Available online at www.sciencedirect.com

ScienceDirect

Computer Vision and Image Understanding 108 (2007) 207–229

Computer Vision and Image Understanding

www.elsevier.com/locate/cviu

Local velocity-adapted motion events for spatio-temporal recognition

Ivan Laptev [a,*], Barbara Caputo [b], Christian Schüldt [c], Tony Lindeberg [c]

[a] IRISA/INRIA, Campus de Beaulieu, 35042 Rennes, France
[b] IDIAP, Rue de Simplon 4, P.O. Box 592, 1920 Martigny, Switzerland
[c] Computational Vision and Active Perception Laboratory (CVAP), School of Computer Science and Communication KTH,
S-100 44 Stockholm, Sweden

Received 7 May 2005; accepted 6 November 2006
Available online 18 March 2007
Communicated by James Maclean

Abstract

In this paper, we address the problem of motion recognition using event-based local motion representations. We assume that similar patterns of motion contain similar events with consistent motion across image sequences. Using this assumption, we formulate the problem of motion recognition as a matching of corresponding events in image sequences. To enable the matching, we present and evaluate a set of motion descriptors that exploit the spatial and the temporal coherence of motion measurements between corresponding events in image sequences. As the motion measurements may depend on the relative motion of the camera, we also present a mechanism for local velocity adaptation of events and evaluate its influence when recognizing image sequences subjected to different camera motions.

When recognizing motion patterns, we compare the performance of a nearest neighbor (NN) classifier with the performance of a support vector machine (SVM). We also compare event-based motion representations to motion representations in terms of global histograms. A systematic experimental evaluation on a large video database with human actions demonstrates that (i) local spatio-temporal image descriptors can be defined to carry important information of space-time events for subsequent recognition, and that (ii) local velocity adaptation is an important mechanism in situations when the relative motion between the camera and the interesting events in the scene is unknown. The particular advantage of event-based representations and velocity adaptation is further emphasized when recognizing human actions in unconstrained scenes with complex and non-stationary backgrounds.

Keywords: Motion; Local features; Motion descriptors; Matching; Velocity adaptation; Action recognition; Learning; SVM

1. Introduction

Video interpretation is a key component in many potential applications within video surveillance, video indexing, robot navigation and human–computer interaction. This wide area of application motivates the development of generic methods for video analysis that do not rely on specific assumptions about the particular types of motion, environments and imaging conditions.

In recent years many successful methods were proposed that learn and classify motion *directly* from image measurements [6,53,11,12,47,62,48,8,3,46]. These direct methods are attractive due to the possibility of training motion models from the video data alone. In particular, using such methods recognition of human activities was shown to be possible without constructing and matching elaborated models of human bodies [11,62,3].

Direct methods to video analysis often rely on the dense motion measurements. To enable subsequent recognition with such methods, it is essential for the measurements in the test and the training data to correspond to some extent. A simple approach to ensure such correspondence is to accumulate all measurements in the video using global descriptors. Global representations, however, depend on the background motion and do not scale well to complex scenes. To avoid the background problem, many methods

[*] Corresponding author.
E-mail addresses: ilaptev@irisa.fr (I. Laptev), bcaputo@idiap.ch (B. Caputo), crilla@nada.kth.se (C. Schüldt), tony@nada.kth.se (T. Lindeberg).

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2006.11.023

deploy motion-based segmentation and compute motion descriptors in segmented regions. Complex scenes with non-rigid backgrounds and motion parallax, however, often make motion-based segmentation unreliable and distract subsequent recognition.

In this work, we focus on a *local* approach to motion recognition. One of the main goals of our method is to avoid the need of segmentation and to enable motion recognition in complex scenes. As a motivation, we observe that local space-time neighborhoods often contain discriminative information. A few examples of such neighborhoods for image sequences with human actions are illustrated in Fig. 1. Here, the similarity of motion in corresponding neighborhoods can be observed despite the difference in the appearance and the gross motions of people performing the same type of action. At the same time, the dissimilarity of image data is evident for non-corresponding neighborhoods. From this example it follows that some of the spatio-temporal neighborhoods may provide sufficient information for identifying corresponding space-time points across image sequences. Such correspondences could be useful for solving different tasks in video analysis. In particular, local correspondence in space-time could be used to formulate methods for motion recognition that do not rely on segmentation and, hence, could be applied to complex scenes.

To investigate this approach and to find corresponding points in space-time, we exploit the spatial and the temporal consistency or *coherence* of motion measurements between pairs of space-time neighborhoods. Considering all the pairs of neighborhoods for a given pair of sequences is computationally hard. Moreover, neighborhoods with simple motions and simple spatial structures may be ambiguous and may not allow for reliable matching when using local image information only. To address this problem, we select informative neighborhoods with low accidental similarity by maximizing the local spatio-temporal variation of image values over space and time. The detection of such neighborhoods, denoted here as *local motion events*, has been recently proposed by Laptev and Lindeberg [26] and is summarized in Section 3.

Local motion events (or simply events) are defined in this paper by the position and the shape of associated space–time neighborhoods. Both the shape and the position of events in video may depend on the recording conditions such as the relative distance and the relative velocity of the camera with respect to the object. Hence, to exploit

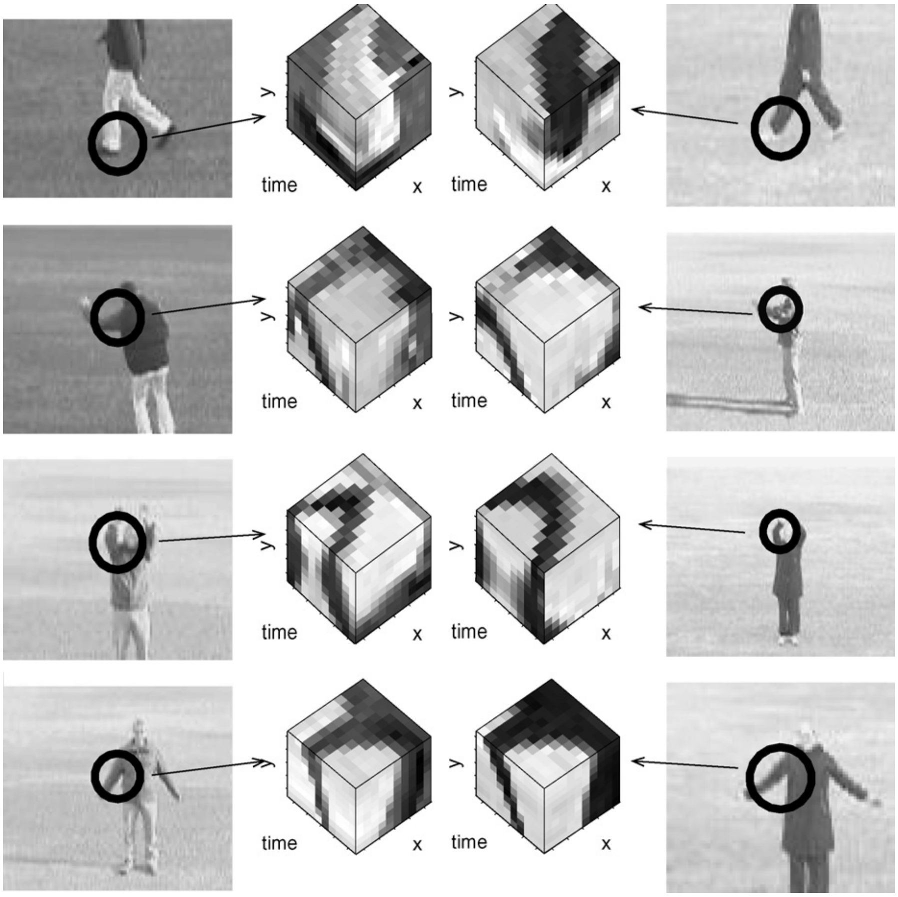


Fig. 1. Local space–time neighborhoods for corresponding space–time points in image sequences: ‘‘walking’’, ‘‘boxing’’ and ‘‘hand waving’’. The motion in corresponding neighborhoods is similar despite variations of the cloth and the global motion of people on the left and on the right. The neighborhoods from different actions have dissimilar motion.

the inherent motion properties of events, it is important to detect events independently of external transformations that effect the image sequences. Invariance of local motion events with respect to the scaling transformations has been previously addressed in [26]. Here, we extend this work and investigate event detection under Galilean transformations arising from the relative motion of the camera. A method for detecting motion events independently of the scale and Galilean transformations is presented in Section 3.

To match corresponding events in image sequences, we evaluate the coherence of motion measurements at pairs of space–time neighborhoods. For this purpose in Section 4 we formulate a set of alternative motion descriptors capturing motion information in the neighborhoods of detected events. Using these descriptors together with associated similarity measures we demonstrate the matching of corresponding events across image sequences in Section 5. Based on the estimated correspondences, we then define a nearest neighbor (NN) classifier and a support vector machine (SVM) classifier as two alternative methods for recognizing instances of motion classes. Fig. 2 summarizes the four steps of the method in this paper.

In Section 6 we evaluate different steps of the method. In particular the influence of local velocity adaptation as well as the choice of motion descriptors and recognition methods is analyzed on the problem of recognizing human actions in simple scenes. Results of human action recognition in complex scenes are then presented in Section 6.4. We conclude the paper with the discussion in Section 7.

This work is partly based on results previously presented in [27,28,51].

2. Related work

This work is related to several domains including motion-based recognition, local feature detection, adaptive filtering and human motion analysis. In the area of motion-based recognition, a large number of different schemes have been developed based on various combinations of visual tasks and image descriptors; see e.g. the monograph by Shah and Jain [52] and the survey paper by Gavrila [14] for overviews of early works. Concerning more recent approaches, Yacoob and Black [60] performed tracking and recognition using principal component analysis and parameterized models of optic flow. Hoey and Little [17] presented a related approach using Zernike polynomial expansions of optic flow. Bobick and Davis [5] recognized human actions against a static background by computing templates of temporal differences and characterizing the resulting motion masks in terms of moments. Zelnik-Manor and Irani [62] as well as Chomat et al. [8] recognized activities using probabilistic models of spatio-temporal receptive fields while Laptev and Lindeberg [29] extended this approach to histograms of locally velocity-adapted receptive fields. Another statistical, non-parametric approach for motion recognition in terms of temporal multi-scale Gibbs models was proposed by Fablet and Bouthemy [12]. Efros et al. [11] presented a recognition scheme in terms of positive and negative components of stabilized optic flow in spatio-temporal volumes. Several other papers have recently explored the structure of space–time volumes for motion representation and action recognition [4,61,21,15]. Regarding event detection, a more close approach to ours by Rao et al. [47] represented and recognized motion in terms of events detected as maxima of curvature of motion trajectories. Different to this method, our approach enables direct detection and matching of motion events without relying on tracking and detection of motion trajectories.

Detection of motion events in space–time is related to interest point detection in static images. Different formulations for interest points have been presented and used in the past [16,36,19,39]. Interest points and their image neighborhoods provide part-based representations of images with possibility to invariance to photometric transformations as well as to similarity and affine transformations [35,32,36,41]. Part-based image representations have been successfully applied to image and video indexing [50,54], wide base-line matching [56,41,55], object recognition [36,13] and other applications. Interest points in space–time have been recently proposed for motion representation in [26]. Here, we extend this work and apply space–time interest points to motion recognition. Our method for velocity-adaptation of motion events is particularly related to the methods of adaptive spatio-temporal filtering that have been considered in [43,33,29].

The motion descriptors introduced in Section 4 build upon several previous works. The use of the N-jet for expressing visual operations was proposed by Koenderink and van Doorn [23] and the first application to spatio-temporal recognition was presented in [8]. The use of histograms of receptive field responses goes back to the work by Schiele and Crowley [49] as well as Zelnik-Manor and Irani [62], and the use of PCA for optic flow was proposed by Black and Jepson [2]. The use of complementary position information in histograms is closely related to the position dependency in the SIFT descriptor by Lowe [36]. Recently, Ke and Sukthankar [20] added a local principal component analysis to the SIFT descriptor. The performance of local descriptors in spatial domain was experi-

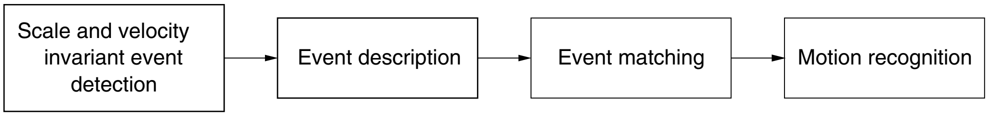


Fig. 2. Overview of the main steps of the method.

mentally evaluated by Mikolajczyk and Schmid in [42]. Here, we follow this experimental approach and evaluate local motion descriptors on the task of motion recognition.

3. Galilean- and scale-adapted event detection

Space–time interest points have recently been proposed to capture and represent local events in video [26]. Such points have stable locations in space–time and correspond to moving two-dimensional image structures at moments of non-constant motion (see Fig. 3a). A direct approach to detect such points consists of maximizing a measure of the local variations in the image sequence $f(x,y,t)$ over both space (x,y) and time t. To formulate such an approach, consider a scale-space representation $L(\cdot,\Sigma) = f * g(\cdot,\Sigma)$ generated by the convolution of f with a spatio-temporal Gaussian kernel

$$g(x,y,t;\Sigma) = \frac{1}{\sqrt{(2\pi)^3 \det(\Sigma)}} \exp\left(\left(-\tfrac{1}{2}(x,y,t)\Sigma^{-1}(x,y,t)^{\mathrm{T}}\right)\right)$$

with a 3×3 covariance matrix Σ. The image variation in a Σ-neighborhood of a space–time point $(\cdot)$ can now be measured by a second-moment matrix composed from spatio-temporal gradients $\nabla L = (L_x, L_y, L_t)^{\mathrm{T}}$

$$\mu(\cdot;\Sigma) = g(\cdot;s\Sigma) * (\nabla L(\nabla L)^{\mathrm{T}}) = \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{12} & \mu_{22} & \mu_{23} \\ \mu_{13} & \mu_{23} & \mu_{33} \end{pmatrix} \qquad (1)$$

integrated within a Gaussian window with the covariance $s\Sigma$ and some constant $s > 1$. Neighborhoods with μ of rank 3 correspond to points with significant variations of image values over both space and time. Points that maximize these variations can be detected by maximizing all eigenvalues $\lambda_1, \ldots, \lambda_3$ of μ or, similarly, by searching the maxima of the interest operator H [26] over (x,y,t)

$$H = \det(\mu) - k\ \mathrm{trace}^3(\mu) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3 \qquad (2)$$

where $k = 0.0005$ is a threshold on the discrepancy between $\lambda_1, \ldots, \lambda_3$.

