KTH Numerical Analysis
and Computer Science

Local Spatio-Temporal Image Features for Motion Interpretation

IVAN LAPTEV

Doctoral Thesis
Stockholm, Sweden 2004

Cover: “People in motion”
Video cut, Ivan Laptev, 2004

TRITA-NA-0413
ISSN 0348-2952
ISRN KTH/NA/R--4/13--SE
ISBN 91-7283-793-4
CVAP 289

KTH Numerisk analys och datalogi
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen fredagen den 11 juni 2004 i Kollegiesalen, Administrationsbyggnaden, Kungl Tekniska högskolan, Valhallavägen 79, Stockholm.

© Ivan Laptev, april 2004

Tryck: Universitetsservice US AB

Abstract

Visual motion carries information about the dynamics of a scene. Automatic interpretation of this information is important when designing computer systems for visual navigation, surveillance, human-computer interaction, browsing of video databases and other growing applications.

In this thesis, we address the issue of motion representation for the purpose of detecting and recognizing motion patterns in video sequences. We localize the motion in space and time and propose to use local spatio-temporal image features as primitives when representing and recognizing motions. To detect such features, we propose to maximize a measure of local variation of the image function over space and time and show that such a method detects meaningful events in image sequences. Due to its local nature, the proposed method avoids the influence of global variations in the scene and overcomes the need for spatial segmentation and tracking prior to motion recognition. These properties are shown to be highly useful when recognizing human actions in complex scenes.

Variations in scale and in relative motions of the camera may strongly influence the structure of image sequences and therefore the performance of recognition schemes. To address this problem, we develop a theory of local spatio-temporal adaptation and show that this approach provides invariance when analyzing image sequences under scaling and velocity transformations. To obtain discriminative representations of motion patterns, we also develop several types of motion descriptors and use them for classifying and matching local features in image sequences. An extensive evaluation of this approach is performed and results in the context of the problem of human action recognition are presented.

In summary, this thesis provides the following contributions: (i) it introduces the notion of local features in space-time and demonstrates the successful application of such features for motion interpretation; (ii) it presents a theory and an evaluation of methods for local adaptation with respect to scale and velocity transformations in image sequences and (iii) it presents and evaluates a set of local motion descriptors, which in combination with methods for feature detection and feature adaptation allow for robust recognition of human actions in complex scenes with cluttered and non-stationary backgrounds as well as camera motion.

Acknowledgments

This work has been done during the great time I spent in Computational Vision and Active Perception Laboratory (CVAP), KTH. First of all, I would like to thank my supervisor Tony Lindeberg for introducing me to the exciting field of vision and scale-space theory, for his support and stimulation, and for providing me with a firm source of knowledge whenever I needed it. My thanks go also to Jan-Olof Eklundh and to Henrik Christensen for providing a stimulating research environment, for their inspiring enthusiasm and personal support.

This work would not be the same without the support from all people at CVAP, thanks for the great atmosphere, open discussions, fun and friendship. In particular, thank you Johan for lots of stimulating talking about computer vision, and the visions about all other important aspects of life. Thank you Carsten and Gerit for your positive thinking and for the great time we spent together. Thank you Barbara for pushing me forward, – although you never read my papers, you knew, it will work! Thank you Peter and Lars for sharing thoughts on the art and vision. Thank you Josephine for the nice parties. Thanks to all people who red this thesis and contributed with very valuable comments. Thanks to the CVAP climbing team, Ola, Jonas, Johan and Calle for the great fun and for the safe climbing, "davaj naverh"!

I thank my parents and my sister Katja for the support and understanding. At last I thank Nastja for the patience, love, and endless support from the beginning to the end. Believing is power. You are my motivation.

Contents

Contents **vii**

1 Introduction **1**
1.1 Contributions of the thesis 3
1.2 Organization of the thesis 4
1.3 List of publications 5

I Background **7**

2 Related work on motion interpretation **9**
2.1 Structural methods 10
2.2 Appearance-based methods using motion templates 11
2.3 Statistical appearance-based methods 12
2.4 Event-based motion interpretation 14
2.4.1 Relevant research in psychology 15
2.4.2 The approach in this thesis 16

3 Computational theory **17**
3.1 Gaussian scale-space 19
3.1.1 Scale and orientation estimation 21
3.1.2 Local descriptors 24
3.1.3 Affine scale-space 24
3.1.4 Affine adaptation 25
3.2 Gaussian spatio-temporal scale-space 28
3.2.1 Spatio-temporal scale transformation 28
3.2.2 Galilean transformation 30
3.3 Time-recursive spatio-temporal scale-space 32
3.4 Motion estimation 34
3.4.1 Local least squares 35
3.4.2 Galilean interpretation 35

II Contributions **37**

4 Local space-time features **39**
4.1 Detection . . . 40
4.2 Velocity correction . . . 42
4.3 Examples of detected features . . . 43
4.4 Discussion . . . 46

5 Scale and velocity adaptation **49**
5.1 Spatio-temporal scale selection . . . 50
5.1.1 Scale selection in space-time . . . 50
5.1.2 Scale-adapted local space-time features . . . 52
5.1.3 Experiments . . . 54
5.2 Velocity adaptation . . . 57
5.2.1 Velocity and scale adaptation of events . . . 59
5.2.2 Discussion . . . 61
5.3 Dense velocity adaptation . . . 62
5.3.1 Mechanism for dense velocity adaptation . . . 63

6 Motion descriptors **67**
6.1 Local space-time descriptors . . . 68
6.1.1 Spatio-temporal N-Jets . . . 71
6.1.2 Principal Component Analysis . . . 72
6.1.3 Histogram-based descriptors . . . 75
6.1.4 Summary . . . 77
6.2 Dissimilarity measures . . . 77
6.3 Motion representations . . . 79
6.3.1 Greedy matching . . . 79
6.3.2 Quantization of local events . . . 80

7 Evaluation **83**
7.1 Stability under scale variations . . . 85
7.1.1 Spatial scale . . . 85
7.1.2 Temporal scale . . . 89
7.2 Stability under velocity variations . . . 89
7.3 Evaluation of local motion descriptors . . . 94
7.3.1 Jet-based descriptors . . . 98
7.3.2 Histogram-based descriptors . . . 99
7.3.3 Comparison to other methods . . . 100
7.4 Evaluation of dense velocity adaptation . . . 103
7.4.1 Experimental setup . . . 104
7.4.2 Discriminability of histograms . . . 106
7.4.3 Discriminability measure . . . 106
7.4.4 Dependency on scales . . . 106

7.4.5 Summary and discussion 109

8 Towards applications **111**
8.1 Recognition of human actions 112
8.2 Sequence matching 116
8.2.1 Walking model 116
8.2.2 Model matching 118
8.2.3 Results and discussion 119

9 Summary and discussion **121**
9.1 Future work 123

Bibliography **125**

Chapter 1

Introduction

Motion in images carries important information about the external world and changes of its structure. In the life of animals, both the potential danger and the source of food frequently comes in association with object motion and requires immediate reaction. Hence, it is not surprising that some of the primitive biological visual systems such as these of frogs (Maturana et al., 1960) have been found to be particularly sensitive to motions in specific directions. Moreover, psychological studies of humans uncover sophisticated mechanisms of motion perception, such as recognizing individuals from their gait patterns (Cutting et al., 1978) and perceiving the three-dimensional structure of objects from motion (Wallach and O'Connell, 1953).

Computational theories of vision address the area of motion analysis in a number of different ways. One type of methods attempts to estimate the spatial properties of rigid objects from motion. Horn (1987) showed that the *motion field*, i.e. the projection of 3-D motion of points in the scene onto the image plane, almost always allows for unique reconstruction of the 3-D structure of a rigid scene. As the motion field is not directly accessible from images, the problem of reconstruction, known as "structure from motion", is non-trivial and has been extensively studied during the past decades. Another approach explores the relative motion between objects and their backgrounds in order to detect the regions of objects in images. This problem, known as "figure-ground segmentation", is related to the Gestalt law of "common fate" (Wertheimer, 1958) stating that portions of an image that move together tend to be parts of the same object or surface.

Besides rigid motions, there is a large class of non-rigid motions including articulated motion, motion of elastic materials, fluids and gases. To classify complex motions in image sequences, Polana and Nelson (1997) suggested to consider their repeatability in time and space and defined three classes of motion: *temporal textures*, *activities* and *motion events*. Temporal textures are defined by statistical regularities over both space and time and include examples such as sea waves, motion of clouds and trees, fluttering leaves, motion of birds in a flock, etc. Activities

are defined by repeatable structures over time but not over space and correspond to scenes with people walking, dancing, talking, as well as to individual motions of animals such as snakes, birds, insects, etc. Finally, motion events consist of isolated simple motions without either spatial or temporal repetition and include events such as throwing a ball, entering a room, starting a car, as well as human gestures and changes in facial expressions.

Complex motions can often be characterized by distinctive dynamical properties in terms of relative motion, changes in velocities and appearance and disappearance of structures due to occlusions. As shown in a number of previous works, such properties can often be captured from image sequences and be used for motion interpretation directly, i.e. without an intermediate step of geometric reconstruction. This idea, as opposed to the interpretation of static images, is known as "motion-based recognition" (Cedras and Shah, 1995) and constitutes the basis for a novel and developing direction in computer vision. The potential applications of motion-based recognition include video browsing and retrieval, surveillance, human-computer interactions, robot navigation, sports, recognition of sign language, lip reading and others.

Many interesting ideas for the solution of motion-based recognition have been proposed. For example, Heeger and Pentland (1986) used spatio-temporal fractals to recognize image areas with turbulent motion, Allmen and Dyer (1990) used the curvature of spatio-temporal curves and surfaces for interpretation of cyclic motions, Doretto, Chiuso, Wu and Soatto (2003) represented and recognized dynamical textures using auto-regressive models. For articulated motions, the trajectories of object parts such as legs and arms have been frequently used for interpretation of human actions (Gavrila and Davis, 1996; Bregler, 1997; Yacoob and Black, 1999). Motion templates in terms of image differences or optical flow have been used for representing, learning and recognizing activities (Fleet et al., 2000; Hoey and Little, 2000; Bobick and Davis, 2001; Efros et al., 2003; Viola et al., 2003). Statistics of spatio-temporal filter responses have been used for recognizing human actions (Chomat and Crowley, 1999; Zelnik-Manor and Irani, 2001) as well as for video indexing and retrieval (Fablet et al., 2002).

Although many impressive results have been reported during the last decade (see Chapter 2 for related work), most of the current methods share a number of common weaknesses. First of all, most of the methods assume that localization of motion patterns is done *prior* to recognition. In the case of template matching, an external module is often required for estimating positions, velocities and scales of the pattern in the image. Other methods require even more specific information, such as accurate object boundaries and a localization of its parts. Whereas motion-based segmentation can often be used to solve the localization problem in scenes with simple backgrounds, scenes with complex, non-stationary backgrounds and multiple motions make the robust localization problematic. In such situations, it is not the presence but the type of motion that is the primal cue for both localization and interpretation of motion patterns. Hence, in the context of motion-based recognition, the ability to perform robust localization prior to recognition cannot

be assumed in general.

Another issue concerns the generalization of methods to different types of motion and the possibility of automatic learning of new motion classes from training data. Whereas many of the current methods use specific representations for each motion class, it is desirable to construct and use general purpose representations that are capable of expressing a variety of motion classes.

