, ν_2 and ν_3 of the Galilean diagonalized second-moment matrix. The entity $\nu_1 + \nu_2 = \operatorname{trace} A$ serves as a measure of the local image contrast (the average gradient magnitude within the support region of the window function), and the constant c_1 serves as a soft threshold for avoiding the division with small values.[7]

Figure 6 shows the result of computing such contrast normalized spatio-temporal interest operators J_k with corresponding unnormalized entities I_k, for two colour image sequences.[8] As can be seen, the contrast normalized spatio-temporal interest operators result in comparably stronger operator responses at the moving objects as well as in small regions around them. Hence, we propose contrast normalization as a complementary mechanisms when using the Galilean-corrected interest operators for computing motion masks for subsequent motion recognition.

8 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 of spatio-temporal interest operators considered here, however, the notion of Galilean diagonalization is of much 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) for performing local normalization of space-time structures for subsequent spatio-temporal recognition.

Concerning actual spatio-temporal scale-space concept, we have here considered Gaussian smoothing kernels. While an implementation based on this notion is not time-causal, time-causality is straightforward to achieve by replacing the Gaussian temporal smoothing operation by time-causal recursive filters. In on-going work, we

[7] In this work, we have estimated c_1 as a constant times the average gradient magnitude over a typical image sequence, i.e. $c_1 = c_2 \overline{\operatorname{trace} A}$, with $c_2 = 1/4$. Of course, other robust contrast estimation schemes could also be used.

[8] In these experiments, we used $\alpha = 3/4$ and slightly coarser scales $\sigma = 2$, $\tau = 5/4$ and $\gamma = 3$.

are developing scale selection mechanisms to be used in conjunction with these spatio-temporal interest operators, as well as spatio-temporal recognition schemes that work upon regions of interest in space-time defined from the responses of these operators.

A Appendix

A.1 Galilean invariance property of Galilean diagonalization

In this appendix, we give a formal proof of the fact that Galilean diagonalization according to (21) is preserved under Galilean transformations.

Consider a spatio-temporal image $f(p) = f(x, y, t)$ with spatio-temporal scale-space representation $L(p;\ \Sigma)$ defined from $L(\cdot;\ \Sigma) = g(\cdot;\ \Sigma) * f(\cdot)$, where g denotes a spatio-temporal Gaussian kernel (3) and Σ is a covariance matrix according to (2).

Given any velocity vector u, define a Galilean transformed image f'' by $f''(p'') = f(p)$, where $p'' = G(u)\, p$ and $G(u)$ denotes a Galilean transformation with image velocity u. Moreover, define the spatio-temporal scale-space representation L'' of f'' according to $L''(\cdot;\ \Sigma'') = g(\cdot;\ \Sigma'') * f''(\cdot)$. Then, from the transformation property under Galilean transformations, it follows that $L''(p''\ \ \Sigma'') = L(p;\ \Sigma)$ provided that the covariance matrices satisfy $\Sigma'' = G \Sigma G^T$. Analogous to equations (11) and (12), let us define second-moment matrices of L and L'' according to

$$\mu(p;\ \Sigma) = \int_{q \in \mathbb{R}^3} (\nabla L(q))(\nabla L(q))^T\, g(p - q;\ \Sigma)\, dq, \tag{72}$$

$$\mu''(p'';\ \Sigma'') = \int_{q'' \in \mathbb{R}^3} (\nabla L''(q''))(\nabla L''(q''))^T\, g(p'' - q'';\ \Sigma'')\, dq''. \tag{73}$$

Then, the second-moment matrices μ and μ'' are related according to

$$\mu'' = G^{-T}(u)\, \mu\, G^{-1}(u). \tag{74}$$

Let us next assume that v is a velocity vector that transforms μ into a block diagonal matrix. In other words, assume that we have a Galilean transformation $G(v)$ such that

$$\mu' = G^{-T}(v)\, \mu\, G^{-1}(v). \tag{75}$$

Our next goal is to find a Galilean transformation that transforms μ'' into block diagonal form. From equation (74) we can rewrite μ as $\mu = G^T(u)\, \mu''\, G(u)$, which after insertion into (74) gives

$$\mu' = G^{-T}(v)\, G^T(u)\, \mu''\, G(u)\, G^{-1}(v). \tag{76}$$

Since Galilean transformation matrices satisfy $G^{-1}(v) = G(-v)$ as well as $G(u - v) = G(u)\, G(-v)$, it follows that