3.1. Galilean transformations

The formulation of the interest operator H (2) in terms of the eigenvalues of μ implies invariance with respect to three-dimensional rotations of the space–time image f. Whereas two-dimensional rotations are common in the spatial domain, a three-dimensional rotation in space–time does not correspond to any known physical transformation. On the other hand, the temporal domain is frequently effected by Galilean transformations caused by the constant relative motion between the camera and the observed objects [33,29] (see Figs. 3(a) and (b)). A Galilean transformation is a linear coordinate transformation $p' = Gp$ with $p = (x,y,t)^{\mathrm{T}}$ defined by the velocity vector $(v_x, v_y)^{\mathrm{T}}$ and the matrix

$$G(v_x, v_y) = \begin{pmatrix} 1 & 0 & v_x \\ 0 & 1 & v_y \\ 0 & 0 & 1 \end{pmatrix} \qquad (3)$$

A Galilean transformation has a skewing effect on the image function $f'(p') = f(p)$ and the corresponding scale-space

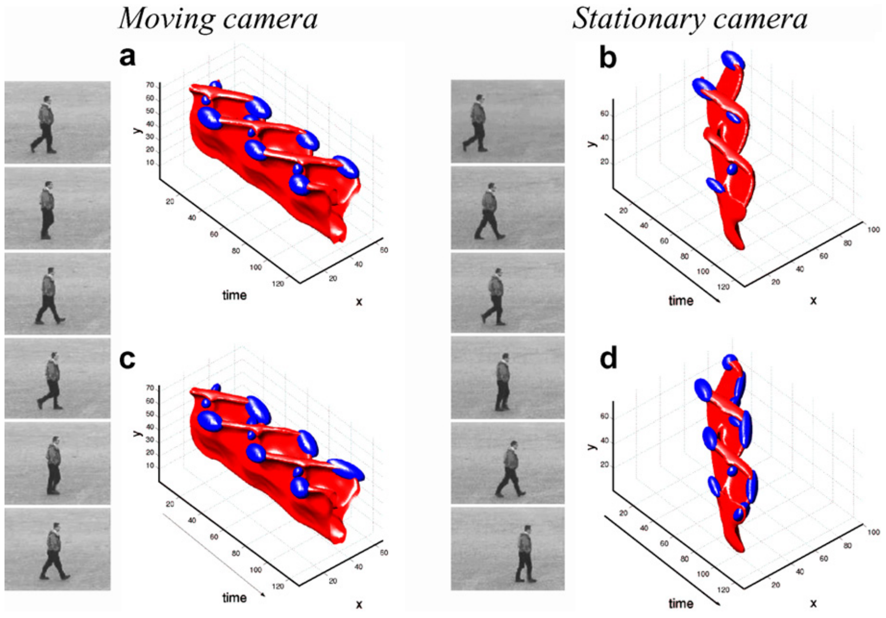


Fig. 3. Detection of local motion events for sequences with different camera motion. Spatio-temporal patterns of a walking person are shown by 3D plots (up-side-down) for (a) and (c) manually stabilized camera and (b) and (d) stationary camera. Motion events (blue ellipsoids) are detected using the original method [26] without velocity adaptation in (a) and (b) and with the proposed method for iterative velocity adaptation in (c) and (d). (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

representation $L'(p';\Sigma') = L(p;\Sigma)$. It can be shown [33], that the covariance matrix Σ of the filter kernel g transforms under G according to $\Sigma' = G\Sigma G^{\mathrm{T}}$ while the spatio-temporal gradient transforms as $\nabla L' = G^{-\mathrm{T}}\nabla L$. Using these properties, the transformation of μ (1) can be derived from [35] as

$$\mu'(p';\Sigma') = G^{-\mathrm{T}}\mu(p;\Sigma)G^{-1} \tag{4}$$

and it follows that μ and, hence, the interest operator H (2) is not preserved under Galilean transformation.

3.2. Velocity adaptation

Our goal is to detect positions and regions of local motion events independently of the relative velocity between the camera and a motion event in the scene. When using the μ-descriptor for event detection, it is essential to cancel the effect of the Galilean transformation and to compute H from the Galilean-invariant second moment matrix. In the case of G being known in advance this can be done by applying the inverse Galilean transformation to μ' as $G^{\mathrm{T}}\mu' G$. For the general case with unknown G we propose to transform μ' into a *standard* Galilean-invariant form. We make the following definitions:

Definition 3.1. Given space–time image functions f_a and f_b, we say that f_b is Galilean-related to f_a if $f_b(Gp) = f_a(p)$ for some Galilean transformation G.

Definition 3.2. Given second-moment matrices μ_a and μ_b, we say that μ_b is Galilean-related to μ_a (denoted here as $\mu_b \overset{G}{\rightarrow} \mu_a$) if μ_a and μ_b can be derived from the corresponding Galilean-related image functions f_a,f_b: $f_b(Gp) = f_a(p)$ using covariance matrices Σ_a, Σ_b: $\Sigma_b = G\Sigma_a G^{\mathrm{T}}$.

It follows that the Galilean-related second moment matrices satisfy (4). It is easy to show that the Galilean relation is transitive, i.e. for second moment matrices μ_a, μ_b, μ_c satisfying $\mu_b \overset{G_{ba}}{\rightarrow} \mu_a$, $\mu_c \overset{G_{cb}}{\rightarrow} \mu_b$ it holds that $\mu_c \overset{G_{ca}}{\rightarrow} \mu_a$ with $G_{ca} = G_{cb}G_{ba}$.

Proposition 3.3. Within the subset of Galilean-related (non-degenerative) second moment matrices there exists a unique matrix with the block-diagonal form

$$\mu'' = \begin{pmatrix} \mu''_{11} & \mu''_{12} & 0 \\ \mu''_{12} & \mu''_{22} & 0 \\ 0 & 0 & \mu''_{33} \end{pmatrix} \tag{5}$$

The proof of Proposition 3.3 is given in Appendix A. Using this proposition we can remove the ambiguity introduced by Galilean transformations if we for a given second-moment matrix μ' find a Galilean-related block-diagonal matrix μ'' (5) and then use it for event detection. For this purpose, we use relation (4) and solve for $G(v_x, v_y)$ that brings μ' into the block-diagonal form

$$\mu'' = G^{-T}\mu' G^{-1} \tag{6}$$

The solution for $G(v_x, v_y)$ is found from the linear system

$$\begin{pmatrix} \mu'_{11} & \mu'_{12} \\ \mu'_{12} & \mu'_{22} \end{pmatrix}\begin{pmatrix} v_x \\ v_y \end{pmatrix} = \begin{pmatrix} \mu'_{13} \\ \mu'_{23} \end{pmatrix} \tag{7}$$

as

$$v_x = \frac{\mu'_{22}\mu'_{13} - \mu'_{12}\mu'_{23}}{\mu'_{11}\mu'_{22} - (\mu'_{12})^2}, \quad v_y = \frac{\mu'_{11}\mu'_{23} - \mu'_{12}\mu'_{13}}{\mu'_{11}\mu'_{22} - (\mu'_{12})^2} \tag{8}$$

To obtain a unique μ'', however, we need to ensure Galilean relation to hold among all μ' derived for all Galilean-related image functions. Such a Galilean relation will be enforced if we compute μ' using

$$\Sigma' = G^{-1}\Sigma'' G^{-T} \tag{9}$$

for G satisfying (6) and for some fixed covariance matrix Σ''. Since G is unknown prior to the computation of μ', however, the Galilean relation among μ' and, hence, the unique Galilean-invariant μ'' cannot be obtained directly. An iterative solution to this problem will be presented in Section 3.3. Here, we note that for some initial guess of Σ', the descriptor μ'' obtained in (6) can be regarded as approximately invariant under Galilean transformations. Hence, we define a *velocity-corrected* interest operator in terms of μ'' (6) as

$$H_{\mathrm{corr}} = \det(\mu'') - k\ \mathrm{trace}^3(\mu'') \tag{10}$$

Remark 3.1. We can note that the solution for $G(v_x, v_y)$ in (7) and (8) is structurally similar to optic flow equations according to Lucas and Kanade [38,18]. Hence, $(v_x, v_y)^{\mathrm{T}}$ can be regarded as a local estimate of image velocity. Note, however, that here we did not use the brightness change constraint deployed in most of the methods for optical flow estimation. The velocity adaptation presented here, hence, can be applied to image functions with the brightness constancy over time violated.

Remark 3.2. By expanding μ'' in (6) using the solution of G (8) the component μ''_{33} of μ'' can be written as

$$\mu''_{33} = \mu'_{33} + \frac{2\mu'_{13}\mu'_{12}\mu'_{23} - \mu'_{22}(\mu'_{13})^2 - \mu'_{11}(\mu'_{23})^2}{\mu'_{11}\mu'_{22} - (\mu'_{12})^2} \tag{11}$$

It can be shown that μ''_{33} corresponds to the residual error term of optic flow estimation according to Lucas and Kanade [38] caused by image measurements violating the estimated constant motion model. Hence, μ''_{33} can be regarded as a measure of non-constant motion in the local neighborhood.

3.3. Scale- and velocity-adapted motion events

Following the previous section, estimation of Galilean-invariant μ'' from $\mu'(\cdot, \Sigma')$ requires G satisfying both (6) and (9). Since G is derived from μ' while μ' depends on G, this "chicken-and-egg" problem cannot be solved

```
Detect local motion events $P = \{p_1, ..., p_N\}$, $p = (x, y, t, \Sigma')$,
as positive space-time maxima of $H_{corr}$ (10) using $\Sigma'$ derived
from (9),(12),(3), with velocity values $v_x = v_y = 0$ and some initial
values of spatial and temporal scales $\sigma, \tau$
for each $p \in P$
    do  Update spatial and temporal scale values $\sigma \leftarrow \sigma + \delta_\sigma$,
        $\tau \leftarrow \tau + \delta_\tau$ such that the normalized Laplacian $\nabla^2_{norm} L =$
        $\sigma^2 \tau^{1/2}(L_{xx} + L_{yy}) + \sigma \tau^{3/2} L_{tt}$ obtains extremum in the
        scale neighborhood: $\delta_\sigma \in (-\varepsilon_\sigma, \varepsilon_\sigma)$, $\delta_\tau \in (-\varepsilon_\tau, \varepsilon_\tau)$
        (see [26] for more details)
        Update the velocity values $v_x, v_y$ according to (8)
        Update $\Sigma'$ (9), $\mu''$ (6) using new estimates of $v_x, v_y, \sigma, \tau$
        Update the position $x, y, t$ by maximizing $H_{corr}$ (10) over
        space and time
    until convergence of parameter values $\sigma, \tau, v_x, v_y, x, y, t$ or
        max. number of iterations
```

Fig. 4. Algorithm for scale- and velocity-adapted detection of local motion events.

directly. An iterative solution for Galilean-invariant μ'' can be formulated as follows. Without loss of generality we assume

$$\Sigma'' = \begin{pmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_0^2 & 0 \\ 0 & 0 & \tau_0^2 \end{pmatrix} \qquad (12)$$

and compute Σ' (9) for some initial guess of $G(v_x, v_y)$ (e.g. with $v_x = 0, v_y = 0$). At each iteration we then (i) re-estimate $G(v_x, v_y)$ from (8), (ii) update Σ' with (9) and (iii) re-detect the position of the event by maximizing H_{corr} (10) using new estimation of μ''. In addition to Galilean-invariance we can achieve local adaptation to the changes of the spatial and the temporal scale of the image function using the scale-selection approach in [26]. For this purpose, we at each iteration update the scale parameters σ, τ and stop iterating when the velocities, the scales and the position of the event converge to stable values. The algorithm for detecting scale- and velocity-adapted motion events is summarized in Fig. 4.

Remark 3.3. Iterative velocity adaptation of motion events described above bears close similarity with the adaptation of spatial interest points [16] with respect to affine transformations in the image plane [35,41]. In fact, the proposed velocity adaptation in space–time could be combined with the affine adaptation in space by estimating the affine transformation from the spatial part of μ'.

3.4. Qualitative evaluation

While the quantitative evaluation of velocity adaptation will be presented in Section 6.1, we will here discuss some qualitative results. An intuitive idea about the effect of velocity adaptation is illustrated in Fig. 3. Two sequences of a walking person have been recorded with a camera stabilized on the person (Figs. 3(a) and (c)) and a stationary camera (Figs. 3(b) and (d)). As can be seen from the spatio-temporal plots, the space–time structure of sequences differs by a skew transformation originating from different motions of the camera. As result, motion events detected without velocity adaptation using [26] are highly influenced by the relative camera motion (compare detection results in Figs. 3(a) and (b)). On the other hand, velocity-adapted motion events illustrated in Figs. 3(c) and (d) have roughly corresponding positions and shapes.[1] Hence, velocity-adapted motion events can be expected to provide reliable matching of corresponding space–time points in image sequences with different relative motions of the camera. A quantitative evaluation of local velocity adaptation on the task of classifying human actions in image sequences will be presented in Section 6.

Fig. 5 shows more examples of detected motion events for image sequences with human actions. From these results we can visually confirm the stability of detected events with respect to repeating structures in image sequences. Moreover, by analyzing spatio-temporal neighborhoods of detected events in Fig. 6, we observe that different actions give rise to different types of motion events. Hence, the proposed method can be expected to provide

[1] Ellipsoidal shapes of features in Figs. 3(a) and (b) are defined by the covariance matrices Σ'' derived from the iterative velocity and scale adaptation procedure summarized in Fig. 4. It takes about twenty iterations on average for the algorithm to converge.