Finally, the presence of background clutter, multiple motions and occlusions influence the dense motion descriptors (e.g. templates and histograms) and makes the tracking of part-based models difficult. Although no systematic evaluation of this problem has been made, it is likely that this type of distraction is often a problem. In summary, the common limitations are

- *localization* of motion patterns is done *prior to recognition*
- use of *specific representations* for each motion class
- influence of *clutter and multiple motions*

Several works address one or more of these issues. For example, dense representations of motion in terms of histograms (Zelnik-Manor and Irani, 2001) or polynomial basis functions of optical flow (Hoey and Little, 2003) are not restricted to particular types of motions and can be learned from the training data. However, both of these methods rely on a pre-segmentation step. Probably the only method that currently addresses all of the listed limitations is by Song et al. (2003) who use local image features and triangulated graphs for unsupervised learning of human motions in complex scenes.

1.1 Contributions of the thesis

This work presents an attempt to address all of the problems mentioned above in a single framework. The main guideline has been to avoid difficult problems, such as segmentation, and accurate motion estimation by directly using *reliable* information in the spatio-temporal data that is applicable for motion interpretation.

We use the notion of *local motion events*[1] that capture local image structures at moments of non-trivial changes. A related idea has been proposed earlier by Rubin and Richards (1985) who suggested to decompose and represent motions using motion primitives such as starts, stops and discontinuities in velocity and acceleration. Here, we develop this idea and construct a framework for detecting local motion events directly from image sequences. As will be shown by experiments, such events have stable locations in space-time and correspond to local image structures at the moments of appearance, disappearance and non-constant motion. One motivation for using such events is that non-constant motion in many cases corresponds to

[1]The terms "local motion events" and "local space-time features" have the same meaning throughout this thesis.

accelerations, hence, often reveals information about the physical forces acting in the observed scene.

The structure of motion patterns in spatio-temporal data is frequently influenced by variations in spatial scale, temporal frequencies and the relative motion of the camera. In order to compensate for these variations and to detect motion events independently of the corresponding image acquisition conditions, we develop a method for adapting motion events to their intrinsic scales and velocities. The information about the intrinsic parameters is obtained from spatio-temporal neighborhoods of the events and is then used in an iterative update procedure (see Chapter 5). The presented approach for local spatio-temporal adaptation is not limited to the adaptation of local motion events but can be used in other methods as an alternative to global camera stabilization and spatio-temporal alignment.

Next, we develop and evaluate a set of local motion descriptors that capture and describe the spatio-temporal structure of events. By experiments on real image sequences, we show that by using such descriptors it is possible to classify different events and to match similar events in different image sequences. In particular, by a comparative study (see Chapter 7) we find that motion descriptors in terms of local, position-dependent histograms give the best performance in terms of matching. A similar result has been reported in the field of spatial recognition, where a local histogram-based *SIFT* descriptor (Lowe, 1999) has been found to outperform other types of local image descriptors (Mikolajczyk and Schmid, 2003).

Adapted motion events supplemented by discriminative descriptors provide a potential basis for representing complex motions in video. Here, we propose two types of event-based video representations and evaluate them on the problem of detecting and recognizing human actions. We explore the advantages of event-based representations in terms of locality and invariance. Given the invariance property with respect to scales and velocity, we show that motion recognition can be performed over different scales and camera motions without need for segmentation or alignment. Moreover, the sparseness of the event-based representations makes them tolerant to the presence of complex, non-stationary backgrounds and occlusions. This property is demonstrated by successful detection and recognition of human actions in complex scenes.

1.2 Organization of the thesis

The thesis is organized as follows. Chapter 2 presents a brief overview of related work on motion interpretation. Next, Chapter 3 describes a theory which is used as a basis for analyzing spatio-temporal data later in this work. In particular, Sections 3.1 and 3.3 present the theory of Gaussian scale-space and its extension to the spatio-temporal domain while Section 3.4 considers estimation of motion from image sequences. Chapter 4 introduces a method for detecting local motion events and illustrates by examples how the resulting features often correspond to meaningful events in image sequences. Chapter 5 presents a theory for local adaptation with respect to scale and velocity and describes an algorithm for detecting scale

and velocity co-variant local space-time features. Chapter 6 introduces a set of motion descriptors that are used for describing and classifying local motion events. In Chapter 7 we evaluate the proposed methods for detection, adaptation and description with regard to their stability under scaling and velocity transformations in the data. In Chapter 8, we present applications of local space-time features to the problem of detecting and recognizing human actions in video sequences. Finally, Chapter 9 concludes the work with a summary and a discussion of the presented approach and possibilities for extensions.

1.3 List of publications

This thesis is partly based on the following publications:

Laptev, I. and Lindeberg, T. (2004). Local descriptors for spatio-temporal recognition, *Proc. ECCV Workshop on Spatial Coherence for Visual Motion Analysis*, Prague, Czech Republic, *to appear*

Laptev, I. and Lindeberg, T. (2004). Velocity adaptation of space-time interest points, *Proc. International Conference on Pattern Recognition*, Cambridge, UK, *to appear*

Lindeberg, T., Akbarzadeh, A. and Laptev, I. (2004). Galilean-corrected spatio-temporal interest operators, *Proc. International Conference on Pattern Recognition*, Cambridge, UK, *to appear*

Schüldt, C., Laptev, I. and Caputo, B. (2004). Recognizing human actions: a local SVM approach, *Proc. International Conference on Pattern Recognition*, Cambridge, UK, *to appear*

Laptev, I. and Lindeberg, T. (2003). Interest points in space-time, *Proc. Ninth International Conference on Computer Vision*, Nice, France, pp. 432–439.

Laptev, I. and Lindeberg, T. (2003). Interest point detection and scale selection in space-time, *in* L. Griffin and M. Lillholm (eds), *Scale-Space'03*, Vol. 2695 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, pp. 372–387.

Laptev, I. and Lindeberg, T. (2004). Velocity-adaptation of spatio-temporal receptive fields for direct recognition of activities: An experimental study, *Image and Vision Computing* **22**(2): 105–116.

Publications beyond this thesis include:

Laptev, I. and Lindeberg, T. (2003). A distance measure and a feature likelihood map concept for scale-invariant model matching, *International Journal of Computer Vision* **52**(2/3): 97–120.

Laptev, I. Mayer, H., Lindeberg, T., Eckstein, W., Steger, C. and Baumgartner, A. (2000). Automatic extraction of roads from aerial images based on scale-space and snakes, *Machine Vision and Applications* (12): 23–31.

Part I

Background

Chapter 2

Related work on motion interpretation

This thesis is related to several areas in computer vision including, motion-based recognition, representation and estimation of motion, interpretation of human actions, recognition in terms of local measurements, scale-space theory, automatic scale selection and anisotropic filtering. In this chapter we give an overview of some of the main contributions in these areas with emphasis on motion representation and motion-based recognition. In the next chapter we present parts of the scale-space theory and describe methods for motion estimation with close relation to the methods used later in this work.

Motion in image sequences has been explored in many different ways. The existing methods can roughly be divided into two classes concerning either rigid motion or non-rigid motion. Rigid scenes provide strong geometrical constraints that can be used for reconstructing three-dimensional structure of scenes from multiple views and for matching different views in images. The geometry of rigid scenes has been a subject of extensive research in computer vision and several text books have recently been published on this topic (Faugeras, 1993; Hartley and Zisserman, 2000; Faugeras et al., 2001). In the context of geometric reconstruction, motion can be used to obtain correspondences between different views of the same scene. Horn (1987) showed that the motion field, which is the projection of motion of 3-D points onto the image plane, almost always allow for unique rigid geometric reconstruction. In practice, however, the motion field is not directly available from image sequences and has to be estimated. Several methods have been proposed to solve the problems of motion estimation and recovery of three-dimensional structure for example by tracking sets of sparse image features, e.g. (Hartley, 1994; Beardsley et al., 1997; Soatto and Perona, 1998; Nistér, 2001) or by "direct methods" using optic flow, e.g. (Hanna, 1991; Kumar et al., 1994; Irani et al., 2002).

In this thesis, we mainly consider non-rigid motion; hence the remaining overview is restricted to this area. Whereas rigid motion is completely described by

the fixed three-dimensional structure of the scene and the relative motion of the camera, non-rigid motion also originates from dynamic changes of objects in the world. Non-rigid motion therefore has more degrees of freedom than rigid motion, which makes the interpretation process more difficult. Changes in scene structure, however, are usually restricted by physical properties of the objects and the type of motion at a given moment of time. Hence, these restrictions can be used to constrain the interpretation process and to determine the three-dimensional structure of a non-rigid object and/or the type of its motion. In particular, recognizing the type of motion directly from image sequences, e.g. walking, running, cycling, etc, has often shown to be more important and less hard than exact geometric reconstruction. This type of methods, opposed to the recognition of static objects, is known as *motion-based recognition* and has been investigated by many researchers in the past, see e.g. (Cedras and Shah, 1995; Shah and Jain, 1997) for overviews.

Among the numerous examples of non-rigid motions, including the motion of water, trees, clouds and animals, etc., the activity of people is a most interesting class of non-rigid motions for practical applications. Recognition of human actions is particularly important for applications in surveillance, human-computer interaction and browsing of video databases. Pose estimation and three-dimensional reconstruction of human bodies, gestures and faces can be used for medical studies, sports and animations. Whereas extensive review of recent work on human motion analysis can be found in (Aggarwal and Cai, 1999; Gavrila, 1999; Moeslund and Granum, 2001), we will here focus on the main types of existing methods and outline their advantages and limitations. The following review will in particular emphasize the aspect of motion representation involved in different methods.

2.1 Structural methods

Structural methods use parametrized models describing geometric configurations and relative motions of parts in the motion pattern. For human bodies, the parts typically correspond to the limbs whose positions are estimated and tracked using image measurements. Much of research has been focused on modeling and tracking human bodies using three-dimensional models (Gavrila and Davis, 1995; Rohr, 1997; Bregler and Malik, 1998; Deutscher et al., 2000; Sidenbladh and Black, 2001; Sminchisescu and Triggs, 2003). The problem of three-dimensional reconstruction of non-rigid motion is difficult due to the large number of free parameters that have to be estimated. For human bodies, the estimation of joint angles between body parts typically requires the estimation of 20-30 parameters whose values are not directly available from the image data. To solve the problem, a number of constraints are usually introduced to reduce the search. These constraints are usually imposed on geometric configurations and on motions. Harder constraints, such as assuming particular types of motion, e.g. walking or running, make the parameter estimation less ambiguous but restrict the domain of application. Initial estimation of parameters (initialization) and recovery from failures of tracking (re-initialization) are

hard problems which are seldom addressed by current methods. To address the initialization problem, Sullivan and Carlsson (2002) used the repetitive nature of human actions in sport sequences and combined three-dimensional tracking with appearance-based recognition of key frames.

Motion recognition does not always require three-dimensional reconstruction and can be approached using structural models in terms of image features. Tracking of image features over time and analyzing the resulting trajectories has been used for human activity recognition (Baumberg and Hogg, 1996; Bregler, 1997; Blake and Isard, 1998; Song et al., 2003), for gesture recognition (Bobick and Wilson, 1995; Isard and Blake, 1998; Black and Jepson, 1998b) and for analyzing facial expressions (Bascle and Blake, 1998). Bregler (1997) used blobs of coherent motion and applied Hidden Markov Models for supervised learning and recognition of human actions. Song et al. (2003) used tracks of point features combined in triangulated graphs to automatically learn human actions from unlabeled training data. Isard and Blake (1998) and Black and Jepson (1998b) applied particle filtering for tracking and recognizing hand gestures. Contour-based models have been used to learn and to track silhouettes of human bodies in (Baumberg and Hogg, 1996; Elgammal et al., 2003) and to capture and to recognize facial expressions by Bascle and Blake (1998).

The use of structural models provides explicit locations of parts which leads to advantages for applications such as human-computer interfaces and motion animation. Tracking of parts, however, is a hard problem due to the high degree of freedom of structural models and due to the possibility of ambiguous interpretations in complex scenes. As a result, the existing structural methods have not been shown to provide reliable motion interpretation in unconstrained situations with multiple motions, heterogeneous backgrounds and partial occlusions. Moreover, most of the existing methods use specific models without providing automatic methods for model construction. An exception to this is the method by (Song et al., 2003), which gives a hope for extending structural methods to a wide range of applications.