$$\mu' = G^{-T}(v - u)\, \mu''\, G^{-1}(v - u) \tag{77}$$

and we have that the Galilean transformation $G(v - u)$ brings μ'' into block diagonal form. Thus, the property of block diagonalization is preserved under Galilean transformations. Specifically, the velocity vector associated with the Galilean transformation that brings a second-moment matrix into block diagonal form is additive under superimposed Galilean transformations. Therefore, if we normalize local space-time

structures using a local Galilean transformations determined from the requirement that the second-moment matrix should be block diagonal, it follows that the result after normalization will always be the same, irrespective of any superimposed Galilean transformation. From this view-point, the notion of Galilean diagonalization can be regarded as a canonical way of normalizing local space-time structures.

Note that although a similar result could be expected from the viewpoint of optic flow computations according to the method by (Lukas & Kanade 1981), we have in this proof not made any assumption that the local spatio-temporal image structures within the support region of the window function should represent a local translational model. (The optic flow estimation method by Lukas and Kanade is derived from such an assumption.) Therefore this result applies to arbitrary types of space-time structures and spatio-temporal events.[9]

A pre-requisite for carrying out this proof, however, is that the spatio-temporal second moment matrices used for computing the second-moment matrices are related according to $\Sigma'' = G \Sigma G^T$ (and for the Galilean diagonalization that $\Sigma' = G \Sigma G^T$). Thus, perfect Galilean invariance can only be expected if the shapes of the spatio-temporal smoothing kernels are coupled according to this relation. Otherwise, the relation will only be approximate. One way of achieving full Galilean invariance is therefore by considering scale-space smoothing over the full family of space-time kernels. An alternative approach is to adapt the spatio-temporal smoothing kernels to the local space-time image structures (Lindeberg & Gårding 1997, Lindeberg 1997, Nagel & Gehrke 1998, Mikolajczyk & Schmid 2002, Laptev & Lindeberg 2004*b*).

A.2 Interpreting Galilean diagonalization with average normal flow

In situations when motion constraint equation is valid, Galilean diagonalization implies that a weighted average of the normal flow vectors will be zero within the support region of the window function used for computing the second-moment matrix. To state and prove this property, let $\nabla_{space} L = (L_x, L_y)^T$ denote the spatial gradient vector and let u denote the optic flow. Then, the optic flow constraint equation can be written

$$L_t + (\nabla_{space} L)^T u = 0. \tag{78}$$

By multiplying this expression by the spatial gradient vector $\nabla_{space} L$ and integrating over the support region of the window function, we obtain

$$\int_{x,y,t \in \mathbb{R}^3} (\nabla_{space} L) \left(L_t + (\nabla_{space} L)^T u \right) g(x, y, t;\ \Sigma)\, dx\, dy\, dt = 0 \tag{79}$$

which with a more compact averaging operator E can written as

$$E((\nabla_{space} L)(L_t + (\nabla_{space} L)^T u)) = 0 \tag{80}$$

[9] The only assumption we have made above is that the purely spatial component of the second-moment matrix is non-singular, i.e., that $\mu_{xx} \mu_{yy} - \mu_{xy}^2 \neq 0$. If this assumption is violated, then the velocity vector u in the Galilean transformation $G(u)$ that diagonalizes μ is not uniquely determined, and we have a situation with a local aperture problem. This indeterminacy will, however, not effect the Galilean normalization, since the indeterminacy will not effect the transformed pattern. Therefore, we can for example choose the pseudo inverse to determine the velocity vector u from (23), and we will obtain either $\nu_1 = 0$ or $\nu_2 = 0$.

The case $\nu_1 = \nu_2 = 0$ is trivial, since $\nu_1 = 0$ and $\nu_2 = 0$ imply $L_x = L_y = 0$ in the entire support region of the window function, and therefore that $\mu_{xt} = \mu_{yt} = 0$.