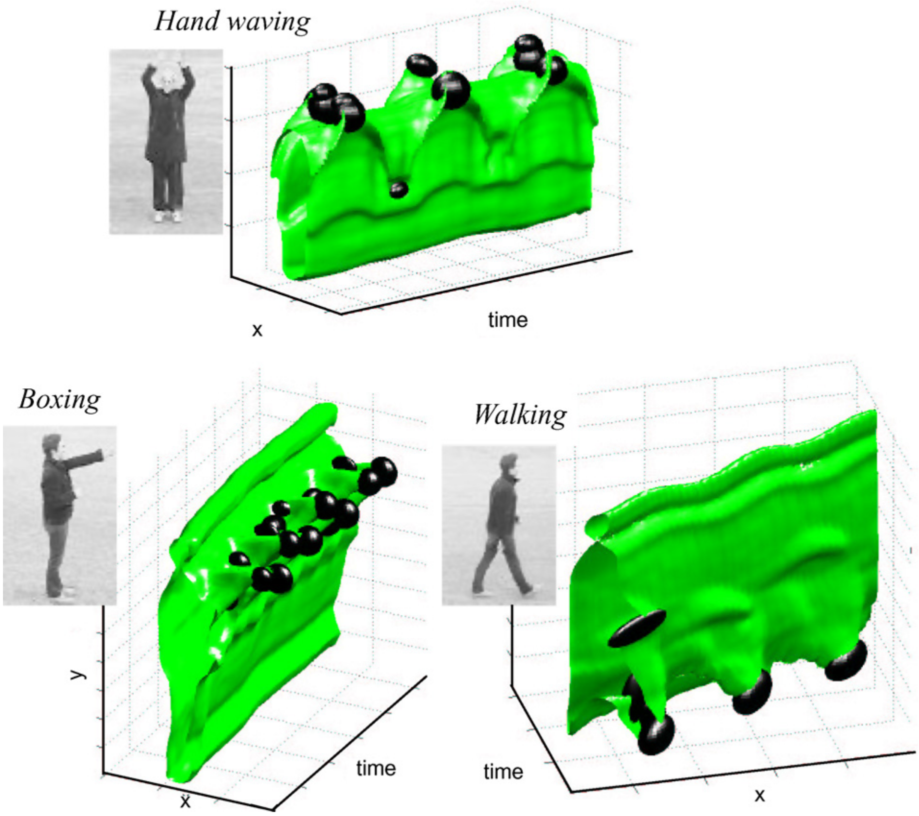


Fig. 5. Examples of scale- and velocity-adapted local motion events. The illustrations show one image from the image sequence and a level surface of image brightness over space–time. The events are illustrated as dark ellipsoids.

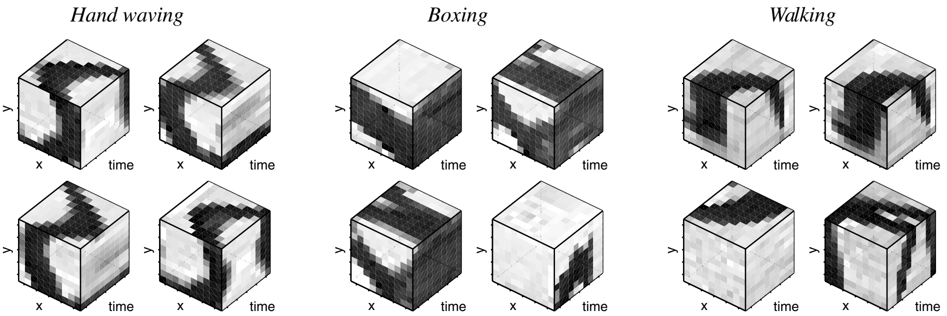


Fig. 6. Examples of spatio-temporal image patches corresponding to neighborhoods of local events detected for different actions in Fig. 5.

promising event candidates for the purpose of matching corresponding space–time points across image sequences.

4. Local descriptors for space–time neighborhoods

This section presents a set of alternative spatio-temporal image descriptors for the purpose of matching corresponding events in video sequences. To enable the matching, the event descriptors should be both discriminative and invariant with respect to irrelevant variations in image sequences. The method of previous section will be used here to adapt local motion events to scale and velocity transformations. Other variations, however, such as the individual variations of motion within a class might be more difficult to formalize since the criteria of optimality may depend on the task. For this reason we here take an empirical approach and define a set of alternative event descriptors whose performance will then be evaluated and compared in practice. The design of these descriptors is inspired by related work in the spatial domain [23,22,36,42] and in the spatio-temporal domain [60,17,62].

4.1. Image measurements

Differential measures are a common tool for describing local structure of the image [22]. As a basis for defining spatio-temporal image descriptors, we shall here make use of Gaussian derivatives. Taking advantage of the scale and velocity estimation in the previous section, we compute scale and velocity *adapted* Gaussian derivatives using the estimate of covariance matrix Σ' from the image sequence f. Provided the correct estimation of Σ', the responses of Gaussian derivatives will be invariant to scale and velocity variations in f if computed from $f''(p) = f(\Sigma'^{-\frac{1}{2}}p)$ that is a normalized image sequence obtained by transforming f to the common coordinate frame. In practice we achieve this by warping the neighborhoods of each detected event in the event-centered coordinate system by the linear coordinate transformation $p'' = c\Sigma'^{-\frac{1}{2}}p$ for some constant $c > 1$ using trilinear interpolation.

To construct invariant spatio-temporal event descriptors in terms of scale and velocity adapted Gaussian derivatives, we then consider the following type of image measurements:

- *N-jets* [23] up to order $N = 4$ (see Fig. 7) evaluated at the center of the detected motion events
$$J(\cdot;\Sigma') = \{L_x, L_y, L_t, L_{xx}, L_{xy}, \ldots, L_{tttt}\} \tag{13}$$
- *Gradient vector fields* obtained by computing vectors of adapted spatio-temporal gradients $(L_\xi\ L_\eta\ L_\zeta)^{\mathrm{T}}$ at every point in a local neighborhood of a motion event.
- *Optic flow fields* computed in the neighborhoods of motion events according to (8) from second-moment matrices defined in terms of adapted spatio-temporal gradients.

There is a number of qualitative similarities as well as differences between these types of image measurements. The N-jet contains a truncated encoding of the complete space–time image structure around the motion event, with an implicit encoding of the optic flow. Gradient vector field also approximates the space–time structure around motion events but without computing higher order derivatives that might be sensitive to noise. By explicitly computing the optic flow, we obtain a representation that is invariant to local contrast in the image domain, at the cost of possible errors in the flow estimation step. In addition to the optic flow, the N-jets and spatio-temporal gradients also encode the local spatial structure, which may either help or distract the recognition scheme depending on the relation between the contents in the training and the testing data. Hence, it is of interest to investigate all three types of image measurements.

4.2. Types of image descriptors

To combine dense flow measurements into image descriptors we consider:

- *Histograms* of either spatio-temporal gradients or optic flow computed at several scales. The histograms will be computed either for the entire neighborhood of a motion event, or over several $(M \times M \times M)$ smaller neighborhoods around the motion event. For the latter case, here referred to as *position dependent histograms*, local coordinates are measured relative to the position and the shape of the detected motion events (see Fig. 8). Local measurements are weighted using a Gaussian window function where we for simplicity com-

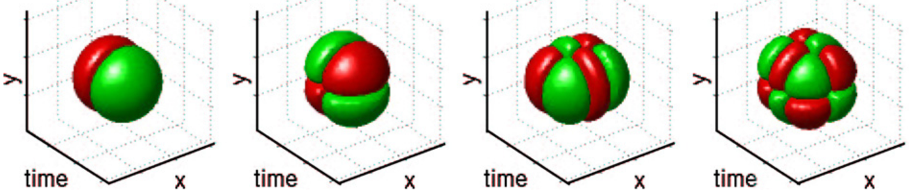


Fig. 7. Examples of impulse responses of spatio-temporal derivatives used to compute N-jet descriptors. The responses are illustrated by threshold surfaces with colors corresponding to different signs of responses. From left to right: $L_t, L_{yt}, L_{xxt}, L_{xytt}$. (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

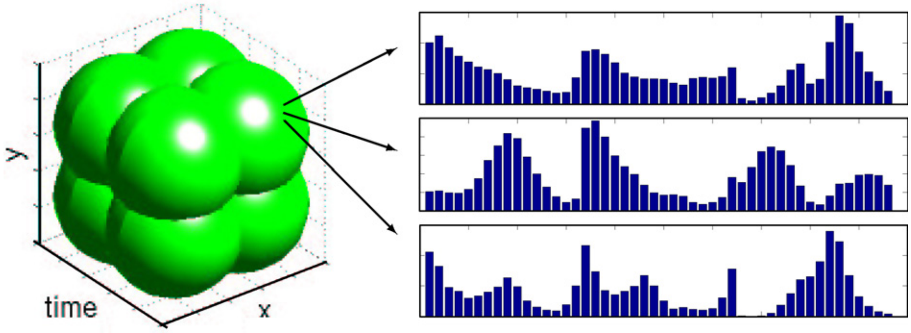


Fig. 8. Examples of position dependent histograms (right) computed for overlapping Gaussian window functions (left).

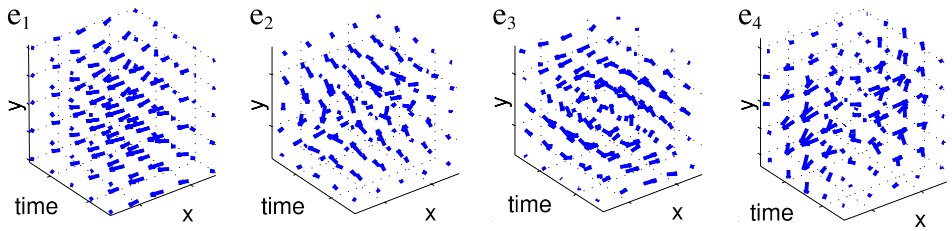


Fig. 9. The four most significant eigenvectors obtained by performing PCA on spatio-temporal gradient fields computed at the neighborhoods of motion events. Although the interpretation of the three-dimensional vector fields is somewhat difficult, we can observe increasing levels of details for the eigenvectors with lower eigenvalues.

pute one-dimensional (marginal) histograms by integrating the responses separately for each component of either the spatio-temporal gradient field or the optic flow field.

- *Principal component analysis* (PCA) of either optic flow or spatio-temporal gradient vectors computed over local scale and velocity normalized spatio-temporal neighborhoods around the motion events. The principal components are computed from local motion events detected in the training data, and the data is then projected to a lower-dimensional space defined by the eigenvectors corresponding to the largest eigenvalues. (see Fig. 9).

4.3. Spatio-temporal image descriptors

By combining the above mentioned notions in different ways, we consider the following types of descriptors for a space–time event $p(x,y,t;\ \Sigma')$ with position (x,y,t) and the neighborhood defined by Σ' (9) in terms of scale values σ, τ and velocity values v_x, v_y:

2Jets, 4Jets: N-jet of order 2 or 4 computed at (x_0,y_0,t_0) at the scale (σ_0,τ_0) according to (13)

MS2Jets, MS4Jets: Multi-scale N-jet of order 2 or 4, computed at (x_0,y_0,t_0) at all 9 combinations of 3 spatial scales $(\sigma_0/2,\sigma_0,2\sigma_0)$ and 3 temporal scales $(\tau_0/2,\tau_0,2\tau_0)$

STG-PDHIST: Local position dependent histograms of first-order partial derivatives

STG-HIST: Local position independent histograms of first-order partial derivatives

OF-PDHIST: Local position dependent histograms of optic flow

OF-HIST: Local position independent histograms of optic flow

STG-HIST: Local principal component analysis of spatio-temporal gradients vectors

OF-HIST: Local principal component analysis of optic flow

We also consider a global histogram-based descriptor as a reference with respect to the previous global schemes for spatio-temporal recognition:

Global-STG-HIST: Global histograms of first-order partial spatio-temporal derivatives computed over the entire image sequence using 9 combinations of 3 spatial scales and 3 temporal scales. This descriptor is closely related to [62]

To obtain affine contrast invariance, the N-jets as well as the spatio-temporal gradient vectors are normalized to unit l_2-norm. For the principal component analysis of the spatio-temporal gradient fields, the affine contrast normalization is performed at the level of scale normalized image volumes. Additional details of implementation of motion descriptors are summarized in Appendix B.

5. Matching and recognition

To find corresponding events based on the information in motion descriptors, it is necessary to evaluate the similarity of the descriptors. In this work, we use three alternative dissimilarity measures for comparing descriptors defined by the vectors d_1 and d_2:

- The normalized scalar product

$$S(d_1,d_2) = 1 - \frac{\sum_i d_1(i)d_2(i)}{\sqrt{\sum_i d_1^2(i)}\sqrt{\sum_i d_2^2(i)}} \tag{14}$$

- The Euclidean distance

$$E(d_1,d_2) = \sum_i (d_1(i) - d_2(i))^2 \tag{15}$$

- The χ^2-measure

$$\chi^2(d_1,d_2) = \sum_i \frac{(d_1(i) - d_2(i))^2}{d_1(i) + d_2(i)} \tag{16}$$

The normalized scalar product and the Euclidean distance can be applied for comparing any type of local space–time descriptors introduced above. The χ^2-measure

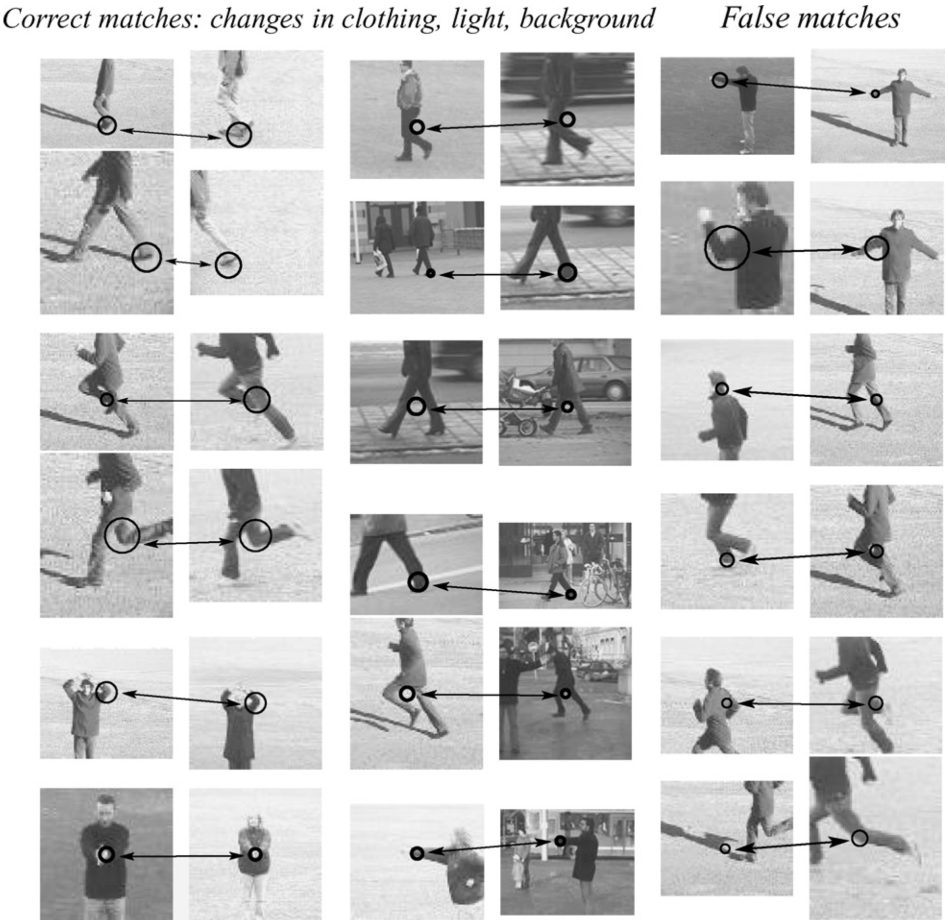


Fig. 10. Examples of matched local space–time features in sequences with human actions. The matches are generated by minimizing the dissimilarity measure (14) between STG-PDHIST descriptors.

will be used to compare histogram-based descriptors only: STG-HIST, OF-HIST, STG-PDHIST, OF-PDHIST.