2.2 Appearance-based methods using motion templates

An alternative to the structural approaches consists of modeling the changes of object appearance in images sequences. This method is closely related to appearance-based methods in the spatial domain and consists of comparing dense maps of spatio-temporal image measurements in the training set and in the test set. Nan et al. (1997) used image intensities in the spatio-temporal image volumes and constructed models for speech in terms of "eigensequences". By applying the principal component analysis to encode image variations within sequences, this method was able to represent complex motions of lips with a few parameters which values were used for the classification.

The values of image intensity frequently depend on many external conditions such as lightning, individual variations of clothing and other factors. To obtain

independence of these conditions when recognizing motion, several methods used image descriptors in terms of motion information only. Among the different techniques developed in this context, Bobick and Davis (2001) constructed temporal templates in terms of Motion History Images where the intensity represented a function of recency of motion. Motion History Images were obtained by segmenting motion patterns in each frame of a sequence and were tested on the task of recognizing aerobics exercises.

To avoid accurate spatial segmentation, several authors used optic flow for representing and matching dynamic image sequences (Black et al., 1997; Yacoob and Black, 1998; Hoey and Little, 2000; Hoey and Little, 2003; Efros et al., 2003). Black et al. (1997) and Yacoob and Black (1998) used principal component analysis to encode optic flow in each frame by a linear combination of orthogonal basis flow fields. The coefficients associated with each basis flow were recorded over time and were used for representing and recognizing image sequences of faces and human bodies in motion. A similar approach was presented by (Hoey and Little, 2000; Hoey and Little, 2003) who proposed a general basis for representing optic flow using Zernike polynomials. The advantage of this method is that the lower order motions are encoded separately from the higher order motions. This allows for separation between relative motions of the camera and class specific motions, such as facial expressions. Efros et al. (2003) also used optic flow, but in contrast to the other described methods constructed spatio-temporal models of human actions where the positive and the negative components of the flow were smoothed and then used for matching and recognizing particular actions in image sequences of ballet, tennis and football.

In comparison to structural methods, appearance-based methods have a lower degree of freedom resulting in the lower ambiguity of matching. These methods also allow for automatic construction of motion models and can therefore be used for representing and recognizing a large variety of motion classes. The structure in these appearance-based representations, however, is not present explicitly. Moreover, these methods rely on either spatial segmentation, spatial alignment or spatio-temporal registration of image sequences prior to recognition. This requirement is a drawback and is similar to the initialization problem of structural methods mentioned above. Moreover, the appearance-base methods consider all motions in the spatio-temporal window and can therefore be disturbed by occlusions as well as irrelevant motions in scenes with non-static backgrounds.

2.3 Statistical appearance-based methods

Both the structural methods and the appearance-based methods search for explicit correspondences between the models and the image structures in test sequences. Finding correspondences is generally known as a hard problem due to the common presence of outliers, missing data and variations of patterns within a class. If the task, however, is to classify motion patterns only, the requirement of finding

explicit correspondences can be relaxed. In this case, position-independent image measurements can be used in combination with statistical or non-parametric methods for classification. In a simplified formulation, such methods answer the question "What" motion is present in the scene rather than the questions "What and Where" (appearance-based methods) or "What, Where and How" (structural methods).

Several non-parametric methods for motion interpretation have been described in the literature. Polana and Nelson (1992) computed first and second order statistics of normalized spatio-temporal gradients and used such position-independent quantities to classify "temporal textures", including sequences of flowing water, waving cloth, fluttering paper and others. Extensions of this method were later presented by Chomat and Crowley (1999), who used histograms of spatio-temporal filter responses for recognizing human gestures and activities as well as by Zelnik-Manor and Irani (2001) who used marginalized histograms of spatio-temporal gradients computed at multiple temporal scales to classify human actions and to cluster events in image sequences with repetitive activities. The events in this method were defined by activities performed over a limited time interval.

In (Doretto, Chiuso, Wu and Soatto, 2003), variations of image values over time were assumed to be realizations of a second-order stationary stochastic process with a Gauss-Markov model. This approach can be seen as a similar to the histogram-based methods above with the difference that the distribution of the temporal changes in images was here described by a parametric statistical model. The parameters of the distribution were estimated from each image sequence and were used for classification of dynamic textures, such as flowing water, fire, smoke and others. An advantage of this method is that, besides motion recognition, it also allows for synthesis of dynamic textures.

Fablet et al. (2002) used a conceptually similar method and described temporal variations of quantized spatio-temporal gradients using temporal coocurrences and non-parametric causal Gibbs models. This method was applied for the classification of video sequences and for the retrieval of video episodes by examples. The considered image sequences corresponded to 20 different scenarios including basketball, hockey, rugby, windsurfing, highways and others.

Statistical methods allow for automatic learning and classification of motions in image sequences without computing *explicit* correspondences. Such methods, however, still require the presence of *implicit* correspondences between the image measurements; i.e. local image measurements have to correspond within the considered spatio-temporal regions. This implies that statistical methods do not avoid the segmentation problem and depend on background motions, camera motions, occlusions, etc. In scenes with static backgrounds, spatial segmentation can be achieved using the presence of motion as a cue (Zelnik-Manor and Irani, 2001). In complex scenes, however, such an approach can be expected to have difficulties. An interesting approach in this respect has been presented by Doretto, Cremers, Favaro and Soatto (2003) who combined segmentation and recognition of dynamic textures in a single statistical framework. Local velocity adaptation of spatio-temporal derivatives has been presented by Laptev and Lindeberg (2004c) to compensate for

camera motions and as an alternative to global methods for camera stabilization, which can be unreliable in scenes with multiple motions.

In this thesis, we will use different versions of appearance-based methods and statistical methods to formulate local descriptors of motion in local spatio-temporal neighborhoods of image sequences as described in Chapter 6. In combination with methods for local event detection and adaptation, such descriptors will be shown to allow for motion recognition in complex scenes as described in Chapters 7 and 8.

2.4 Event-based motion interpretation

All of the methods described above rely on a correspondence (implicit or explicit) of spatio-temporal image patterns in the test set and in the training set. The computation of explicit correspondences has an advantage, since it allows for the localization of motion patterns in the sequence and can be used to reject outliers such as background motion. To obtain explicit correspondences, most of the described methods use the *spatial* variations of moving patterns as a discriminative cue for matching. *Temporal* variations, however, also provide discriminative information which can be used for matching and localizing motion patterns over time. In combination, spatial and temporal cues can be used to enhance the matching by localizing image structures with specific spatio-temporal properties. Such structures have often been regarded as motion events or motion units.

An early idea of event-based motion representation in computer vision has been discussed in (Rubin and Richards, 1985; Engel and Rubin, 1986). In this work, the authors considered significant changes in motion as "motion boundaries" and identified motion boundaries, such as smooth starts, smooth stops, pauses, impulse starts and impulse stops. Motion boundaries (or motion events) were defined by discontinuities in the curvature of motion trajectories. The authors motivated their approach using considerations from kinematics and argued that motion boundaries correspond to force discontinuities in the real world. Moreover, the approach was compared to psychophysical experiments where the motion perception of people was found to be related to the concept of motion boundaries (Runeson, 1974). The approach is also related to the well-known experiment on Moving Light Displays (MLD), where the ability of humans to perceive structure and motion from the motion of sparse point features attached to a human body has been demonstrated by Johansson (1976). The method, however, was not tested on real images.

A similar approach was later presented by Gould and Shah (1989), who defined a Trajectory Primal Sketch and used discontinuities in velocity and acceleration to represent the motion. In (Rangarajan et al., 1992), motion events were found as zero-crossings of smoothed trajectories in terms of velocity and direction of motion, and were used to match image patterns of walking people. The trajectories were extracted manually from several body points. Recently, a further development of this approach and its application to automatic interpretation of hand actions was presented by (Rao et al., 2002). In this work, the hand of a person was automatic-

ally tracked in image sequences and motion events in the resulting trajectory were matched to trajectories in training sequences.

Besides changes in the trajectories of a single object, interactions between objects such as splits and unifications also correspond to meaningful events that often can be described by verbs such as "pick up", "place", "attach", "detach", "touch" and so forth. To detect this kind of events, Brand (1997) used causal reasoning about motions and collisions of surfaces. Using psychological arguments, he emphasized the importance of visual events, such as appearance, disappearance, inflation, deflation, flash, acceleration and discontinuity. He then considered the task of parsing image sequences into primitive events using examples of instructional videos, such as disassembling a computer in a constrained environment. The method for event detection, however, was based on the spatial segmentation of moving objects and is therefore not directly suitable for complex scenes. Related work using event-based reasoning with connections to linguistics and Newtonian mechanics was presented in (Kuniyoshi and Inoue, 1993; Siskind and Morris, 1996; Mann et al., 1997).

Another type of approach for temporal segmentation of motions into events was presented by (Rui and Anandan, 2000; Zelnik-Manor and Irani, 2001). Rui and Anandan (2000) considered global descriptors of image sequences in space in terms of optic flow. The variations in the flow were then analyzed over time and significant temporal changes in the flow were regarded as action boundaries. This approach is related to the detection of motion discontinuities in temporal trajectories used in (Rubin and Richards, 1985; Engel and Rubin, 1986; Rangarajan et al., 1992; Rao et al., 2002). The authors applied their method to temporal segmentation of video sequences of people performing common household activities such as making a bed. The result of this approach is particular interesting, since the resulting action boundaries were found to coincide with the results of manual segmentation of the same sequences performed in psychological experiments by Zacks et al. (2000).

Zelnik-Manor and Irani (2001) used global statistical spatio-temporal image descriptors to describe actions in spatio-temporal image volumes (see Section 2.3). The image descriptors were computed for subsequences and then clustered to detect subsequences with similar events. The method was applied to outdoor video sequences with people performing actions, such as walking, jogging, hand-waving as well to a tennis sequence. The result of clustering resulted in a successful classification of subsequences into actions, for example for a tennis sequence the system reliably detected events corresponding to strokes, steps and hops.

2.4.1 Relevant research in psychology

The idea of motion interpretation in terms of events has been studied in psychology for a long time (Newtson et al., 1977; Zacks et al., 2000; Lassiter et al., 2000; Tversky et al., 2002). In experiments performed in early 70's, Newtson et al. (1977) found that people tend to segment ongoing behavior into consistent episodes of actions, also called units or events. Moreover, it was found that people could extract more information from the ongoing behavior if they were asked to segment

the activities into smaller units. Inversely, perceivers who lacked prior knowledge about the observed behavior tended to segment finer units of actions compared to experienced perceivers if both groups were given a task to remember and to describe the behavior. Given the consistency of units selected by different people, this evidence strongly support the idea that people perceive and possibly represent and recognize the motion (at least partly) in terms of events with well-defined temporal boundaries. Besides the computational arguments discussed in the beginning of this section, these psychological arguments serve as a additional motivation for event-based motion recognition when constructing an artificial visual system.

2.4.2 The approach in this thesis

The existing computational methods for event-based motion interpretation either rely on spatial segmentation and tracking or use global image measurements. This makes the existing methods potentially unreliable in complex unconstrained scenes with multiple independent motions, objects and events. Moreover, the existing methods detect different kind of events, for example motion boundaries (Rubin and Richards, 1985), which can be used to detect discontinuities in velocity of an object but are not suitable to detect the appearance of the object, such as when a person enters the room.