According to our convention, Galilean diagonalization is achieved by when $E(L_x L_t) = E(L_y L_t) = 0$. In vector notation we have $E((\nabla_{space} L) L_t) = 0$. Hence, Galilean diagonalization implies

$$E((\nabla_{space} L)(\nabla_{space} L)^T u) = 0 \tag{81}$$

which can be interpreted as a weighed matrix average of the optic flow vectors being zero. To interpret this relation further, let us split the optic flow vector u into one component $u_\parallel$ parallel to the gradient vector and one component $u_\perp$ perpendicular, i.e. $u = u_\parallel + u_\perp$. Then, since $(\nabla_{space} L)^T u_\perp = 0$, it follows that

$$E((\nabla_{space} L)(\nabla_{space} L)^T u_\parallel)) = 0 \tag{82}$$

Due to the fact that $\nabla^T_{space} L$ and $u_\parallel$ are parallel, we have $(\nabla^T_{space} L) u_\parallel = |\nabla_{space} L||u_\parallel|$ and $(\nabla^T_{space} L)|u_\parallel| = |\nabla^T_{space} L| u_\parallel$. Hence, we can rewrite the previous relation as

$$E(|\nabla_{space} L|^2 \, u_\parallel) = 0 \tag{83}$$

which means that after Galilean diagonalization, the following weighted average of the normal flow vectors $u_\parallel$ will be zero:

$$\int_{x,y,t \in \mathbb{R}^3} |\nabla_{space} L|^2 \, u_\parallel \, g(x, y, t; \; \Sigma) \, dx \, dy \, dt = 0. \tag{84}$$

References

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

Bigün, J., Granlund, G. H. & Wiklund, J. (1991), 'Multidimensional orientation estimation with applications to texture analysis and optical flow', *IEEE Trans. Pattern Analysis and Machine Intell.* **13**(8), 775–790.

Davis, J. & Bobick, A. (1997), The representation and recognition of action using temporal templates, *in* 'Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition', Puerto Rico, pp. 928–934.

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

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

Hall, D. (2001), Viewpoint Independent Recognition of objects from local appearance, Phd thesis, INRIA Rhone-Alpes, 655 Ave de l'Europ, 38330 Montbonnot-St. Martin, France.

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

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

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

Laptev, I. & Lindeberg, T. (2003), Interest point detection and scale selection in space-time, *in* L. Griffin, ed., 'Proc. Scale-Space'03', Vol. 2695 of *Lecture Notes in Computer Science*, Springer-Verlag, Isle of Skye, Scotland, pp. 372–387.

Laptev, I. & Lindeberg, T. (2004*a*), Velocity adaptation of spatio-temporal interest points. *ICPR'04* (to appear).

Laptev, I. & Lindeberg, T. (2004*b*), 'Velocity-adapted spatio-temporal receptive fields for direct recognition of activities', *Image and Vision Computing* **22**(2), 105–116.

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. (1997), Linear spatio-temporal scale-space, *in* B. M. ter Haar Romeny, L. M. J. Florack, J. J. Koenderink & 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 from: http://www.nada.kth.se/cvap/abstracts/cvap257.html

Lindeberg, T. (2002), Time-recursive velocity-adapted spatio-temporal scale-space filters, *in* P. Johansen, ed., 'Proc. 7th European Conference on Computer Vision', Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, pp. 52–67.

Lindeberg, T. & Fagerström, D. (1996), Scale-space with causal time direction, *in* 'Proc. 4th European Conf. on Computer Vision', Vol. 1064, Springer Verlag, Berlin, Cambridge, UK, pp. 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, *in* 'Proc. 7th Int. Conf. on Computer Vision', Corfu, Greece, pp. 1150–1157.

Lukas, B. D. & Kanade, T. (1981), An iterative image registration technique with an application to stereo vision, *in* 'Image Understanding Workshop'.

Mikolajczyk, K. & Schmid, C. (2002), An affine invariant interest point detector, *in* 'Proc. 7th European Conference on Computer Vision', Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, pp. I:128–142.

Nagel, H. & Gehrke, A. (1998), Spatiotemporal adaptive filtering for estimation and segmentation of optical flow fields, *in* 'Proc. 5th European Conf. on Computer Vision', Springer-Verlag, Freiburg, Germany, pp. 86–102.

Niyogis, S. A. (1995), Detecting kinetic occlusions, *in* 'Proc. 5th Int. Conf. on Computer Vision', Cambridge, MA, pp. 1044–1049.

Wildes, R. & Bergen, J. (2000), Qualitative spatio-temporal analysis using an oriented energy representation, Vol. 1843, pp. II:768–784.

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

Zetzsche, C. & Barth, E. (1991), 'Direct detection of flow discontinuities by 3-D curvature operators', *Pattern Recognition Letters* **12**(12), 771–779.