Using the dissimilarity measures, we can match local events in the image sequences by searching for the pairs of events with the lowest dissimilarity of their descriptors. Fig. 10 presents matched events for some sequences with human actions. To generate the matches, we here used the STG-PDHIST descriptor in combination with the normalized scalar product. In Fig. 10(left) we observe that matches are found for similar parts (legs, arms and hands) at moments of similar motion. Note that the local nature of the descriptors enables a correct matching of similar events despite variations in the background, illumination and the cloth of people. Pure local information, however, is not always sufficient to discriminate between different types of actions and events as illustrated in Fig. 10(right).

5.1. Classification

Until now we have focused on the task of representing and matching individual events in image sequences. Given the problem of motion recognition, it is natural to combine evidence from several motion events for the purpose of final classification. In this section we define two alternative representations using combinations of motions events in image sequences. For these representation we then formulate a NN and a SVM classifier that will be used for motion recognition in the rest of this article.

Motion events originating from the same pattern of motion are likely to have joint properties within an image sequence in terms of relative positions in space–time. Such properties could be used to disambiguate the mismatches of local events and, hence, to increase the performance of recognition. Stable modelling of space–time arrangements of motion events, however, is not trivial due to the presence of noise and individual variations of motion patterns within a class. A similar problem of representing static objects in images using constellations of local features has been recently addressed for the task of object recognition in [37,30,13,54]. Currently, however, there exists no general solution to this problem.

To avoid stability issues, in this work we consider local properties of motion events only. Given n local events with descriptors d_i in the image sequence, we define two representations as:

LME: Unordered set of local descriptors (“bag of features”)

$$\mathscr{D} = \{d_1, \ldots, d_n\} \tag{17}$$

LMEHist: Histogram of quantized descriptors

$$\mathscr{H} = (h_1, \ldots, h_n) \tag{18}$$

with each histogram bin h_i corresponding to one quantization label. Quantization is obtained by K-mean clustering of descriptors in the training set. For a test sequence, each descriptor is assigned a quantization label $i = 1,\ldots,n$ corresponding to the label of the nearest cluster.

5.1.1. Nearest neighbor classifier

Given a set of training sequences for each motion class, we use a NN classifier to recognize motion classes in test sequences as follows. For LMEHist representations we compute the dissimilarity between histograms $\mathscr{H}$ (18) of the training and the test sequences using one of the dissimilarity measures in (14)–(16). The class of the test sequence is then determined by the class of the training sequence with the lowest dissimilarity.

For sequences represented by unordered sets of event descriptors $\mathscr{D}$ (17), we adopt the following greedy matching approach. Given sets of motion descriptors $\mathscr{D}_1$ and $\mathscr{D}_2$ in two sequences, the dissimilarity measure is evaluated for each pair of features (d_i^1, d_j^2), $d_i^1 \in \mathscr{D}_1$, $d_j^2 \in \mathscr{D}_2$ according to (14)–(16). The pair of events with the minimum dissimilarity is matched and the corresponding descriptors are removed from $\mathscr{D}_1$ and $\mathscr{D}_2$. The procedure is repeated until no more feature pairs can be matched, either due to a threshold on the dissimilarity or the lack of data. The dissimilarity between two image sequences is then defined by the sum of dissimilarities of N individual event matches. Given training sequences and the test sequence, the class of the test sequence is determined by the class of the training sequence with the lowest dissimilarity.

5.1.2. Support vector machines

Support vector machines (SVMs) are state-of-the-art large margin classifiers which have recently gained popularity within visual pattern recognition ([58,59] and many others). In this section we first give a brief overview of binary classification with SVMs (for the extension to multi-class settings and further details we refer the reader to [9,57]); then, we address the problem of using local descriptors in an SVM framework. For this purpose, we will describe a family of kernel functions that, in spite of not being Mercer kernels, can be effectively used in the framework of action recognition.

Consider the problem of separating the set of training data $(\boldsymbol{x}_1, y_1),\ldots,(\boldsymbol{x}_m, y_m)$, where $\boldsymbol{x}_i \in \Re^N$ is a feature vector and $y_i \in \{-1,+1\}$ its class label. If we assume that the two classes can be separated by a hyperplane $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, and that we have no prior knowledge about the data distribution, then the optimal hyperplane (as to say the one with the lowest bound on the expected generalization error) is the one which maximizes the margin [9,57]. The optimal values for $\boldsymbol{w}$ and b can be found by solving the following constrained minimization problem:

$$\min_{\boldsymbol{w},b} \tfrac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{subject to } y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geqslant 1, \quad \forall i = 1, \ldots, m \tag{19}$$

Solving (19) using Lagrange multipliers α_i gives a classification function

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{w} \cdot \boldsymbol{x} + b\right) \tag{20}$$

where α_i and b are found by the SVC learning algorithm [9,57]. Most of the α_is' take the value of zero; those $\boldsymbol{x}_i$ with non-zero α_i are the "support vectors". In cases where the two classes are non-separable, the solution can be found as for the separable case except for a modification of the Lagrange multipliers into $0 \leqslant \alpha_i \leqslant C$, where C is the penalty for the misclassification. To obtain a non-linear classifier, one maps the data from the input space $\Re^N$ to a high dimensional feature space $\mathscr{H}$ by $\boldsymbol{x} \to \Phi(\boldsymbol{x}) \in \mathscr{H}$, such that the mapped data points of the two classes are linearly separable in the feature space. Assuming there exists a kernel function K such that $K(\boldsymbol{x}, \boldsymbol{z}) = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{z})$, a non-linear SVM can be constructed by replacing the inner product $\boldsymbol{x} \cdot \boldsymbol{z}$ in the linear SVM by the kernel function $K(\boldsymbol{x}, \boldsymbol{z})$

$$f(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right) \tag{21}$$

This corresponds to constructing an optimal separating hyperplane in the feature space.

SVMs have proved effective for recognition of visual patterns like objects and categories using global descriptors [45,7,31]. Particularly, several authors have shown that Gaussian kernels

$$K_{\mathrm{Gauss}}(\boldsymbol{x}, \boldsymbol{z}) = \exp\{-\gamma \sum_i \|x_i - z_i\|^2\} \tag{22}$$

or χ^2 kernels [1]

$$K_{\chi^2}(\boldsymbol{x}, \boldsymbol{z}) = \exp\{-\gamma\chi^2(\boldsymbol{x}, \boldsymbol{z})\}, \quad \chi^2(\boldsymbol{x}, \boldsymbol{z}) = \sum_i \frac{|x_i - z_i|^2}{|x_i + z_i|} \tag{23}$$

perform well in combination with histogram representations [7,45]. Hence, in this paper we use kernels (22) and (23) within SVM when recognizing motion in image sequences represented by the histograms of quantized events $\mathscr{H}$ (18).

Now we turn to the problem of using SVMs with local motion events. Given the representation of two image sequences by the sets of local descriptors $\mathscr{D}_1$ and $\mathscr{D}_2$ (17), it is possible to use $\mathscr{D}_1$ and $\mathscr{D}_2$ as input for SVMs via a class of local kernels defined as [58]

$$K_L(\mathscr{D}_1,\mathscr{D}_2)=\tfrac{1}{2}[\hat{K}(\mathscr{D}_1,\mathscr{D}_2)+\hat{K}(\mathscr{D}_2,\mathscr{D}_1)] \tag{24}$$

with

$$\hat{K}(\mathscr{D}_1,\mathscr{D}_2)=\frac{1}{n_1}\sum_{i=1}^{n_1}\operatorname*{arg\,max}_{j=1,\ldots,n_2}\{K_l(d_i^1,d_j^2)\} \tag{25}$$

Different choices are possible for the local feature similarity kernel K_l, for instance:

$$K_l=\exp\left\{-\gamma\left(1-\frac{\langle \boldsymbol{x}-\boldsymbol{\mu}_x|\boldsymbol{z}-\boldsymbol{\mu}_z\rangle}{\|\boldsymbol{x}-\boldsymbol{\mu}_x\|\|\boldsymbol{z}-\boldsymbol{\mu}_z\|}\right)\right\} \tag{26}$$

or the Gaussian and χ^2 kernel given in (22) and (23). The family of kernels given in (24) and (25) relies on the matching of corresponding events in $\mathscr{D}_1$ and $\mathscr{D}_2$. For each local event with descriptor d_i^1 in the first sequence, Eq. (25) enforces the search for the best matching event with descriptor d_j^2 in the second sequence according to a similarity measure given by the local kernel K_l. Local kernels can enforce either one-to-one or one-to-many matching. It is also possible to enforce a threshold on the similarity value given by K_l, so to consider significant event matches only.

Despite the claim in [58], the kernels in (24) and (25) are not Mercer kernels. However, they have been empirically demonstrated to give highly useful results in visual applications such as object recognition and categorization [58,45]. Hence, we use this family of kernels for recognizing motion in image sequences represented by local motion events (LME).

6. Evaluation

In this section, we evaluate methods described in Sections 3–5, respectively. We perform evaluation using video sequences with six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed by different subjects in scenes with homogeneous and complex backgrounds. Scenes with homogeneous backgrounds (see Fig. 11) are used initially to evaluate velocity-invariance of adapted motion events in Section 6.1, the performance of event descriptors in Section 6.2 and the performance of event-based motion recognition in Section 6.3. Finally in Section 6.4 we evaluate the performance of event-based action recognition in complex scenes.

6.1. Evaluation of velocity adaptation

To evaluate the velocity adaptation procedure described in Section 3, we here study event detection for image sequences distorted by different amount of camera motion. In particular, we evaluate (i) the repeatability of motion events, (ii) the stability of event descriptors and (iii) recognition performance for different velocities of the camera. As test data we consider a subset of image sequences from the database in Fig. 11 transformed by the Galilean transformation $G(v_x,v_y)$ in (3) with $v_x=\{.5,1.0,1.5,2.0,2.5\}$ and $v_y=0$. The transformation was achieved by warping the original image sequences with trilinear interpolation and in this way simulating horizontal translation of the camera.[2] We compare event detection with and without velocity adaptation and evaluate the following methods:

Horig:	maxima of the space–time operator H (2) without neither scale nor velocity adaptation
Hcorr:	maxima of the velocity-corrected operator H_{corr} (10) without iterative scale and velocity adaptation
HcorrSc:	maxima of H_{corr} with iterative scale adaptation only according to [26]
HcorrVel:	maxima of H_{corr} with iterative velocity adaptation and no scale adaptation (algorithm in Fig. 4 without step 3)
HcorrScVel:	maxima of H_{corr} in combination with iterative scale and velocity adaptation according to the algorithm in Fig. 4

6.1.1. Repeatability

To evaluate the stability of event detection under Galilean transformations, we compute the number of corresponding (or repeating) events detected in different Galilean transformed sequences of the same scene. For this purpose, given the known value of G for each sequence, we transform the positions of the detected events into the original coordinate frame by $\tilde{p}=G^{-1}p$ and match $\tilde{p}$ with the position of the events detected in the original image sequence. The repeatability rate is then computed as a ratio between the number of matched features and the total number of features in both sequences.

Fig. 12 illustrates the repeatability averaged over all sequences in the test set and computed for different velocity transformations and for different methods of event detection. As can be seen, the curves cluster into two groups corresponding to high re-detection rates for events *with* iterative velocity adaptation and to lower re-detection rates for events *without* velocity adaptation. Hence, we confirm that velocity adaptation is an essential mechanism for stable detection of the events under velocity transformations in the data. By comparing the results of *Horig* and *Hcorr*, we also observe a slightly better repeatability of events

[2] Simulation of camera motion by means of interpolated warping was chosen here due to practical considerations. Obtaining real video data for the experiments in this section would require the recording of human actions to be done simultaneously with several cameras translating at different and well-controlled velocities. Although this type of data acquisition is possible in principle, it would require specific equipment that was not available at our disposal. We believe that the artifacts of interpolation do not effect the results of this section in a principled way.