In this thesis, we will propose an alternative approach for event detection and event-based motion interpretation where we will detect events using *local* information in image sequences *only* (see Chapters 4-6). Hence, the resulting approach will avoid the problems associated with the spatial segmentation. Moreover, it will be independent from the global variations in the scene. The evaluation of the presented event-based scheme for motion interpretation will be demonstrated for scenes with unconstrained environments in Chapter 8.

Chapter 3

Computational theory

Interpretation of visual information involves comparison of images taken under different conditions and at different moments in time. As images representing the same scene or a class of objects might be very different depending on the view, the lightning, etc., there is a need for invariant representations that emphasize the important properties of the image while suppressing irrelevant variations.

One important source of variations originates from the fact that images are usually obtained by the perspective projection of light onto a planar sensor with a finite resolution. This makes images dependent on the position of the camera in the scene as well as on the internal camera parameters such as the focal length and the resolution of the sensor. In particular, the distance to the observed object and the values of the internal camera parameters effect the *scale* of the object in the image. Moreover, changes in the orientation of an object relative to the camera introduce perspective deformations to the image. The motion of objects relative to the camera effects the spatio-temporal structure of the video data.

In genera, the transformations that effect the image data are not known in advance. In order to estimate the properties of the data independently of the external conditions, one fundamental approach consists of constructing representations of the image data that are *closed* under an interesting set of transformations. Given such a representation, it is then possible to estimate the values of the transformation parameters that relate pairs of images or video sequences and to obtain the desired invariance. This idea is illustrated in Figure 3.1.

For the case of scale variations, we can construct a multi-scale representation of an image by introducing a scale parameter σ that refers to scaled versions of the same image with different amount of scaling. Such a representation with a continuous σ-parameter was introduced by Witkin (1983) and has been developed into the scale-space theory (Koenderink, 1984; Yuille and Poggio, 1986; Lindeberg, 1994; Florack, 1997) which has been extensively studied during the last two decades. Many different types of scale-spaces have been proposed (ter Haar Romeny, 1994) including linear scale-spaces (Koenderink, 1984), non-linear scale-spaces (Perona

$$\begin{array}{ccccc} \widetilde{D} & - & F & \rightarrow & \widetilde{R} \\ \uparrow & & & & \downarrow \\ T_D & & & & T_D^{-1} \\ \downarrow & & & & \downarrow \\ D & - & F & \rightarrow & R \end{array}$$

Figure 3.1: Images D and $\widetilde{D}$ are related by a transformation T_D. Given image representations R and $\widetilde{R}$ constructed by F and closed under T, $T_D \in T$, it is possible to estimate T_D from R and $\tilde{R}$ and to "undo" the effect of transformation.

and Malik, 1990; Alvarez et al., 1992; Florack et al., 1995; Black et al., 1998), morphological scale-spaces (Alvarez and Morel, 1994) and recently proposed locally orderless images (Koenderink and van Doorn, 1999). In particular, the Gaussian scale-space generated by the convolution of images with a Gaussian kernel has been shown to be a natural choice for a multi-scale representation with respect to its mathematical properties and close relations to biological vision (Koenderink, 1984; Young, 1987; DeAngelis et al., 1995). In Section 3.1, we review parts of the Gaussian scale-space theory as well as a method for automatic scale selection (Lindeberg, 1998b) which makes it possible to compare image structures independently of their scales.

Concerning perspective transformations, their effect depend not only on the external and internal parameters of the camera, but also on the three-dimensional structure of the scene. Hence, unless the three-dimensional structure is known or estimated, it is in general not possible to construct image representations that are closed under perspective transformations. However, given the assumption of locally smooth surfaces, the perspective transformation can locally be approximated by affine transformations. An extension of an isotropic Gaussian scale-space to an affine scale-space (Lindeberg and Gårding, 1994; Griffin, 1996; Lindeberg and Gårding, 1997) that is closed under affine transformations is presented in Section 3.1.3. Similarly to the notion of scale selection, Section 3.1.4 presents an approach to automatic estimation of the affine transformation from the data in order to relate local image structures affected by different perspective transformations.

Finally, the extension of the scale-space concept into the spatio-temporal domain (Koenderink, 1988; Lindeberg and Fagerström, 1996; Florack, 1997; Lindeberg, 1997a; ter Haar Romeny et al., 2001) is presented in Section 3.2. Here the representation for temporal image sequences (video) is generated by the convolution with a three-dimensional Gaussian kernel with one scale parameter σ for the spatial domain and another, independent, scale parameter τ for the temporal domain. Moreover, to obtain closedness under constant motion of the camera relative to the scene, the separable spatio-temporal scale-space can be extended to a non-separable

Galilean scale-space (Lindeberg, 2002). Automatic adaptation of scales and velocities of local space-time image structures is one of the main contributions of this thesis and will be presented in Chapter 5.

Since the Gaussian function has infinite support in all directions while the temporal domain is causal, i.e. information is not available for the future, the Gaussian spatio-temporal scale-space is not well suited for real-time applications. To address this problem, Lindeberg and Fagerström (1996) introduced a causal scale-space generated by the convolution of the temporal domain with a set of recursive filters coupled in cascade, which in the limit case can be described by Poisson-type scale-space (Lindeberg, 1997b). The Poisson scale-space can be seen as a discrete counterpart to the Gaussian scale-space on a discrete temporal domain that respects temporal causality. This type of the scale-space is reviewed in Section 3.3.

At the end of this chapter, Section 3.4 shortly reviews a theory for estimating motion from images. Due to the huge scope of the related work and the existence of extensive reviews on the subject, e.g. (Barron et al., 1994; Jähne et al., 1999), we will here only summarize a few main results as well as a specific method for motion estimation that will be used later in this work.

3.1 Gaussian scale-space

Given an N-dimensional signal $f\colon \mathbb{R}^N \to \mathbb{R}$, its Gaussian scale-space representation $L\colon \mathbb{R}^N \times \mathbb{R}^+ \to \mathbb{R}$ can be defined by the convolution of f

$$L(x;\, \sigma^2) = \int_{x \in \mathbb{R}^N} f(\xi) g(x - \xi;\, \sigma^2) d\xi, \tag{3.1}$$

with a Gaussian kernel

$$g(x;\, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}} e^{-(x^T x)/2\sigma^2}, \tag{3.2}$$

where the variance σ^2 corresponds to the continuous scale parameter. An equivalent definition can be obtained by the solution of the diffusion equation

$$\partial_{\sigma^2} L = \frac{1}{2} \nabla^2 L \tag{3.3}$$

with initial condition $L(\cdot;\, 0) = f$. Intuitively, by increasing the scale parameter introduces more smoothing to the data and simplifies its structure. Similarly, the level of details in a scene decreases when viewed from different distances (e.g. consider watching leaves, trees and a forest).

Previous work has shown that the uniqueness of the Gaussian kernel can be derived from different sets of constraints defined on L. One set of such constraints introduced by Koenderink (1984) consists of *causality*, *isotropy* and *homogeneity*. Causality means that new level surfaces must not be created when the scale parameter is increased. Isotropy and homogeneity mean that all spatial positions and

scale levels must be treated in a similar manner. Other properties of the Gaussian scale-space are the *semi-group* property

$$g(\cdot;\, {\sigma^2}_1) * g(\cdot;\, {\sigma^2}_2) * f(\cdot) = g(\cdot;\, {\sigma^2}_1 + {\sigma^2}_2) * f(\cdot) \tag{3.4}$$

and the *scaling* property

$$L(x;\, t) = L'(sx;\, s^2\sigma^2), \tag{3.5}$$
$$L'(\cdot;\, s^2\sigma^2) = f'(\cdot) * g(\cdot;\, s^2\sigma^2), \quad f'(sx) = f(x)$$

which can be verified using the scaling property of N-dimensional Gaussian kernel

$$g(\xi;\, \sigma^2) = s^N g(s\xi;\, s^2\sigma^2). \tag{3.6}$$

The scaling property implies that the uniform scaling of the signal f corresponds to a *shift* in scale in the scale-space representation L. Hence, the Gaussian scale-space is closed under uniform scaling transformations.

The analysis of functions often requires their differentiation. Whereas for discrete functions the differentiation is an ill-posed operation, the derivatives of L can still be computed in a well-posed manner even for discrete functions $f\colon \mathbb{Z}^N \to \mathbb{R}$. This can be seen from the commutative property of the convolution and the differentiation

$$(\partial_{x^n} f) * g = \partial_{x^n}(f * g) = (\partial_{x^n} g) * f. \tag{3.7}$$

Hence, instead of computing derivatives of possibly discontinuous f, we can differentiate the continuous Gaussian function and convolve f by the Gaussian derivative. This property is very important, since it allows us to perform differential geometric analysis on any bounded function f. Note, that in practice the Gaussian kernel and its derivatives should be discretized both with respect to the coordinates x and the scale parameter σ^2.

The commutative property of differentiation and convolution also implies that the Gaussian derivatives satisfy the diffusion equation (3.3) and, hence, all the properties of the Gaussian scale-space including the semi-group

$$g_{x^m}(\cdot;\, {\sigma^2}_1) * g_{x^n}(\cdot;\, {\sigma^2}_2) * f(\cdot) = g_{x^{n+m}}(\cdot;\, {\sigma^2}_1 + {\sigma^2}_2) * f(\cdot), \tag{3.8}$$

and scaling property

$$L_{x^n}(x;\, t) = s^n L'_{x'^n}(x';\, \sigma'^2) \tag{3.9}$$
$$x' = sx, \quad \sigma'^2 = s^2\sigma^2$$

where n, m denote the order of differentiation. As can be seen, the scaling property for scale-space derivatives (3.9) differs from (3.5) by a factor s^n in the amplitude. However, if we introduce the dimensionless coordinates ξ

$$\xi = x/\sigma, \quad \xi' = x'/\sigma' \tag{3.10}$$

and the corresponding *scale-normalized* derivative operators

$$\partial_{\xi^n} = \sigma^n \partial_{x^n}, \quad \partial_{\xi'^n} = \sigma'^n \partial_{x'^n}, \tag{3.11}$$

the scaling property of scale-space derivatives exactly corresponds to the scaling property of the scale-space representation L

$$L_{\xi^n}(x;\, \sigma^2) = L'_{\xi'^n}(x';\, \sigma'^2). \tag{3.12}$$

Finally, the isotropy of the Gaussian scale-space L (3.1) makes it closed under rotations. For two-dimensional images $f(x, y)$ this implies that L is closed under rotations in the image plane

$$L(x, y;\, \sigma^2) = L'(u, v;\, \sigma^2) \tag{3.13}$$

where (u, v) is a coordinate system obtained by the rotation of (x, y) with angle α

$$\begin{pmatrix} u \\ v \end{pmatrix} = R(\alpha) \begin{pmatrix} x \\ y \end{pmatrix}, \quad R(\alpha) = \begin{pmatrix} \cos\alpha & sin\alpha \\ -\sin\alpha & cos\alpha \end{pmatrix} \tag{3.14}$$

Closedness under rotations also applies to spatial scale-space derivatives

$$L_{x^n y^m}(x, y;\, \sigma^2) = L'_{u^n v^m}(u, v;\, \sigma^2), \tag{3.15}$$

$$L'(\cdot;\, \sigma^2) = f'(\cdot) * g(\cdot;\, \sigma^2), \quad f'(u, v) = f(x, y)$$

where the derivative operators in the rotated and in the original coordinate systems are related according to

$$\begin{aligned} \partial_u &= \partial_x \cos\alpha + \partial_y \sin\alpha \\ \partial_v &= -\partial_x \sin\alpha + \partial_y \cos\alpha \end{aligned} \tag{3.16}$$

3.1.1 Scale and orientation estimation

The Gaussian scale-space representation is closed under rotations and scaling transformations. However, the representation itself does not provide information on how an image f relates to its transformed version f'. In other words, given L and L', we do not know the scaling s and the rotation R that put these two representations in correspondence. One way of estimating s and R can be to maximize the correlation between L and L' over all possible scalings and rotations. Another, more efficient way, is to try to estimate the *intrinsic* scales and orientations for each representation *separately* using the information in L and L' respectively. Here, we follow the latter approach and describe several methods for estimating the orientation and the scale of images locally using information in their local neighborhoods.

One common approach to estimate the local orientation in images is to consider the direction of the spatial gradient $\nabla L = (L_x, L_y)^T$ at each image point and to estimate the local coordinate transformation R (3.14) using

$$\cos\alpha = \frac{L_x}{\sqrt{L_x^2 + L_y^2}}, \quad \sin\alpha = \frac{L_y}{\sqrt{L_x^2 + L_y^2}}. \tag{3.17}$$

This method works well for patterns with well-defined orientations such as edges. However, the estimation of gradient orientation for nearly isotropic patterns such as blobs can be unstable. Some image patterns, e.g. stars, can be assigned several but limited number of stable orientations. In such situation one can estimate and assign several orientations to the same image point. Lowe (2004) proposed to consider histograms of orientations of image gradients accumulated in local image neighborhoods and to estimate local orientations by the significant maxima in such histograms. This approach has shown to be very robust in practice.

Estimation of orientations involves computation of image gradients $\nabla L(x, y; \sigma^2)$ and depends on the scale parameter σ^2. Hence, the estimation of scale is crucial for obtaining invariance with respect to both scalings and rotations of the image. An approach for estimating scales from local image neighborhoods has been proposed by Lindeberg (1993), (Lindeberg, 1998b) and is based on the evolution properties of normalized scale-space derivatives $\sigma^{n+m} L_{x^m y^n}(x, y; \sigma^2)$ over scales.

Consider a two-dimensional Gaussian function $g(x, y; \sigma_0^2)$ as a prototype of a blob-like image pattern with the radius proportional to the standard deviation σ_0. Using the semi-group property (3.4), the scale-space representation of this signal is obtained as

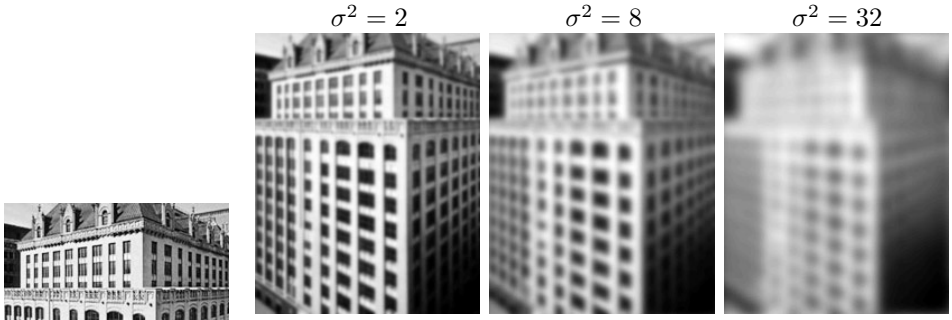
$$L(x, y; \sigma^2) = g(x, y; \sigma_0^2 + \sigma^2) \tag{3.18}$$

and its normalized Laplacian has the form

$$\nabla^2_{norm} L(\cdot; \sigma^2) = (\sigma_0 + \sigma)^2 (g_{xx}(\cdot; \sigma_0^2 + \sigma^2) + g_{yy}(\cdot; \sigma_0^2 + \sigma^2)). \tag{3.19}$$

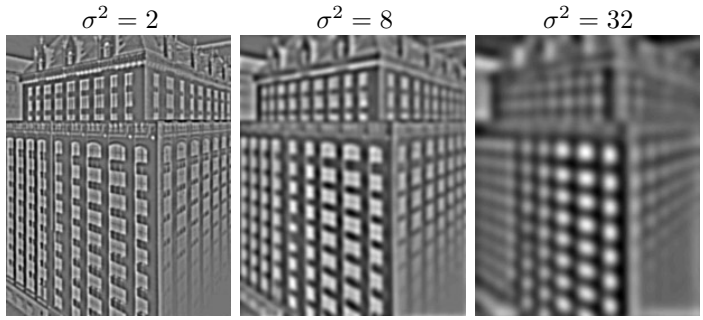
By differentiating (3.19) with respect to the scale parameter σ, it can be shown that the response of the normalized Laplacian operator $\nabla^2_{norm} L$ at the center of the Gaussian blob obtains *extremum* at the scale $\sigma{=}\sigma_0$. This result implies that given a blob-like image structure, we can estimate its size σ_0 by maximizing the response of $(\nabla^2_{norm} L)^2$ over scales. Such a strategy has shown to be very effective in practice for many image structures that are only approximatively similar to the ideal Gaussian blob. Moreover, a similar idea of maximizing normalized differential entities over scales has been proposed for other types of image structures such as ridges, corners and diffuse edges as well as for dense scale estimation (Lindeberg, 1998a).

The illustration of scale selection in practice is shown in Figure 3.2. Here, the original image of a building consists of many blob-like image structures (windows) with different sizes due to the variations in depth of the corresponding three-dimensional scene. The scale-space representation and the responses of the Laplacian operator for this image are shown for three different scale levels at the top-right of Figure 3.2. As can be noted, the responses of the normalized Laplacian operator give preference to different image structures at different scales depending on the size of these structures. By searching the maxima of the normalized Laplacian over both *space* and *scale* we find points shown by the circles at the bottom of Figure 3.2. As can be seen, the scales of detected blobs (proportional to the radius of circles) roughly correspond to the size of windows in the image. Moreover, the responses of the normalized Laplacian over scales clearly confirm its scale-selective property.

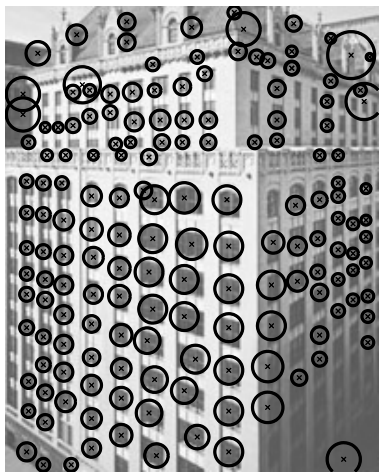


Scale-space representation at different scale levels

Original image



Responses of Laplacian operator at different scale levels



Laplacian features with scale selection

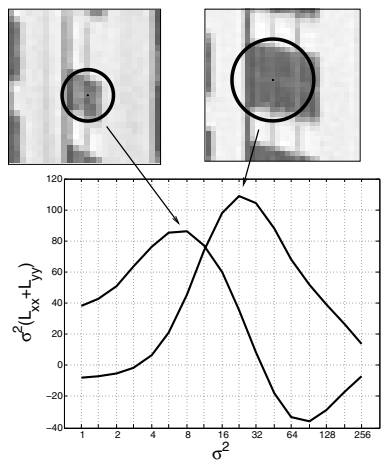


Figure 3.2: Top: Original image of a building, its scale-space representation and responses of the Laplacian operator at different scale levels. Bottom: Scale-space maxima of the normalized Laplacian operator estimate the positions and scales of blob-like structures in images. The signatures of $\nabla^2_{norm} L$ over scales show the scale-selective behavior of the normalized Laplacian operator.

3.1.2 Local descriptors

With the described methods, we can estimate the scale and the orientation of local image structures. The locality of these methods is an advantage, since any global method will in general depend on different subsets of information visible in different images. Moreover, Nielsen and Lillholm (2001) demonstrate that local features such as blobs and edges detected in scale-space have high information content and can be used for e.g. image reconstruction. However, the obtained estimates of features are not unique and we can have ambiguities over rotations, scales and positions in the image. Hence, in order to estimate the correspondence between images, we have to obtain the correspondence between their local image structures. This introduces the problem known as *matching*. Matching has been investigated in a wide range of different contexts including image and video indexing (Schmid and Mohr, 1997; Sivic and Zisserman, 2003), wide base-line matching (Tuytelaars and Van Gool, 2000; Mikolajczyk and Schmid, 2002; Tell and Carlsson, 2002), object recognition (Lowe, 1999; Fergus et al., 2003) optic flow estimation and tracking (Smith and Brady, 1995; Bretzner and Lindeberg, 1998) and others.

One way of approaching the matching problem consists of describing local neighborhoods of detected local image structures and then searching for the structures in different images with the most similar descriptors. One type of possible descriptors is a local jet (Koenderink, 1984) composed of image derivatives up to some order

$$\mathcal{J}(u, v; \sigma_0^2) = (\sigma_0 L_u, \sigma_0^2 L_{uu}, \sigma_0^2 L_{uv}, \sigma_0^2 L_{vv}, ..., \sigma_0^{n+m} L_{u^n v^m}). \tag{3.20}$$

Computation of derivatives for the estimated values of scales and orientations as well as the use of scale normalization makes it possible to compare local image structures at different scales and orientations.

Another interesting choice of descriptors is a position-dependent histogram of image gradients computed in the local neighborhood of a detected image structure. This type of descriptor is known as *SIFT* (Lowe, 1999) and has been shown to give very good performance in practice. An overview of different types of local descriptors and their evaluation can be found in (Mikolajczyk and Schmid, 2003).

3.1.3 Affine scale-space

The perspective projection of locally smooth surfaces onto an image plane can locally be approximated by the affine transformation. The affine transformation is a linear mapping $E\colon \mathbb{R}^2 \rightarrow \mathbb{R}^2$ that can be decomposed into two rotational components $R(\alpha)$, $R(\beta)$ (3.14) and one non-uniform scaling S as

$$E = R(\beta)R(-\alpha)SR(\alpha), \quad S = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix}. \tag{3.21}$$

To see how this transformation affects the scale-space representation, let us re-write the Gaussian kernel (3.2) for a two-dimensional case ($N = 2$) using the covariance

matrix Σ:

$$g(p;\,\Sigma) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-(p^T\Sigma^{-1}p)/2}, \quad \Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \tag{3.22}$$

where $p = (x, y)^T$ denotes image coordinates. Given the affine transformation of the coordinates $p' = Ep$, it can be easily verified that the Gaussian kernel transforms according to

$$g(p;\,\Sigma) = \det(E)g(Ep;\,E\Sigma E^T) = \det(E)g(p';\,\Sigma'). \tag{3.23}$$

We observe the analogy of this transformation with the earlier scaling transformation of the Gaussian in (3.6). The important difference, however, is that the transformed covariance matrix Σ' will no longer be of the form $\Sigma = sI$ for the identity I in general. This implies that the Gaussian scale-space described earlier and generated by the isotropic Gaussian kernel with $\Sigma = sI$ is *not* closed under affine image transformations. To extend the Gaussian scale-space to be closed under arbitrary affine transformations we consider L generated by the convolution of the image f with a Gaussian kernel having an *arbitrary* covariance matrix $\Sigma \in SPSD(2)$[1]

$$L(\cdot;\,\Sigma) = f(\cdot) * g(\cdot;\,\Sigma) \tag{3.24}$$

Concerning the differentiation, using the coordinate transformation $p' = Ep$ and the chain rule it can be verified that the derivative operators transform under E as

$$\begin{pmatrix} \partial_x \\ \partial_y \end{pmatrix} = E^T \begin{pmatrix} \partial_{x'} \\ \partial_{y'} \end{pmatrix}. \tag{3.25}$$

In particular, it follows that the scale-space gradient ∇L transforms under E according to

$$\nabla L(p;\,\Sigma) = E^T \nabla' L'(p';\,\Sigma'), \tag{3.26}$$

where $p' = Ep$, $\Sigma' = E\Sigma E^T$ and ∇' denotes differentiation with respect to p'.