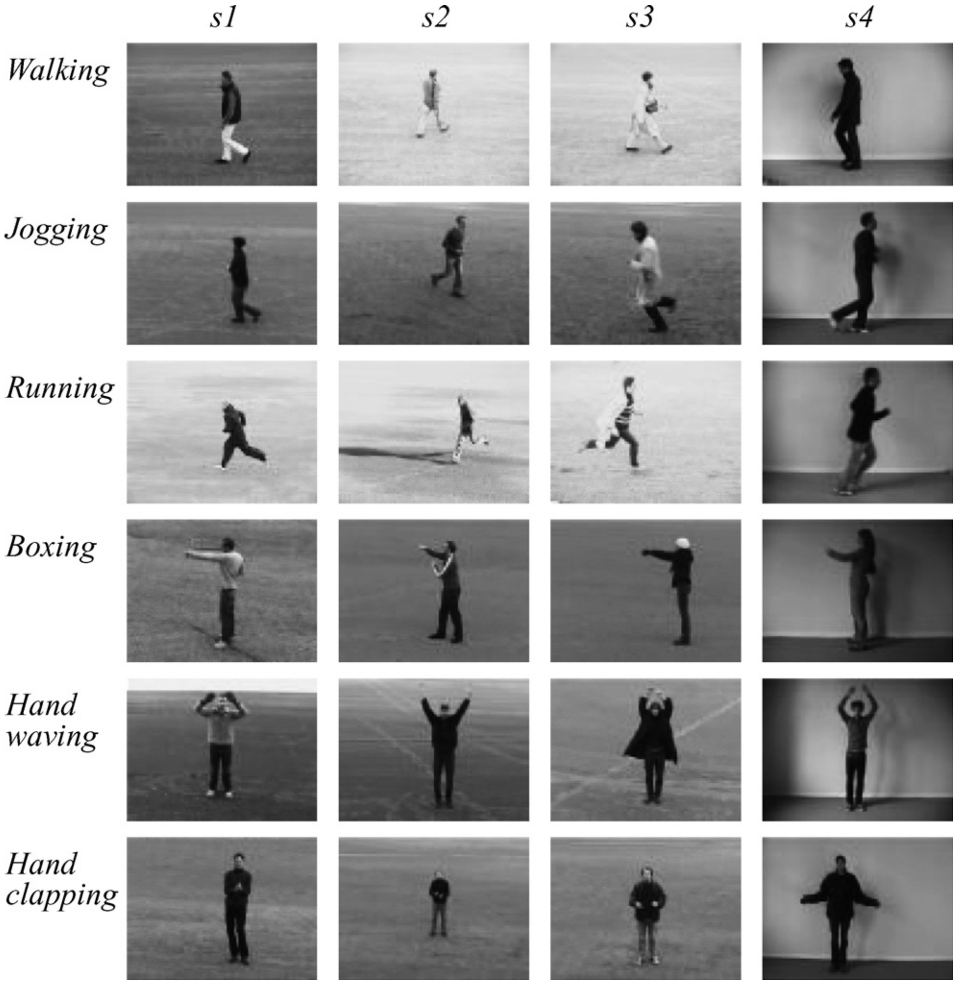


Fig. 11. Example frames from the human action database [51] with six classes of actions (walking, jogging, running, boxing, hand waving, hand clapping) performed by 25 subjects in four scenarios: (*s*1) outdoors, (*s*2) outdoors with scale variation, (*s*3) outdoors with different clothes and (*s*4) indoors. The database contains 2391 sequences in total. All sequences were taken over homogeneous backgrounds with a static camera with 25 fps frame rate. The sequences were down-sampled to the spatial resolution of 160 × 120 pixels and have a length of four seconds in average. The database is publicly available from http://www.nada.kth.se/cvap/actions/.

detected using the velocity-corrected operator (*Hcorr*). To restrict the number of evaluated detectors, we will use velocity-corrected detection only when evaluating the stability of image descriptors and the performance of recognition.

6.1.2. Stability of descriptors

Galilean transformations effect the shape of events in space–time and influence the values of the space–time descriptors. To compensate for velocity transformations, the covariance matrices Σ' of the filter kernels could be adapted to velocity values estimated either iteratively according to the algorithm in Fig. 4 or in "one-step" (8). The first approach is truly invariant under velocity transformations and is natural when computing image descriptors for velocity-adapted events (*HcorrVel*, *HcorrScVel*). The other approach is less demanding in terms of computations, at the cost of approximative invariance to velocity transformations. Such an approach is natural to combine with events detected without iterative velocity adaptation. For the evaluation we compute 4Jet-descriptors using: (i) filter kernels with iterative velocity adaptation for velocity-adapted events *HcorrVel*, *HcorrScVel*; (ii) filter kernels with one-step velocity adaptation for events *HcorrSc*, *Hcorr*; (iii) separable filter kernels without velocity adaptation for non-adapted events *Horig*, *Hcorr* here denoted as *HorigV0*, *HcorrV0*.

The stability of the descriptors is evaluated by computing the average Euclidean distance between pairs of descriptors for corresponding events. The pairs of corresponding events are determined as in the repeatability test above. The results of the evaluation are illustrated in Fig. 12(b). As can be seen, the most stable method with the least Euclidean distance between the descriptors corresponds to the combination of events and descriptors computed with the iterative velocity adaptation (*HcorrVel*, *HcorrScVel*). The performance of the descriptors with approximative velocity adaptation (*HcorrSc*, *Hcorr*) is

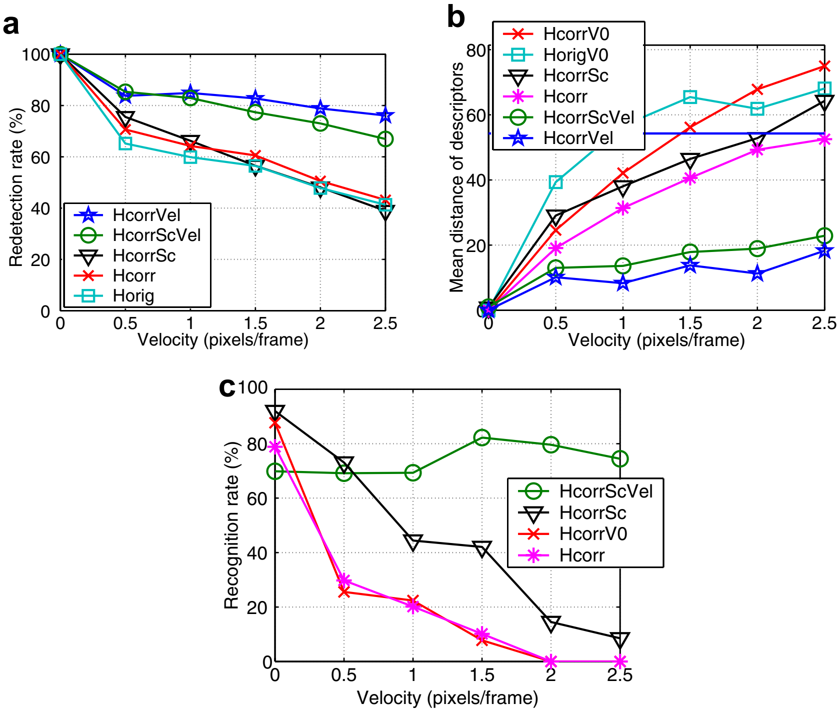


Fig. 12. Evaluation of local motion events under Galilean transformations. (a) Repeatability of motion events for different values of the velocity; (b) mean Euclidean distance between 4Jet-descriptors of corresponding events in the original and in the velocity-warped sequences. The horizontal line in the graph corresponds to the mean distance between all descriptor pairs in the sequences; (c) average performance of action recognition subject to velocity transformations in the data.

better than for descriptors without velocity adaptation (*HorigV0*, *HcorrV0*), however, it is outperformed significantly by the methods involving iterative velocity adaptation. Hence, a more accurate estimation of true Galilean transformations using the iterative velocity adaptation appears to be crucial for obtaining stability under velocity transformations.

6.1.3. Recognition performance

Besides the stability of image descriptors and the repeatability of event detection, reliable matching and motion recognition also requires the motion events to be discriminative. Here, we evaluate the discriminative power of the velocity-adapted events and the stability of the recognition performance under Galilean transformations. We consider an action in a test sequence as correctly recognized if it corresponds to the action of a person in the most similar training sequence. The similarities between sequences are computed using greedy matching in combination with the Euclidean distance metric and 4Jet-descriptors. The sequences for the test and the training sets here correspond to a subset of outdoor scenes $s1$ illustrated in Fig. 11. Different subjects were used in the training and in the test sets while the recognition performance was averaged over 100 random permutations with respect to the subjects.

Fig. 12(c) illustrates the results of motion recognition for different velocity transformations and for different types of adaptation of motion events. As can be seen, the stable curve under different velocity transformations corresponds to the iterative velocity adaptation of motion events and descriptors (*HcorrScVel*). However, the best recognition performance is achieved for the velocity value $v_x = 0$ for methods without iterative velocity adaptation. An explanation for the maximum at $v_x = 0$ is that both the training sequences and the original test sequences were recorded with a stationary camera. Hence, the velocities of the people in the test and training sequences are similar. Moreover, the relatively low recognition rate of *HcorrScVel* at $v_x = 0$ can be explained by the *loss of discriminative power* associated with the velocity adaptation. Velocity is indeed an important cue when discriminating between, for example, a walking and a running person. Since velocity adaptation cancels this information from the local descriptors, it is not surprising that *HcorrScVel* performs slightly worse than the other methods when the velocity in the training and in the test sets coincide. Hence, the stability with respect velocity transformations is here achieved at the cost of a slight decrease in the recognition performance. This property will become even more evident in the next section.

6.2. Evaluation of local motion descriptors

In this section, we evaluate and compare the motion descriptors introduced in Section 4. For this purpose we perform motion recognition experiments using 192 sequences with six types of actions from the database in Fig. 11 performed by eight different individuals. To assess the generalization performance of the method we present recognition results for different number of randomly selected training subjects. We used a NN classifier and three dissimilarity measures according to Section 5. Since the recognition performance was found to be dependent on velocity adaptation (Section 6.1.3), we performed separate experiments using either scale-adapted events or events detected with iteratively adapted scales and velocities.

The results of experiments using different types of local motion descriptors as well as different dissimilarity measures are shown in Fig. 13. Due to the large number of tested descriptors, we show only one descriptor within each descriptor class that maximizes the recognition performance within the class. We observe that the recognition rates are rather high for most of descriptor classes while the highest value 96.5% is obtained for OF-PDHIST descriptor in combination with the Euclidean distance measure. Independently of the dissimilarity measure and the type of local measurements (STG or OF), the position-dependent histograms result in the best performance when using scale-adapted events (see Fig. 13(left)). This result coincides with a similar result in the spatial domain, where the conceptually similar *SIFT*-descriptor [36] was found to outperform other local image descriptors when matching local events in static images [42].

Concerning the other descriptors, we observe that optic flow (OF) in combination with PCA does not perform well in most of these experiments. Moreover, the OF descriptors are consistently worse than the STG descriptors in the experiments using velocity-adapted events (see Fig. 13(right)). A reasonable explanation for the last observation is the reduced discriminative power of the OF descriptors due to the velocity adaptation. We note a rather stable recognition performance for all the methods depending on the number of training subjects.

The recognition performance for the events detected without iterative velocity adaptation (see Fig. 13(left)) is somewhat better than for the events with iterative velocity adaptation (see Fig. 13(right)). Similar to Section 6.1.3 this

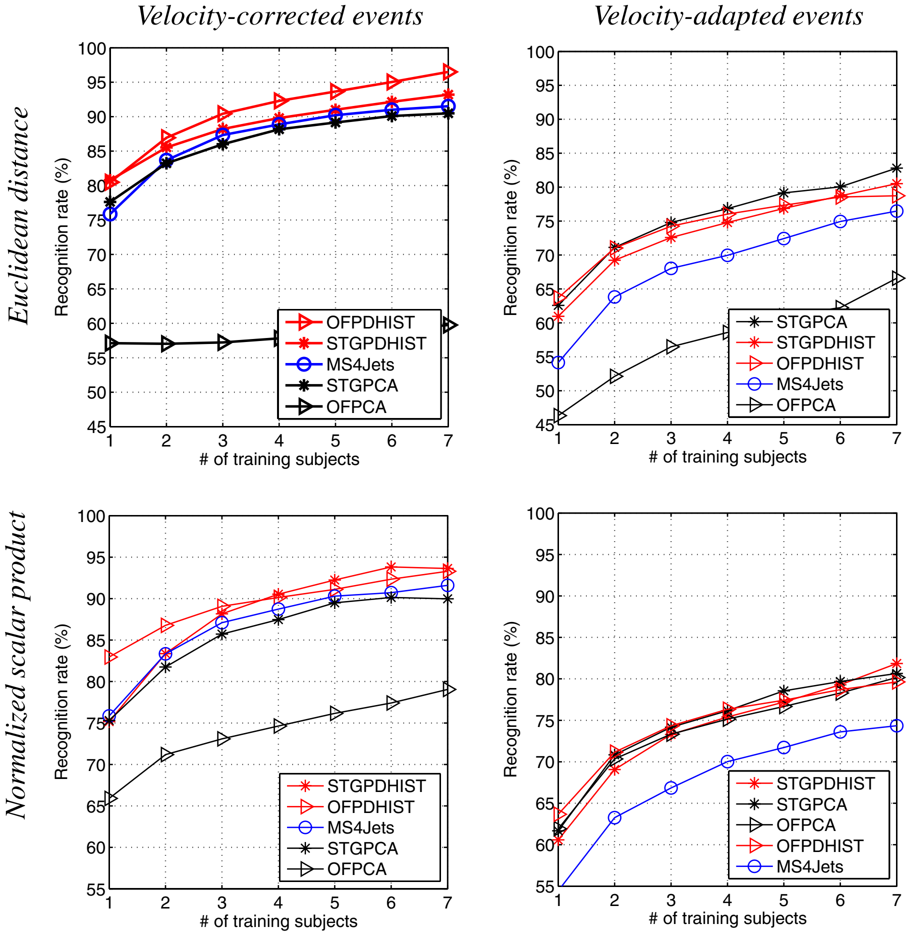


Fig. 13. Results of recognizing human actions in a subset of $s1$ sequences of the database in Fig. 11. The recognition performance is reported for different number of training subjects when using either (top) the Euclidean distance; or (bottom) the normalized scalar product for event comparison. (Left column) Recognition rates obtained for scale-adapted events with complementary velocity correction; (right column) recognition rates obtained for scale- and velocity-adapted events. All recognition results are averaged over 500 random perturbations of the dataset with respect to the subject. The results are shown only for the descriptor maximizing the recognition performance within the descriptor class (e.g. MS4Jets is chosen among MS4Jets, MS2Jets, 4Jets and 2Jets).

can be explained by the fact that the camera was stationary in both the training and the test sequences. Although velocity adaptation appears to be unnecessary in this case, the advantage of velocity adaptation for motion recognition will be emphasized in Section 6.4 when performing experiments on image sequences with different amount of camera motion.