3.1.4 Affine adaptation

A Gaussian kernel with an arbitrary 2×2 covariance matrix Σ has three free parameters (note, that Σ is symmetric). Hence, the scale-space representation L of an image generated by this kernel has five dimensions which makes it difficult to compute L in practice. However, if the goal is to estimate an affine transformation that has been applied to an image, we can use a search procedure in $L(x, y;\,\Sigma)$ that requires computation of L at some points only.

One procedure for affine adaptation has been introduced by Lindeberg and Gårding (1994) and is based on the transformation properties of the windowed second moment matrix defined as

$$\mu(p;\,\Sigma) = \int_{\xi \in \mathbb{R}^2} w(p - \xi) \left(\nabla L(\xi;\,\Sigma)(\nabla L(\xi;\,\Sigma))^T\right) d\xi \tag{3.27}$$

[1] $SPSD(2)$ stands for the cone of symmetric positive semidefinite 2×2 matrices.

where w is a window function, usually taken as a Gaussian $w(x, y) = g(x, y, ; s\Sigma)$ for some constant s. The second moment matrix contains information about the distribution of image gradients ∇L in the local neighborhood defined by w. Using the transformation property of ∇L (3.26), it can be shown (Lindeberg and Gårding, 1997) that under a linear transformation E, μ transforms as

$$\mu(p; \Sigma) = E^T \mu'(p'; \Sigma') E \tag{3.28}$$

where μ' is computed according to (3.27) using a transformed and normalized window function $w'(p'; \Sigma') = \det(E)^{-1} g(p'; \Sigma')$.

Assume now that $\mu(p; \Sigma) = M$ is in the standard form, i.e. $M = I$. Then, the corresponding second moment matrix $\mu'(p'; \Sigma') = M'$ in a transformed image will be related to M as

$$M' = E^{-T} M E^{-1} = (EE^T)^{-1}. \tag{3.29}$$

Using this relation, we obtain an estimate for E as

$$E = M'^{-1/2}. \tag{3.30}$$

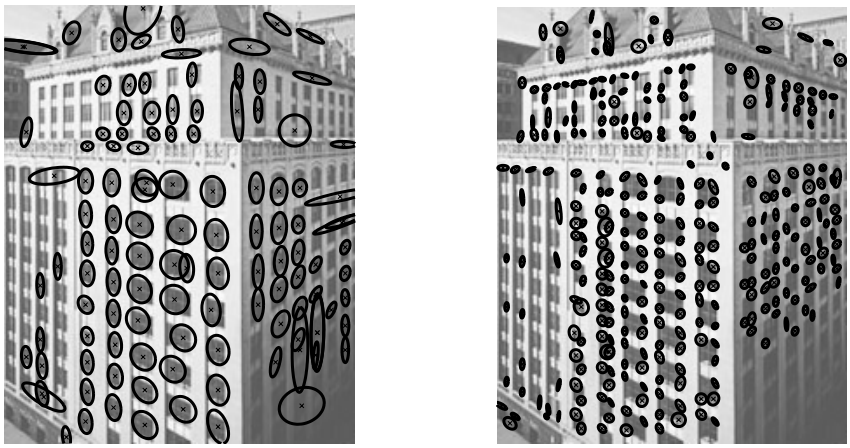
Several observations about this estimate should be made. Firstly, the magnitude of the second moment matrix depends generally on the changes of the image brightness. Hence, the estimate of E in (3.30) is only defined up to an arbitrary scaling factor σ. Secondly, E in (3.30) is defined up to an arbitrary rotation R since $M' = (EE^T)^{-1} = (ERR^T E^T)^{-1}$. Finally, the relation (3.30) holds only if M' is computed using the adapted covariance matrix $\Sigma' = \sigma^2 E \Sigma E^T = \sigma^2 EE^T$, hence, the so called fixed point relation $\Sigma' = \sigma^2 M'^{-1}$ must hold.

Since our aim is to estimate the affine transformation E, the correct matrix Σ' cannot be known in advance. However, by starting the estimation of E with an arbitrary Σ'_0 (usually isotropic, e.g. $\Sigma'_0 = \sigma^2{}_0 I$), we can use the obtained matrix M'_0 in order to re-estimate E in a new iteration. Moreover, in order to estimate the scale parameter σ and the rotation component R we can use methods for scale selection and rotation estimation described in Section 3.1.1.

An algorithm for iterative affine adaptation can be summarized as follows:

1. Set $\Sigma'_0 = I$.
2. Normalize for the magnitude $\Sigma'_i = \Sigma'_i / \det(\Sigma'_i)$.
3. Estimate scale $\sigma_i = \text{argmax}_\sigma (\nabla^2{}_{norm} L(\cdot; \sigma^2 \Sigma))^2$ (see Section 3.1.1)
4. Update $\Sigma'_{i+1} = (\mu'(\cdot; \sigma^2{}_i \Sigma'_i))^{-1}$; go to step 2 if not converged.
5. Set $E = \sigma_i R (\Sigma'_i)^{1/2}$ where R is an estimate of the rotation from $L'(\cdot; \sigma^2{}_i \Sigma'_i)$.

One possible criterion for the convergence in step 4 can be an estimate of the isotropy of a matrix $A = (\Sigma'_{i+1})^{-1} \Sigma'_i$. If A is isotropic, i.e. if the ratio of the eigenvalues $\lambda_{max}(A)/\lambda_{min}(A) \approx 1$, the fixed point condition has been reached and the iteration can be terminated.



Laplacian features with affine adaptation *Harris features with affine adaptation*

Figure 3.3: Detection of affine-adapted local features. Note how the estimated shape and scale of features represented by ellipses capture the shape and the size of corresponding image structures. The effect of adaptation can be observed by comparing affine-adapted Laplacian features and scale-adapted features (Figure 3.2).

Feature detection with affine adaptation. The described method for iterative affine adaptation is still rather expensive to be executed for all image locations. Several applications, however, such as image indexing, and recognition (see Section 3.1.2) require affine-invariant matching only at some sparse image locations regarded as local features or interest points. One method to detect local features using extrema of the Laplacian operator has already been presented in Section 3.1.1. Another popular approach (Förstner and Gülch, 1987; Harris and Stephens, 1988) is based on the second-moment matrix μ (3.27) and emphasizes image locations that maximize variation of image gradients in a local neighborhood. Such points can be detected as positive spatial maxima of the Harris function

$$H = \det(\mu) - k\,\mathrm{trace}^2(\mu) = \lambda_1(\mu)\lambda_2(\mu) - k(\lambda_1(\mu) + \lambda_2(\mu))^2. \tag{3.31}$$

Both the Laplacian operator ${\nabla^2}_{norm}L$ and the Harris function H are defined in terms of Gaussian derivatives and depend on the affine image transformations. To detect local features independently of affine transformations, Lindeberg and Gårding (1997) and Mikolajczyk and Schmid (2002) proposed to combine feature detection either in terms of ${\nabla^2}_{norm}L$ (3.19) or H (3.31) with the previously described scheme for affine adaptation. The adaptation procedure is initialized with the positions of local features detected for initial values of Σ'. Then, at each iteration the position of features is updated using the recently estimated values of Σ'_i and σ_i. The result of feature detection with affine adaptation is illustrated in Figure 3.3. Other schemes for affine-adapted image descriptors have been proposed by (Ballester and Gonzalez, 1998; Baumberg, 2000; Tuytelaars and Van Gool, 2000).

3.2 Gaussian spatio-temporal scale-space

Temporal image sequences (or video) can be represented by a function $f: \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ over two spatial dimensions (x, y) and one temporal dimension t. The transformation properties of the spatial subspace of f are the same as for static images and, hence, the theory described earlier in Sections 3.1.1-3.1.4 still applies. The temporal dimension, however, has a number of specific properties that must be treated separately.

Temporal data often contains events that last over a finite amount of time. Examples of such events are appearance and disappearance of a person in a room, blinking of an eye, water splash, sunset, etc. As different events can have different extents in time, the notion of temporal scales is meaningful. Concerning the transformations, changes in the temporal scale occur due to the changes in the sampling rate of the camera and the temporal frequency of image patterns (e.g. fast vs. slow walking of a person). Note, that the perspective transformation that effects spatial scales does not change temporal frequencies in image sequences and, hence, has no effect on temporal scales. Moreover, temporal scales of events are generally independent from spatial scales for correspondent image structures. Hence, the spatial and the temporal scales in image sequences have to be treated independently.

Another source of variation in the temporal domain originates from the relative motion between the observed pattern and the camera. For an illustration, consider a spatio-temporal pattern of a walking person (Figure 3.4(a)) recorded either with a stabilized camera (Figure 3.4(b)) or a stationary camera (Figure 3.4(c)). As can be seen, spatio-temporal patterns of legs, shown for x-t-slices of a sequence, are influenced by the relative motion of the camera. The transformation of the temporal domain due to relative camera motion can be compared to the transformation of the spatial domain due to the changes in the relative orientation between the camera and the object.

Finally, another property of the temporal domain originates from the fact that the time dimension is causal, i.e. the information in real-time applications cannot be accessed from the future. This limitation introduces restrictions on the set of allowed operations on image sequences that will be discussed in Section 3.3.

3.2.1 Spatio-temporal scale transformation

When comparing events in two image sequences, the values of temporal scales might not be known in advance. Hence, in order to obtain invariance, it is natural to consider representations of image sequences that are closed under possible spatial and temporal scale transformations of the data.

Different types of spatio-temporal scale-space representation have been considered (Koenderink, 1988; Lindeberg and Fagerström, 1996; Lindeberg, 1997a; Florack, 1997; Lindeberg, 2002; ter Haar Romeny, 2003). One possible approach to obtain closedness for spatial and temporal scalings, is to extend the framework of the Gaussian scale-space into a spatio-temporal domain. For this purpose consider

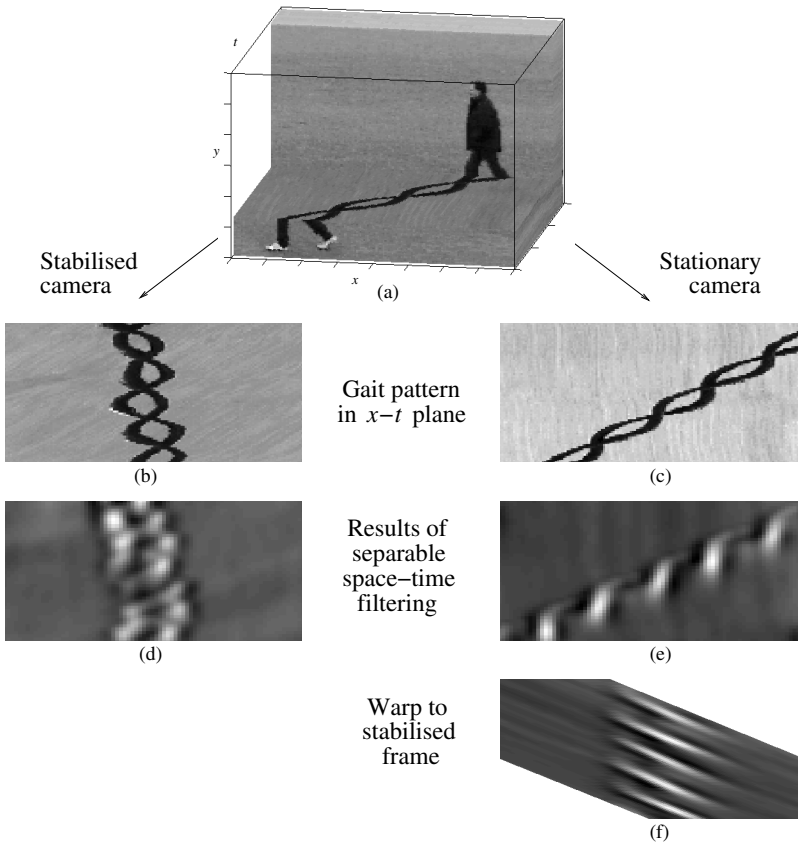


Figure 3.4: Spatio-temporal image of a walking person (a) depends on the relative motion between the person and the camera (b)-(c). If this motion is not taken into account, spatio-temporal filtering (d-e) (here, the second order spatial derivative) results in different responses for different camera motions. Manual stabilization of the pattern in (e) shown in (f) makes the difference more explicit when compared to (d).