6.2.1. Comparison to other methods

In this section, we compare the performance of local motion descriptors to the performance of other related representations of image sequences evaluated on the same dataset. At first, we consider a method in terms of *spatial* local features detected as maxima of Harris operator [16] for every fourth frame in the image sequence. The obtained features are adapted with respect to the spatial scale using the approach in [32,40] and spatial N-Jet descriptors are computed for each feature at the adapted scale. The resulting features and the corresponding descriptors are then used for action recognition in a similar way as local motion events. Such a method is very similar to ours, except that it does not use any temporal information neither for the event detection nor for the computation of the local descriptors. The main motivation for comparing with this approach was to confirm that the temporal information captured by motion events is *essential* for the recognition and that the problem of action recognition in our sequences is non-trivial from the view point of spatial recognition. From the results obtained for this method (Spatial-4Jets) presented in Fig. 14, we confirm that the performance of the local spatial features is here close to chance and that the use of temporal information is essential for this data set.

Two other methods used for comparison are based on *global* histograms of spatio-temporal gradients computed for the whole sequence at points with significant temporal variations of intensity. Such points are estimated by thresholding the first-order temporal partial derivative computed for all points in the sequences (a number of different thresholds were tested and only the best obtained results are reported here). Separable histograms were computed for:

Global-STG-HIST-MS: Normalized components of spatio-temporal gradients $L_x/\|\nabla L\|$, $L_y/\|\nabla L\|$, $L_t/\|\nabla L\|$ at multiple spatial and temporal scales

Global-STG-HIST-ZI: Absolute values of components in Global-STG-HIST-MS $|L_x|/\|\nabla L\|$, $|L_y|/\|\nabla L\|$, $|L_t|/\|\nabla L\|$ at multiple temporal scales only

The representation Global-STG-HIST-ZI was used previously for the task of action recognition in [62]. Global-STG-HIST-MS is an extension of [62] where we additionally take the direction of the spatio-temporal gradients at multiple spatial scales into account. To recognize actions using these two types of global representations, we computed histograms for all the sequences in the dataset and used nearest neighbor classification and dissimilarity measures according to Section 5. The results for both of these methods optimized over three dissimilarity measures are shown in Fig. 14. As can be seen, both methods perform rather well with the better performance for Global-STG-HIST-MS. A representation in terms of local motion events (OF-PDHIST) results in the best performance for the methods compared here.

6.3. Evaluation of action recognition

In this section, we evaluate the performance of action recognition on the full database illustrated in Fig. 11 using both NN and SVM classifiers. To train the SVM classifier all image sequences were divided with respect to the subjects into a training set (8 persons), a validation set (8 persons) and a test set (9 persons). The validation set was used for parameter optimization while all the results are reported for the test set.

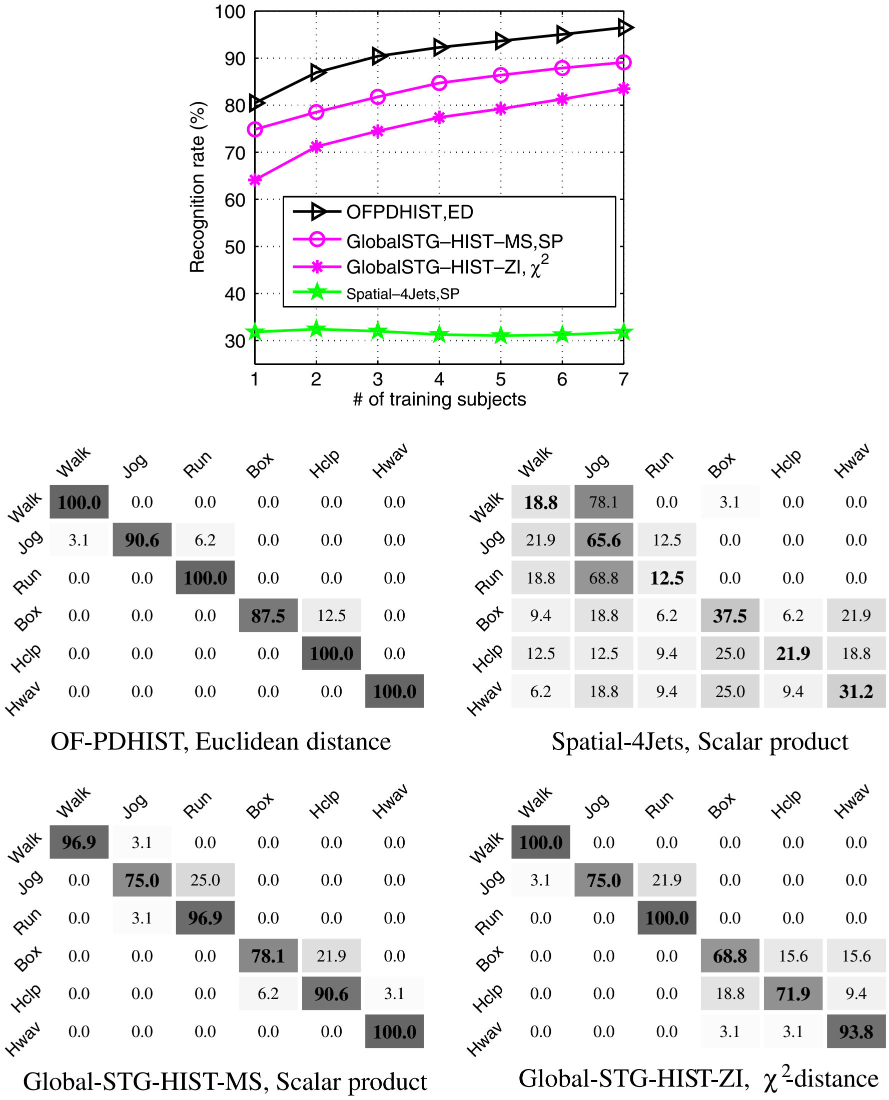
6.3.1. Methods

We compare the results obtained with the combination of three different representations and two classifiers. The representations according to the definitions in Sections 4.3 and 5.1 are: (i) LME (17) with scale- and velocity-adapted OF-PDHIST local motion descriptors; (ii) 128-bin LMEHist histograms (18) defined on 4Jets local motion descriptors and (iii) Global-STG-HIST. For the classification we use: (i) SVM with either local feature kernel [58] in combination with LME or SVM with χ^2 kernel for classifying histogram-based representations LMEHist and Global-STG-HIST (see Section 5.1.2) and (ii) nearest neighbor classification in combination with MLE, MLEHist and Global-STG-HIST according to Section 5.1.1.

6.3.2. Results

Fig. 15(top) shows recognition rates for all of the methods. To analyze the influence of different scenarios we performed training on different subsets of $\{s1\}$, $\{s1, s4\}$, $\{s1, s3, s4\}$ and $\{s1, s2, s3, s4\}$ (see Fig. 11 for the definitions of the subsets). It follows that LME with local SVM gives the best performance for all training sets while the performance of all methods increases with the number of scenarios used for training. Regarding the histogram-based representations, SVM outperforms NN as expected, while LMEHist gives a slightly better performance than Global-STG-Hist.

Fig. 15(bottom) shows confusion matrices obtained with the LME + SVM method. As can be seen, there is a clear separation between the leg actions and the arm actions. Most of the confusion occurs between jogging and running sequences as well as between the boxing and hand clapping sequences. We observed a similar structure on the confusion matrices for the other methods as well. The scenario with scale variations ($s2$) is the most difficult one for all



	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**100.0**	0.0	0.0	0.0	0.0	0.0
Jog	3.1	**90.6**	6.2	0.0	0.0	0.0
Run	0.0	0.0	**100.0**	0.0	0.0	0.0
Box	0.0	0.0	0.0	**87.5**	12.5	0.0
Hclp	0.0	0.0	0.0	0.0	**100.0**	0.0
Hwav	0.0	0.0	0.0	0.0	0.0	**100.0**

OF-PDHIST, Euclidean distance

	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**18.8**	78.1	0.0	3.1	0.0	0.0
Jog	21.9	**65.6**	12.5	0.0	0.0	0.0
Run	18.8	68.8	**12.5**	0.0	0.0	0.0
Box	9.4	18.8	6.2	**37.5**	6.2	21.9
Hclp	12.5	12.5	9.4	25.0	**21.9**	18.8
Hwav	6.2	18.8	9.4	25.0	9.4	**31.2**

Spatial-4Jets, Scalar product

	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**96.9**	3.1	0.0	0.0	0.0	0.0
Jog	0.0	**75.0**	25.0	0.0	0.0	0.0
Run	0.0	3.1	**96.9**	0.0	0.0	0.0
Box	0.0	0.0	0.0	**78.1**	21.9	0.0
Hclp	0.0	0.0	0.0	6.2	**90.6**	3.1
Hwav	0.0	0.0	0.0	0.0	0.0	**100.0**

Global-STG-HIST-MS, Scalar product

	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**100.0**	0.0	0.0	0.0	0.0	0.0
Jog	3.1	**75.0**	21.9	0.0	0.0	0.0
Run	0.0	0.0	**100.0**	0.0	0.0	0.0
Box	0.0	0.0	0.0	**68.8**	15.6	15.6
Hclp	0.0	0.0	0.0	18.8	**71.9**	9.4
Hwav	0.0	0.0	0.0	3.1	3.1	**93.8**

Global-STG-HIST-ZI, χ^2-distance

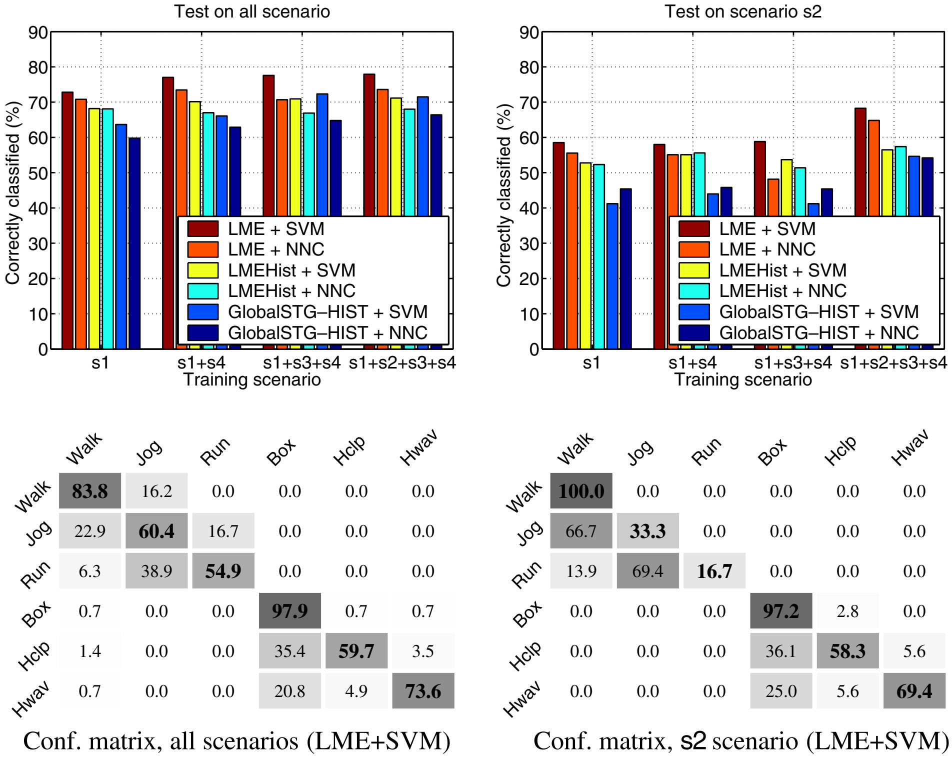
Fig. 14. Comparison of the recognition performance using local space–time events (OF-PDHIST) and other methods in terms of (Spatial-4Jets) spatial interest points with fourth-order spatial Jets descriptors; (Global-STG-HIST-MS) global histograms of normalized spatio-temporal gradients computed at multiple spatial and temporal scales; (Global-STG-HIST-ZI) global histograms of normalized spatio-temporal gradients according to [62]. Results are shown as plots for different number of training subjects and as confusion matrices for experiments with seven subjects in the training set.

the methods. The low spatial resolution of image sequences is likely to be one among other factors disturbing the reliable interpretation in $s2$. The recognition rates and the confusion matrix when testing on $s2$ only are shown in Fig. 15(right).

When analyzing the confusion matrices in Fig. 15(bottom), the confusion between walking and jogging as well as between jogging and running can partly be explained by the high similarities between these classes (running of some people may appear very similar to the jogging of the others). As can be seen from Fig. 15(top, right), the performance of the local motion events (LME) is significantly better than the performance of Global-STG-HIST for all the training subsets. This indicates the stability of recognition with respect to scale variations in image sequences when using scale-adapted local features for action representation. Further, experiments indicating the stability of LME representation under scale changes can be found in [24].

6.4. Action recognition in complex scenes

In this section, we apply motion events to action recognition in complex scenes. We use a test set with 51 image sequences of human actions that have been recorded in city environments (see Fig. 16). The type of recorded actions was the same as in the database of Fig. 11 except for jogging. Most of the sequences contain heterogeneous background as well as background motion caused by for example moving cars. Moreover, about the half of all the sequences were taken with a stationary camera, while the other half with a moving camera that was manually stabilized on the subject. Other variations in these sequences include changes in the spatial scale (sequences 1–3, 17, 22–27, 37), occlusions



	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**83.8**	16.2	0.0	0.0	0.0	0.0
Jog	22.9	**60.4**	16.7	0.0	0.0	0.0
Run	6.3	38.9	**54.9**	0.0	0.0	0.0
Box	0.7	0.0	0.0	**97.9**	0.7	0.7
Hclp	1.4	0.0	0.0	35.4	**59.7**	3.5
Hwav	0.7	0.0	0.0	20.8	4.9	**73.6**

Conf. matrix, all scenarios (LME+SVM)

	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**100.0**	0.0	0.0	0.0	0.0	0.0
Jog	66.7	**33.3**	0.0	0.0	0.0	0.0
Run	13.9	69.4	**16.7**	0.0	0.0	0.0
Box	0.0	0.0	0.0	**97.2**	2.8	0.0
Hclp	0.0	0.0	0.0	36.1	**58.3**	5.6
Hwav	0.0	0.0	0.0	25.0	5.6	**69.4**

Conf. matrix, s2 scenario (LME+SVM)

Fig. 15. Results of action recognition for different methods and scenarios. (Top, left) Recognition rates for test sequences in all scenarios; (top, right) recognition rates for test sequences in the *s*2 scenario; (bottom, left) confusion matrix for LME + SVM for test sequences in all scenarios; (bottom, left) confusion matrix for LME + SVM for test sequences in the *s*2 scenario.