the scale-space representation $L(p; \Sigma)$ obtained by the convolution of the spatio-temporal signal $f(p)$, $p = (x, y, t)^T$ with a separable three-dimensional Gaussian kernel

$$g(p; \Sigma) = \frac{1}{\sqrt{(2\pi)^3 \det(\Sigma)}} e^{-(p^T \Sigma^{-1} p)/2}, \quad \Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \tau^2 \end{pmatrix}. \tag{3.32}$$

Here σ^2 and τ^2 denote spatial and temporal scale parameters respectively and allow for independent scale transformations in space and time. Under the scaling

transformation

$$S = \begin{pmatrix} s_1 & 0 & 0 \\ 0 & s_1 & 0 \\ 0 & 0 & s_2 \end{pmatrix} \tag{3.33}$$

the spatio-temporal Gaussian kernel g, the scale-space representation L and its derivatives transform similarly to the spatial domain (see Section 3.1) as

$$\begin{aligned} g(p;\, \Sigma) &= s^2{}_1 s_2 g(Sp;\, S\Sigma S) \\ L(p;\, \Sigma) &= L'(Sp;\, S\Sigma S) \\ L_{x^m y^n t^k}(p;\, \Sigma) &= (s_1^{m+n} s_2^k) L'_{x^m y^n t^k}(Sp;\, S\Sigma S) \end{aligned} \tag{3.34}$$

where $L'(\cdot;\, S\Sigma S) = f'(\cdot) * g(\cdot;\, S\Sigma S)$ and $f'(Sp) = f(p)$. Moreover, scale normalization of spatio-temporal derivatives

$$L_{x^m y^n t^k, norm}(p;\, \Sigma) = \sigma^{m+n} \tau^k L_{x^m y^n t^k}(p;\, \Sigma) \tag{3.35}$$

makes it possible to compare spatio-temporal image structures independently of observation scales. To obtain invariance under scale transformation we will consider a mechanism for automatic spatio-temporal scale selection in Section 5.1.

3.2.2 Galilean transformation

Constant motion $(v_x, v_y)^T$ between the camera and the object effects points p in the image sequence f according to a linear Galilean transformation

$$p' = Gp, \quad G = \begin{pmatrix} 1 & 0 & v_x \\ 0 & 1 & v_y \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.36}$$

From the transformation property of the Gaussian kernel

$$g(p;\, \Sigma) = \det(G) g(Sp;\, G\Sigma G^T) = \det(G) g(p';\, \Sigma') \tag{3.37}$$

it follows that $\Sigma' = G\Sigma G^T$ does not remain diagonal under Galilean transformation and, hence, the separable scale-space L in (3.32) is *not closed* under Galilean transformation. The problem becomes even more apparent from Figures 3.4(d)-(f) where different responses of separable derivative operators are obtained depending on different velocities of the camera. Note, that this situation is similar to the case of affine transformations in space (see Section 3.1.3).

To extend the separable spatio-temporal scale-space to be closed under Galilean transformation, we consider L generated by the convolution of the image sequence f with a Gaussian kernel having an arbitrary covariance matrix $\Sigma \in SPSD(3)$. Similarly to the differentiation in the case of affine transformation (3.25) spatio-temporal derivative operators transform as

$$\begin{pmatrix} \partial_x \\ \partial_y \\ \partial_t \end{pmatrix} = G^T \begin{pmatrix} \partial_{x'} \\ \partial_{y'} \\ \partial_{t'} \end{pmatrix} = \begin{pmatrix} \partial_{x'} \\ \partial_{y'} \\ v_x \partial'_x + v_y \partial'_y + \partial_{t'} \end{pmatrix} \tag{3.38}$$

and the spatio-temporal gradient transforms according to

$$\nabla L(p;\, \Sigma) = G^T \nabla' L'(p';\, \Sigma'), \tag{3.39}$$

where ∇' denotes differentiation with respect to p'.

The result of applying the transformed (also regarded as "velocity-adapted") derivative operators to a spatio-temporal synthetic image with one spatial and one temporal dimension is illustrated in Figure 3.5. As can be seen, depending on the value of v_x used for adaptation, the filtering is able to emphasize either the moving pattern (Figure 3.5(b)) or the stationary background (Figure 3.5(c)).

If we want to interpret events independently of their relative motion to the camera, one approach is to adapt derivative operators *globally* with respect to the velocity of the events in the field of view. This approach is equivalent to the global camera stabilization followed by separable filtering. As illustrated in Figure 3.5(b), the result of filtering using globally adapted receptive fields with $v_x = -1$ indeed enhances the structure of the moving pattern. However, the stationary pattern is suppressed and it follows that global velocity adaptation is not able to handle multiple motions. Moreover, global velocity adaptation is likely to fail if the external velocity estimation is incorrect (Figure 3.5(d)). This emphasizes the need for local velocity adaptation that will be considered in Sections 5.2 and 5.3.

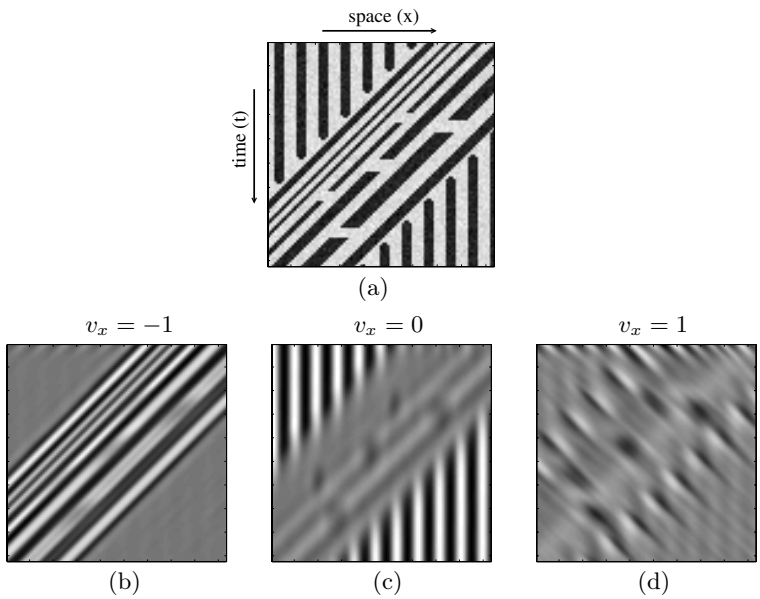


Figure 3.5: (a): Synthetic spatio-temporal pattern. (b)-(d): convolution of (a) with spatio-temporal velocity-adapted second-order derivative operators with $\sigma^2 = 32$, $\tau^2 = 32$ and velocity parameters $v_x = -1, 0, 1$, respectively.

3.3 Time-recursive spatio-temporal scale-space

Real-time vision applications such as automatic car navigation are constrained by the fact that the image data is available only for the past and not for the future. One drawback of the Gaussian temporal scale-space described in the previous section is that the Gaussian kernel has infinite support in both directions and, hence, is not well-suited for the causal temporal data. To address this problem, Koenderink (1988) proposed to transform the time axis to map the present moment to the unreachable infinity and to apply the Gaussian convolution in the transformed domain. Another approach was taken by Lindeberg and Fagerström (1996) who have shown that the only possible way to generate a scale-space that satisfies causality, semi-group and non-creation of local extrema on the discrete time domain is by applying a recursive smoothing operation

$$L^{(k+1)}(t) = \frac{1}{1+\mu}\,(L^{(k)}(t) + \mu L^{(k+1)}(t-1)), \tag{3.40}$$

where k denotes the number of temporal smoothing stages. The corresponding temporal smoothing kernel with coefficients $c_n \geq 0$ obeys temporal causality by only accessing data from the past. Moreover, this kernel is normalized to $\sum_{n=-\infty}^{\infty} c_n = 1$ and has mean value $m = \sum_{n=-\infty}^{\infty} n c_n = \mu$ and variance $\tau^2 = \sum_{n=-\infty}^{\infty} (n-m)^2 c_n = \mu^2 + \mu$. By coupling k such recursive filters (3.40) in cascade, we obtain a filter with mean $m_k = \sum_{i=1}^{k} \mu_i$ and variance $\tau_k^2 = \sum_{i=1}^{k} \mu_i^2 + \mu_i$.

It can be shown that if for a given variance τ^2 we let $\mu_i = \tau^2/K$ become successively smaller by increasing the number of filtering steps K, then the filter kernel approaches the Poisson kernel (Lindeberg, 1997a), which corresponds to the canonical temporal scale-space concept having a continuous scale parameter on a discrete temporal domain. Lindeberg and Fagerström (1996) also show that temporal derivatives of the time-recursive scale-space generated by (3.40) can be computed by finite differences between successive temporal scale levels. An important practical advantage of this is that no other time buffers are necessary for computing temporal scale-space derivatives than the channels L^k of the scale-space representation.

A separable spatio-temporal causal scale-space can be generated by the recursive convolution in the temporal domain and the Gaussian convolution in the spatial domain. The resulting spatio-temporal scale-space will be of four dimensions since no temporal buffering accept the delayed temporal scale levels is necessary for maintaining the scale-space representation. This data set constitutes one time slice of the five-dimensional spatio-temporal representation of the complete history of the visual observer.

Convolution kernels and their derivatives corresponding to a scale-space with one temporal and one spatial dimension are illustrated in Figure 3.6. As can be seen, the increasing value of the temporal smoothing stages K makes the kernels more similar to the corresponding Gaussian kernels on the right computed for the same scale values. An extension of the recursive separable scale-space to the recursive Galilean scale-space has been considered in (Lindeberg, 2002).