(sequences 5, 35, 13, 36, 38) and three-dimensional view variations (sequences 17, 20, 22–24).

For the training, we used 192 sequences of human actions with simple background from the database in Fig. 11. Since the training and the test sequences were recorded with different camera motions, we detected local motion events using iterative adaptation with respect to scale and velocity according to Section 3.3. For each event, we then computed scale- and velocity-adapted local image descriptors according to Section 4. To recognize image sequences we used NN and SVM classifiers in combination with LME event-based representations according to Section 5.1. The recognition rate was then computed as a ratio between the number of correctly classified actions and the number of all sequences in the test set. For the NN classifier, the recognition rate was separately computed for all (valid) combinations of local motion descriptors and the three dissimilarity measures in (14)–(16). Due to computational complexity, the SVM method was only evaluated for the type of descriptor with the best performance of the NN classifier.

Motion events are frequently triggered by the background motion in complex scenes. This behaviour is illustrated on one of our test sequences in Fig. 17(a) where a large number of detected events is caused by the visual interaction of cars and a person. In our recognition framework outlier rejection is made implicitly by enforcing consistent matches of events in the training and the test sequences. When matching events in Fig. 17(a) to the training set with human actions, most of the background events are discarded as illustrated in Fig. 17(b).

The recognition rates for the different types of local motion descriptors are presented in Fig. 18 where for each descriptor the result is optimized over different dissimilarity measures (14)–(16). As can be seen, the highest recognition rate is obtained for the STG-PCA and the STG-PDHIST descriptors. We can note that the same type of descriptors (in the same order) gave the best performance when evaluated on action recognition in the simple scenes using motion events detected with iterative velocity adaptation (see Fig. 13(right)). Given a large number of all tested descriptors, the consistency of these results indicates good generalization of the method for scenes with complex backgrounds.

Confusion matrices for the two best descriptors and the NN classifier are illustrated in Fig. 18(bottom). As can be seen, the performance of STG-PCA is almost perfect for all actions except "running" which is recognized as "jogging" in most of the cases. This confusion can be explained by somewhat diffuse definition of the boundary between these two classes of actions. If "running" and "jogging" actions are merged into one class, the performance of STG-PCA increases to 96%.

We can note that the 2Jets-descriptor with the forth best performance is also the most simple one among all the other alternatives and contains only 9 components. This indicates that the information in the other types of descrip-

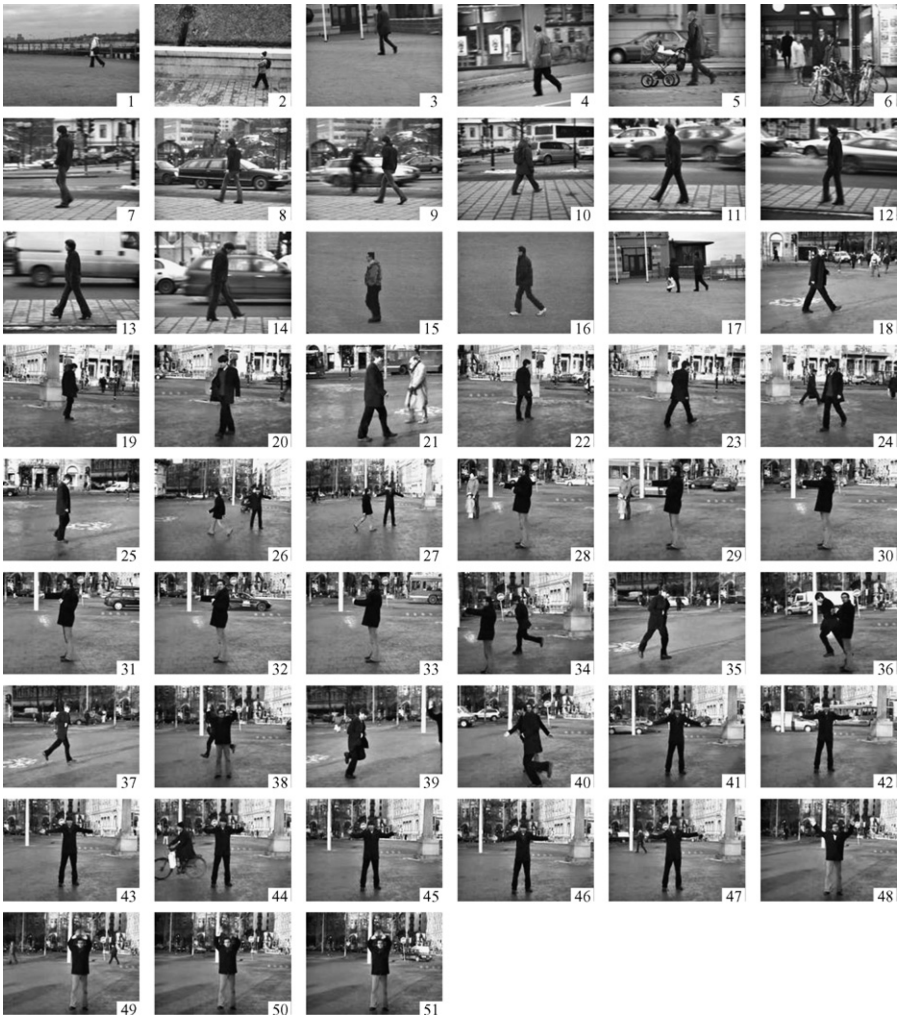


Fig. 16. Image frames from 51 sequences with human actions performed in complex scenes. (1–27): Walking; (28–33): boxing; (34–40): running; (41–47): hand clapping; (48–51): hand waving. The scenes contain variations in terms of heterogeneous, non-static backgrounds, variations in the spatial scale, variations in the camera motions as well as occlusions.

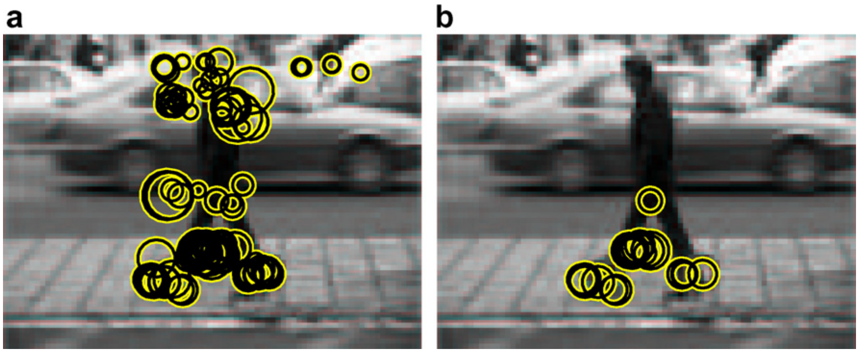
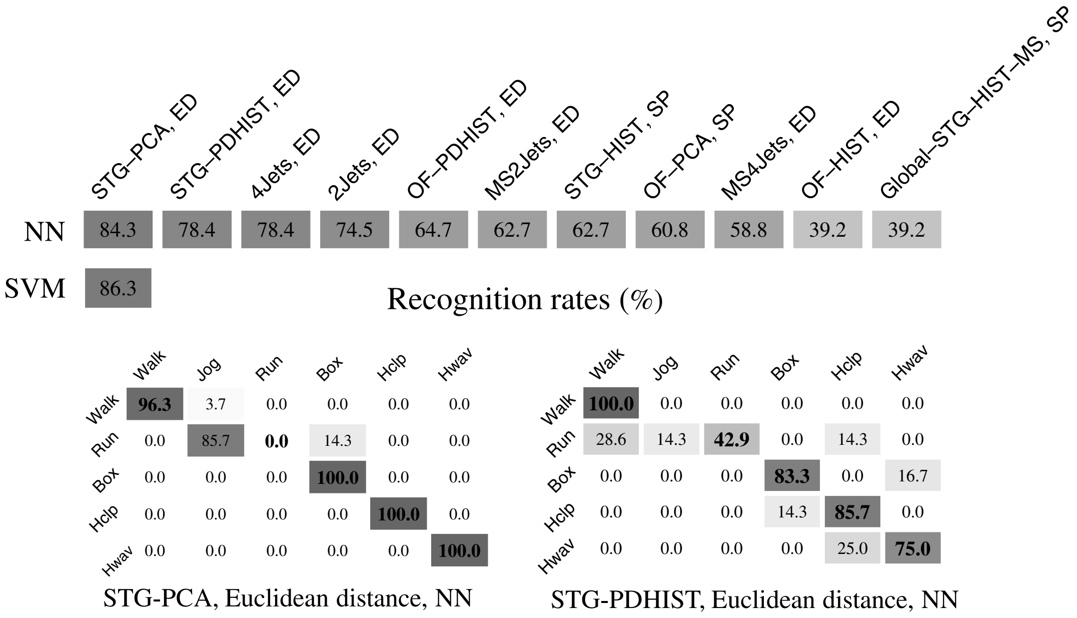


Fig. 17. Illustration of motion events for a walking action with complex background. (a) Time-projection of all detected features onto one frame of a sequence; (b) a subset of features in (a) that do match with events in a similar training sequence.

tors might be highly redundant. Among the histogram-based descriptors, we can note that the position-dependent histograms perform significantly better than position-independent histograms, which is consistent with the results in Section 6.2. When comparing the local measurements, we note that descriptors based on spatio-temporal gradients perform better than descriptors based on optic flow in most of the cases.

Finally, we also compare the performance of the local methods to the performance of the two global methods in



Recognition rates (%)

	STG–PCA, ED	STG–PDHIST, ED	4Jets, ED	2Jets, ED	OF–PDHIST, ED	MS2Jets, ED	STG–HIST, SP	OF–PCA, SP	MS4Jets, ED	OF–HIST, ED	Global–STG–HIST–MS, SP
NN	84.3	78.4	78.4	74.5	64.7	62.7	62.7	60.8	58.8	39.2	39.2
SVM	86.3										

	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**96.3**	3.7	0.0	0.0	0.0	0.0
Run	0.0	85.7	**0.0**	14.3	0.0	0.0
Box	0.0	0.0	0.0	**100.0**	0.0	0.0
Hclp	0.0	0.0	0.0	0.0	**100.0**	0.0
Hwav	0.0	0.0	0.0	0.0	0.0	**100.0**

STG-PCA, Euclidean distance, NN

	Walk	Jog	Run	Box	Hclp	Hwav
Walk	**100.0**	0.0	0.0	0.0	0.0	0.0
Run	28.6	14.3	**42.9**	0.0	14.3	0.0
Box	0.0	0.0	0.0	**83.3**	0.0	16.7
Hclp	0.0	0.0	0.0	14.3	**85.7**	0.0
Hwav	0.0	0.0	0.0	0.0	25.0	**75.0**

STG-PDHIST, Euclidean distance, NN

Fig. 18. Results of recognizing human actions in complex scenes for the sequences in Fig. 16. (Top) Recognition rates for different types of descriptors and classification methods maximized over dissimilarity measures: the Euclidean distance (ED), the normalized scalar product-based measure (SP) and the χ^2 measure as defined in Section 5; (bottom) confusion matrices for two best methods.

terms of histograms of spatio-temporal gradients as described in Section 6.2.1. From Fig. 18, we see that independently of the type of local descriptors, the performance of all tested local methods is better (or equal for OF-HIST) than the performance of the global descriptors. The low performance of the global histograms with the best performance for Global-STG-HIST-MS (39.2%) is not surprising, since such descriptors depend on the motion of the camera, scale variations and the motion in the background. Thus, the results in this experiment confirm the expected advantages of event-based local motion representations in terms of (i) stability to scale and velocity transformations due to the adaptation procedure in Section 3.3 as well as (ii) stability to multiple motions in the scene due to the local nature of the motion descriptors and the matching procedure.

7. Summary and discussion

This paper explored the notion of local motion events for motion recognition. The original motivation for the method was to overcome difficulties associated with motion recognition in complex scenes. Towards this goal, the experiments in Section 6.4 confirmed the expected advantage of event-based motion representations by demonstrating promising results for the task of recognizing human actions in complex scenes.

To obtain invariance with respect to relative camera motion we proposed to adapt motion events to Galilean transformations estimated from the data. This method has been shown to be essential for motion recognition in scenes where methods for motion segmentation and/or camera stabilization may not be reliable. Local velocity adaptation, however, has been achieved at the price of reduced discriminative power of the motion descriptors. Hence, if the relative motion of the camera is known in advance (e.g. for a fixed surveillance camera), a higher discriminative power of the motion descriptors could be obtained if the velocity adaptation stage is discarded. If these specific assumptions are violated, however, we argue that it may be a clear advantage to include an explicit mechanism for local velocity adaptation as we do in this work.

When comparing different types of local motion descriptors, we found the position-dependent histograms to provide the best recognition performance. This result is consistent with the findings in the spatial domain where the related histogram-based SIFT descriptor [36] has been demonstrated to give the best performance in [42]. We have also shown how motion representations in terms of local motion events can be combined with a SVM classifier for an additional increase in recognition performance.

There are several natural extensions of this work. Currently we use a simplifying assumption of a single motion class per image sequence. This assumption, however, is not imposed by the local motion events per se and could be relaxed if re-formulating the recognition methods in Section 5 accordingly.

Another issue concerns relative structure of events in space–time. Whereas here for simplicity reasons all the events have been treated independently, there exists a strong dependency among events imposed by the temporal and the spatial structure of motion patterns. Using this dependency as an additional constraint is expected to increase the recognition performance. A similar idea along with the detection of multiple motions in a scene has been recently addressed in [25] for the special case of periodic motion.

Another domain of possible extensions of this work concerns the observation that the spatio-temporal interest point operator in this work returns information rich and well localized but rather sparse responses in space–time. This behaviour appears to be useful in some applications such as the one presented in this paper or applications of space–time sequence alignment as discussed in [25]. For other applications the sparse nature of motion events in this paper may be too restrictive. Concerning alternative classes of spatio-temporal interest operators, a few complementary methods have been recently proposed and used in [34,10,44]. Investigating other types of space–time interest operators such as the spatio-temporal Laplacian or quadrature filter pairs may be a fruitful direction for future research.

Acknowledgments

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.

Appendix A. Galilean-invariant block-diagonal form

In this section, we prove Proposition 3.3 in Section 3.2. For this purpose we first consider

Lemma A.1. *For each non-degenerative second moment matrix there exists a unique Galilean-related second moment matrix with the block-diagonal form (5).*

Lemma A.1 follows directly from the uniqness of the solution for $G(v_x, v_y)$ (7) and (8) that brings a second-moment matrix into a block-diagonal form according to (6).

Now let non-degenerative second-moment matrices μ_a, μ_b, μ_c be Galilean-related as $\mu_b \xrightarrow{G_{ba}} \mu_a$ by G_{ba} and $\mu_c \xrightarrow{G_{cb}} \mu_b$ by G_{cb}. Let also μ_c be of the block-diagonal form (5). Since the Galilean relation of second-moment matrices is transitive, it follows that $\mu_c \xrightarrow{G_{ca}} \mu_a$ by $G_{ca} = G_{cb}G_{ba}$. By Lemma A.1 we have that μ_c is a unique block-diagonal matrix that is Galilean-related to μ_a. Since μ_c is also Galilean-related to μ_b, we have that two arbitrary Galilean-related second moment matrices have a unique block-diagonal form. This proves Proposition 3.3.

Appendix B. Implementation details of motion descriptors

This section provides technical details for the event descriptors in Section 4.

For a motion event defined by position x, y, t scale values σ, τ and velocity values v_x, v_y, all histograms were computed at 9 combinations of 3 spatial scales $\sigma/2$, σ, 2σ and 3 temporal scales $\tau/2$, τ, 2τ. The global histograms (descriptor Global-STG-HIST) were computed for all combinations of the spatial scales $\sigma \in \{1, 2, 4\}$ and the temporal scales $\tau \in \{1, 2, 4\}$. When accumulating marginalized histograms of spatio-temporal gradients, only image locations with L_t above a threshold (chosen manually by optimizing the recognition performance on the validation set) were allowed to contribute. Moreover, all the marginal histograms were smoothed with binomial filters and were normalized to unit l_1-norm. For the position dependent histograms (descriptors OF-PDHIST and STG-PDHIST), we discretize each of the spatial and the time coordinates in two bins ($M = 2$) and evaluate the position dependent entities using Gaussian weighted window functions centered at $(x \pm \alpha\sigma, y \pm \alpha\sigma, t \pm \beta\tau)$ with $\alpha = 1.5$ and $\beta = 1.5$. The spatial standard deviation of the Gaussian weighting function was 3σ and the temporal standard deviation 3τ. For the position dependent histograms, 16 bins were used for the components of the spatio-temporal gradients or the optic flow, while 32 bins were used for the position independent histograms. Thus, with $M = 2$ the position dependent histograms contain 9 scales × 8 positions × 3 derivatives × 16 bins = 3456 accumulator cells, and position independent histograms contain 9 scales × 3 derivatives × 32 bins = 864 cells. For the local principal component analysis, the gradient vectors and the optic flow were computed in windows of spatial extent $\pm 3\sigma$ and temporal extent $\pm 3\tau$ around the interest points. Prior to the computation of the principal components using $D = 100$ dimensions, the gradient vectors and the optic flow were re-sampled to a $9 \times 9 \times 9$ grid using trilinear interpolation.

References

[1] S. Belongie, C. Fowlkes, F. Chung, J. Malik, Spectral partitioning with indefinite kernels using the nyström extension, in: Proc. Seventh Eur. Conf. on Computer Vision, Copenhagen, DenmarkLecture Notes in Computer Science, vol. 2352, Springer Verlag, Berlin, 2002, p. III:531 ff.

[2] M.J. Black, A.D. Jepson, Eigentracking: Robust matching and tracking of articulated objects using view-based representation, Int. J. Comput. Vis. 26 (1) (1998) 63–84.

[3] M.J. Black, Y. Yacoob, S.X. Ju, Recognizing human motion using parameterized models of optical flow, in: M. Shah, R. Jain (Eds.), Motion-Based Recognition, Kluwer Academic Publishers, Dordrecht, Boston, London, 1997, pp. 245–269.

[4] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space–time shapes, in: Proc. 10th Int. Conf. on Computer Vision, Beijing, China, 2005, pp. II:1395–II:1402.

[5] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001) 257–267.

[6] O. Boiman, M. Irani, Detecting irregularities in images and in video. in: Proc. 10th Int. Conf. on Computer Vision, Beijing, China, 2005, pp. I:462–I:469.

[7] O. Chapelle, P. Haffner, V. Vapnik, SVMs for histogram-based image classification, IEEE Trans. Neural Network 10 (5) (1999).

[8] O. Chomat, J. Martin, J.L. Crowley, A probabilistic sensor for the perception and recognition of activities, in: Proc. Sixth Eur. Conf. on Computer Vision, Dublin, IrelandLecture Notes in Computer Science, vol. 1842, Springer Verlag, Berlin, 2000, pp. I:487–I:503.

[9] N. Cristianini, J.S. Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.

[10] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: VS-PETS, 2005, pp. 65–72.

[11] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: Proc. Ninth Int. Conf. on Computer Vision, Nice, France, 2003, pp. 726–733.

[12] R. Fablet, P. Bouthemy, Motion recognition using nonparametric image motion models estimated from temporal and multiscale co-occurrence statistics, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1619–1624.

[13] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in: Proc. Computer Vision and Pattern Recognition, Madison, Wisconsin, 2003, pp. II:264–II:271.

[14] D.M. Gavrila, The visual analysis of human movement: a survey, Comput. Vis. Image Und. 73 (1) (1999) 82–98.

[15] J.M. Gryn, R.P. Wildes, J.K. Tsotsos, Detecting motion patterns via direction maps with application to surveillance, in: WACV/MOTION, 2005, pp. 202–209.

[16] C. Harris, M.J. Stephens, A combined corner and edge detector, in: Alvey Vision Conference, 1988, pp. 147–152.

[17] J. Hoey, J.J. Little, Representation and recognition of complex human motion, in: Proc. Computer Vision and Pattern Recognition, Hilton Head, SC, 2000, pp. I:752–I:759.

[18] B. Jähne, H. Haußecker, P. Geißler, Signal processing and pattern recognition, Handbook of Computer Vision and Applications, vol. 2, Academic Press, 1999, chapter 13.

[19] T. Kadir, M. Brady, Saliency, scale and image description, Int. J. Comput. Vis. 45 (2) (2001) 83–105.

[20] Y. Ke, R. Sukthankar, PCA-SIFT: a more disctinctive representation for local image descriptors, Technical Report IRP–TR–03–15, Intel, November 2003.

[21] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, in: Proc. 10th Int. Conf. on Computer Vision, Beijing, China, 2005, pp. I:166–I:173.

[22] J.J. Koenderink, A.J. van Doorn, Generic neighborhood operators, IEEE Trans. Pattern Anal. Mach. Intell. 14 (6) (1992) 597–605.

[23] J.J. Koenderink, A.J. van Doorn, Representation of local geometry in the visual system, Biol. Cybern. 55 (1987) 367–375.

[24] I. Laptev, Local Spatio-Temporal Image Features for Motion Interpretation. Ph.D. thesis, Department of Numerical Analysis and Computer Science (NADA), KTH, S-100 44 Stockholm, Sweden, 2004. ISBN 91-7283-793-4.

[25] I. Laptev, S. Belongie, P. Pérez, J. Wills. Periodic motion detection and segmentation via approximate sequence alignment, in: Proc. 10th Int. Conf. on Computer Vision, Beijing, China, 2005, pp. I:816–I:823.

[26] I. Laptev, T. Lindeberg, Space–time interest points, in: Proc. Ninth Int. Conf. on Computer Vision, Nice, France, 2003, pp. 432–439.

[27] I. Laptev, T. Lindeberg, Local descriptors for spatio-temporal recognition, in: First Int. Workshop on Spatial Coherence for Visual Motion Analysis, vol. 3667 of Lecture Notes in Computer Science, Springer Verlag, Berlin, 2004, pp. 91–103.

[28] I. Laptev, T. Lindeberg, Velocity adaptation of space–time interest points, in: Proc. Int. Conf. on Pattern Recognition, Cambridge, UK, 2004, pp. I:52–I:56.

[29] I. Laptev, T. Lindeberg, Velocity-adaptation of spatio-temporal receptive fields for direct recognition of activities: an experimental study, Image Vis. Comput. 22 (2) (2004) 105–116.

[30] B. Leibe, B. Schiele, Interleaved object categorization and segmentation, in: Proc. British Machine Vision Conference, Norwich, GB, 2003.

[31] O. Linde, T. Lindeberg, Object recognition using composed receptive field histograms of higher dimensionality, in: Proc. Int. Conf. on Pattern Recognition, Cambridge, UK, 2004, pp. II:1–II:6 .

[32] T. Lindeberg, Feature detection with automatic scale selection, Int. J. Comput. Vis. 30 (2) (1998) 77–116.

[33] T. Lindeberg, Time-recursive velocity-adapted spatio-temporal scale-space filters, in: Proc. Seventh Eur. Conf. on Computer Vision, Copenhagen, DenmarkLecture Notes in Computer Science, vol. 2350, Springer Verlag, Berlin, 2002, pp. I:52–I:67.

[34] T. Lindeberg, A. Akbarzadeh, I. Laptev, Galilean-corrected spatio-temporal interest operators, in: Proc. 17th Int. Conf. on Pattern Recognition, Cambridge, UK, 2004, pp. I:57–I:62.

[35] T. Lindeberg, J. Gårding, Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure, Image Vis. Comput. 15 (6) (1997) 415–434.

[36] D.G. Lowe, Object recognition from local scale-invariant features, in: Proc. Seventh Int. Conf. on Computer Vision, Corfu, Greece, 1999, pp. 1150–1157.

[37] D.G. Lowe, Local feature view clustering for 3d object recognition, in: Proc. Computer Vision and Pattern Recognition, Kauai Marriott, Hawaii, 2001, pp. I:682–I:688.

[38] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: DARPA Image Understanding Workshop, 1981, pp. 121–130.

[39] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: Proc. British Machine Vision Conference, 2002, pp. 384–393.

[40] K. Mikolajczyk, C. Schmid, Indexing based on scale invariant interest points, in: Proc. Eighth Int. Conf. on Computer Vision, Vancouver, Canada, 2001, pp. I:525–I:531.

[41] K. Mikolajczyk, C. Schmid, An affine invariant interest point detector, in: Proc. Seventh Eur. Conf. on Computer Vision, Copenhagen, DenmarkLecture Notes in Computer Science, vol. 2350, Springer Verlag, Berlin, 2002, pp. I:128–I:142.

[42] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, in: Proc. Computer Vision and Pattern Recognition, 2003, pp. II:257–II:263.

[43] H.H. Nagel, A. Gehrke, Spatiotemporal adaptive filtering for estimation and segmentation of optical flow fields, in: H. Burkhardt, B. Neumann (Eds.), Proc. Fifth Eur. Conf. on Computer Vision, Freiburg, Germany, Lecture Notes in Computer Science, vol. 1407, Springer Verlag, Berlin, 1998, pp. II:86–II:102.

[44] J.C. Niebles, H. Wang, F.F. Li, Unsupervised learning of human action categories using spatial-temporal words, in: Proc. British Machine Vision Conference, 2006.

[45] M.E. Nilsback, B. Caputo, Cue integration through discriminative accumulation, in: Proc. Computer Vision and Pattern Recognition, 2004, pp. II:578–II:585.

[46] R. Polana, R.C. Nelson, Recognition of motion from temporal texture, in: Proc. Computer Vision and Pattern Recognition, 1992, pp. 129–134.

[47] C. Rao, A. Yilmaz, M. Shah, View-invariant representation and recognition of actions, Int. J. Comput. Vis. 50 (2) (2002) 203–226.

[48] Y. Rui, P. Anandan, Segmenting visual actions based on spatio-temporal motion patterns. in: Proc. Computer Vision and Pattern Recognition, vol. I, Hilton Head, SC, 2000, pp. 111–118.

[49] B. Schiele, J.L. Crowley, Recognition without correspondence using multidimensional receptive field histograms, Int. J. Comput. Vis. 36 (1) (2000) 31–50.

[50] C. Schmid, R. Mohr, Local grayvalue invariants for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 19 (5) (1997) 530–535.

[51] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proc. 17th Int. Conf. on Pattern Recognition, Cambridge, UK, 2004, pp. III:32–III:36.

[52] M. Shah, R. Jain (Eds.), Motion-Based Recognition, Kluwer Academic Publishers, Dordrecht, Boston, London, 1997.

[53] E. Shechtman, M. Irani, Space–time behavior based correlation, in: Proc. Computer Vision and Pattern Recognition, San Diego, CA, 2005, pp. I:405–I:412.

[54] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: Proc. Ninth Int. Conf. on Computer Vision, Nice, France, 2003, pp. 1470–1477.

[55] D. Tell, S. Carlsson, Combining topology and appearance for wide baseline matching, in: Proc. Seventh Eur. Conf. on Computer Vision, Copenhagen, DenmarkLecture Notes in Computer Science, vol. 2350, Springer Verlag, Berlin, 2002, pp. I:68–I:83.

[56] T. Tuytelaars, L.J. Van Gool, Wide baseline stereo matching based on local, affinely invariant regions, in: Proc. British Machine Vision Conference, 2000, pp. 412–425.

[57] V. Vapnik, Statistical Learning Theory, Wiley, NY, 1998.

[58] C. Wallraven, B. Caputo, A. Graf, Recognition with local features: the kernel recipe, in: Proc. Ninth Int. Conf. on Computer Vision, Nice, France, 2003, pp. 257–264.

[59] L. Wolf, A. Shashua, Kernel principal angles for classification machines with applications to image sequence interpretation, in: Proc. Computer Vision and Pattern Recognition, 2003, pp. I:635–I:640.

[60] Y. Yacoob, M.J. Black, Parameterized modeling and recognition of activities, Comput. Vis. Image Und. 73 (2) (1999) 232–247.

[61] A. Yilmaz, M. Shah, Recognizing human actions in videos acquired by uncalibrated moving cameras, in: Proc. 10th Int. Conf. on Computer Vision, Beijing, China, 2005, pp. I:150–I:157.

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