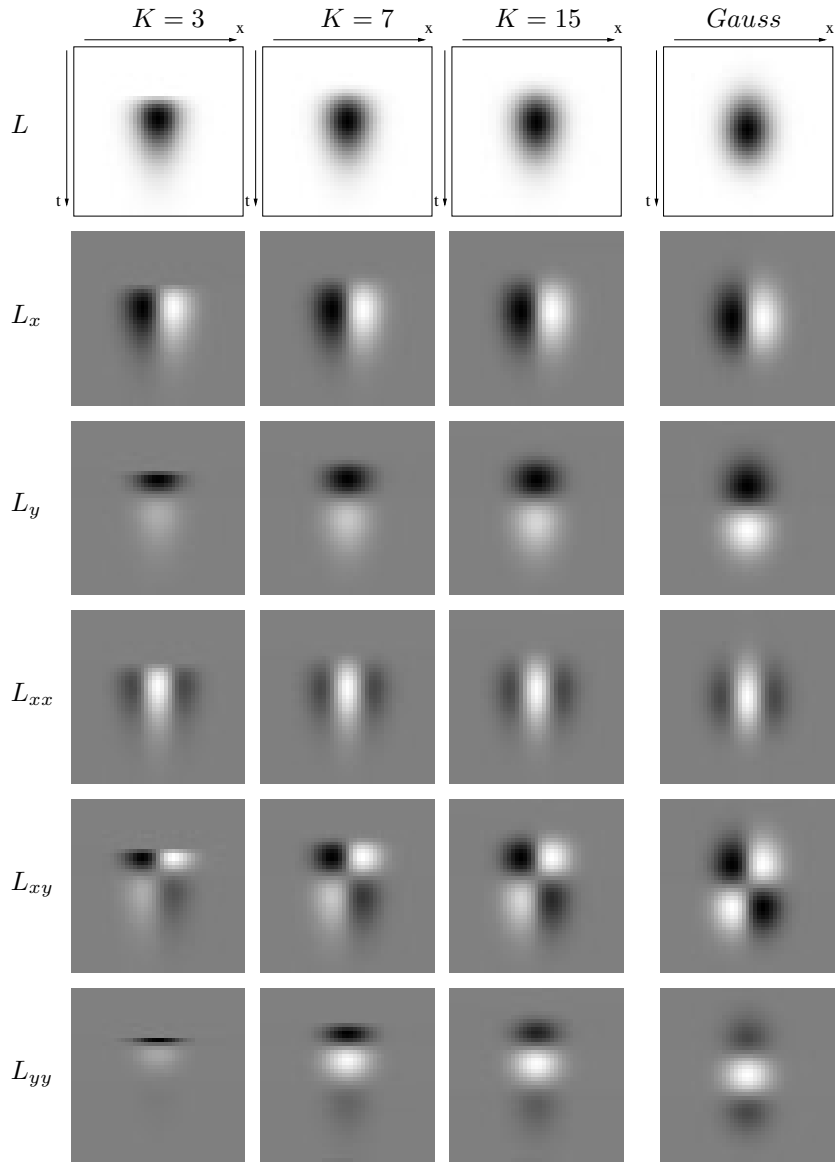


Figure 3.6: Time-recursive convolution kernels with spatial variance $\sigma^2 = 16$ and temporal variance $\tau^2 = 32$. With increasing number of smoothing steps k the kernels and their derivatives will become more similar the corresponding Gaussian kernel and its derivatives.

3.4 Motion estimation

Motion estimation from images concerns with the estimation of a two-dimensional *motion field* originating from the projection of three-dimensional physical motion onto image sequences. The motion field is, however, not directly accessible from images and we can only find its approximation, known as *optical flow* (Gibson, 1950).

Computation of optical flow is usually based on the assumption that changes in images over time are caused by the physical motion in the world and not by other factors such as the motion of a light source. The projection of physical motion results in the displacement of image structures and, hence, the problem of estimating optical flow can be formulated as finding correspondences between points in subsequent frames.

The problem of correspondence has already been addressed in Section 3.1.2 within the context of local feature matching. Hence, one approach to estimate *sparse* optical flow consists of tracking distinctive image features over time. An overview of feature-based matching techniques can be found in (Faugeras, 1993). Beside *sparse* optical flow, many applications such as structure from motion (Horn, 1987) require estimation of *dense* optical flow for each image point. Estimating dense optical flow is known to suffer from a so called "aperture problem" due to the ambiguity of matching in homogeneous image areas and areas with one-dimensional variation of image brightness.

When estimating dense optical flow (v_x, v_y) from an image sequence $f(x, y, t)$, one assumption made by many methods is that all changes in f over time t should be caused by motion. This constraint can be formalized in the Brightness Change Constraint Equation, BCCE (Horn and Schunck, 1981), stating that the total derivative of f with respect to time has to equal zero

$$\frac{df}{dt} = \frac{\partial f}{\partial x} v_x + \frac{\partial f}{\partial y} v_y + \frac{\partial f}{\partial t} = 0. \tag{3.41}$$

By comparing this equation to the transformation property of a temporal derivative operator in (3.38), we can interpret BCCE in terms of compensating the effect of Galilean transformation discussed earlier in Section 3.2. Optical flow (v_x, v_y), however, cannot be computed from (3.41) directly since BCCE constitutes one equation with two unknowns (v_x, v_y). A number of different methods to resolve this problem have been proposed and include differential techniques (Lucas and Kanade, 1981), tensor-based techniques (Bigün and Granlund, 1987; Granlund and Knutsson, 1995), quadrature filter techniques (Adelson and Bergen, 1985) and others. Whereas comprehensive reviews on different methods can be found in (Barron et al., 1994; Jähne et al., 1999), we here describe only one possible alternative that will be of relevance in the rest of this thesis.

3.4.1 Local least squares

Lucas and Kanade (1981) proposed to solve (3.41) by requiring all points in a local neighborhood w (usually a Gaussian) of any point p to have the same velocity (v_x, v_y). Using the notation $\nabla f = (f_x, f_x, f_t)^T$ for a spatio-temporal image gradient and $u = (v_x, v_y, 1)^T$ for velocity, this results in the following least square minimization problem

$$(v_x, v_y) = \text{argmin}_u \int_{\xi \in \mathbb{R}^2} w(p - \xi)[(\nabla f(\xi))^T u]^2 d\xi \tag{3.42}$$

which using least square estimation leads to

$$\int_{\xi \in \mathbb{R}^2} w(p - \xi)[\nabla f(\xi)(\nabla f(\xi))^T] u d\xi = \mu u = 0 \tag{3.43}$$

where μ is a spatio-temporal second-moment matrix

$$\mu = w * \begin{pmatrix} {f_x}^2 & f_x f_y & f_x f_t \\ f_x f_y & {f_y}^2 & f_y f_t \\ f_x f_t & f_y f_t & {f_t}^2 \end{pmatrix} = \begin{pmatrix} \mu_{xx} & \mu_{xy} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix}. \tag{3.44}$$

The solution of (3.43) is obtained as

$$\begin{pmatrix} v_x \\ v_y \end{pmatrix} = A^{-1} b, \qquad A = \begin{pmatrix} \mu_{xx} & \mu_{xy} \\ \mu_{xy} & \mu_{yy} \end{pmatrix}, b = \begin{pmatrix} \mu_{xt} \\ \mu_{yt} \end{pmatrix}. \tag{3.45}$$

3.4.2 Galilean interpretation

The result above can be interpreted in terms of a Galilean transformation G. If we consider transformation of the pattern $f(p) = f'(p')$ for $p' = G^{-1}p$ and a Galilean transformation G defined using the estimated velocities (v_x, v_y), then, using the transformation property of the second-moment matrix $\mu' = G^T \mu G$ (3.28), it can be verified that μ transforms under G according to

$$\mu' = \begin{pmatrix} \mu_{xx} & \mu_{xy} & 0 \\ \mu_{xy} & \mu_{yy} & 0 \\ 0 & 0 & \mu'_{tt} \end{pmatrix}. \tag{3.46}$$

This implies that motion adaptation using velocity estimation according to (3.45) brings a spatio-temporal second moment matrix of a spatio-temporal image pattern into a block-diagonal form. This observation will be important when deriving methods for velocity correction and velocity adaptation in Sections 4.2 and 5.2.

Meanwhile, Figure 3.7 illustrates Local Least Square (LLS) velocity estimation for synthetic spatio-temporal patterns with one spatial and one temporal dimension. Here, all spatio-temporal derivatives were computed using separable Gaussian

derivative operators at fixed spatio-temporal scales. Whereas the ellipses corresponding to the estimated covariance matrix $\mu^{-1} = G^T G$ are shown in Figure 3.7(a), the dense optical flow field according to LLS is shown in 3.7(b). As can be seen, the method works consistently even when the assumption of locally constant velocity is strongly violated as in the case of a lower sequence.

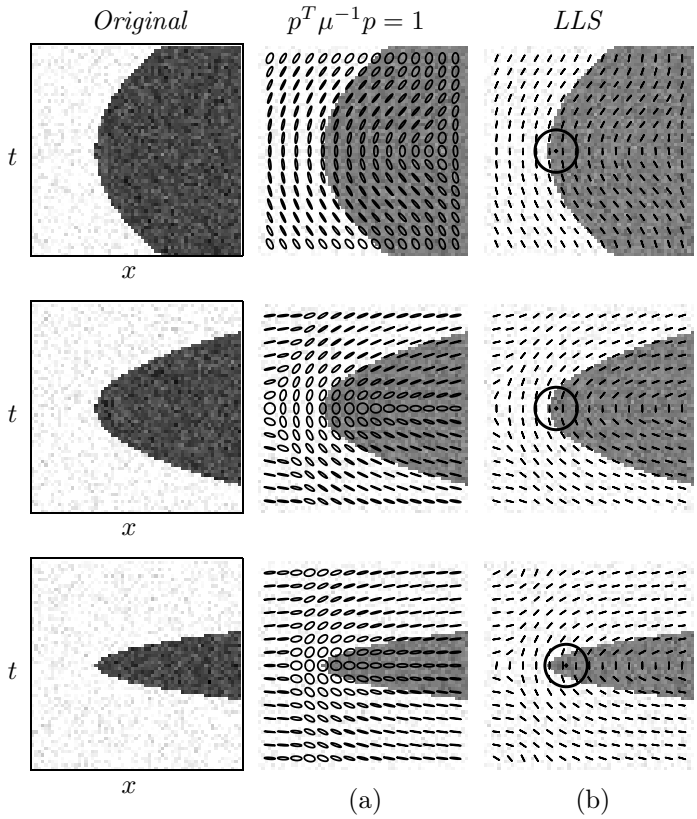


Figure 3.7: Local least square velocity estimation for image patterns with one spatial and one temporal dimension. (a): Ellipses with covariance matrices corresponding to the locally estimated second-moment matrices. (b): Estimated velocities. Note, that chosen velocities are consistent even in the case (lower sequence) when the assumption of locally constant motion is violated

Part II

Contributions

Chapter 4

Local space-time features

Psychological literature accumulated lots of evidence in support of the idea that people tend to segment motion into *action units* or *events* (Newtson et al., 1977; Zacks et al., 2000; Lassiter et al., 2000).

> "... The world presents us with a continuous stream of activity which the mind parses into events. Like objects, they are bounded; they have beginnings, (middles,) and ends. Like objects, they are structured, composed of parts. However, in contrast to objects, events are structured in time..." (Tversky et al., 2002).

Experiments have shown that events are well localized in time and are consistently identified by different people. Moreover, the ability of memorizing activities has shown to be dependent on how fine we subdivide the motion into units (see Section 2.4.1 for more discussion).

In computer vision, the idea of interpreting motion in terms of motion events has been investigated in several works (Brand, 1997; Rui and Anandan, 2000; Zelnik-Manor and Irani, 2001; Yu and Ballard, 2002). Rui and Anandan (2000), for example, demonstrated a method for subdividing video with human activities into meaningful units that have shown to be correlated with the results of manual segmentation obtained in psychological experiments. Concerning applications, segmentation of motion into events provides support for video browsing, video retrieval and video summarization. Moreover, it might also be an efficient approach of representing motion for other visual tasks involving learning and recognition

Whereas the motivation for event detection is sufficiently strong, no reliable and sufficiently general solution to the problem exists. Methods of one type use global measurements and perform well in certain situations but are not suited for scenes with multiple events. In order to localize motion in the scene, methods of another type usually involve spatial segmentation followed by temporal tracking. These methods fully rely on segmentation and tracking which stability in general cannot be assumed in practice (see discussion in Section 2.4).

In this work we argue that detection and localization of events as well as the following interpretation of motion does not necessary require intermediate steps of segmentation and tracking. Instead of solving the correspondence problem over time (tracking) followed by the analysis of temporal trajectories, we here propose to detect events *directly* from spatio-temporal image data. The main idea is to focus on the

distinctive locations in space-time that are robust with respect to detection and discriminative for the purpose of interpretation

Regions in space-time with such properties will be denoted as "local space-time features" or, equivalently, "local motion events". With respect to the detection scheme developed in this work, we will not claim that our features will correspond to the notion of events used in psychological literature. However, we note that both local space-time features and motion units in psychology share similar properties in terms of locality and specific structure of their neighborhoods in space-time.

Following the idea of local space-time features, the rest of this Chapter introduces one possible approach for feature detection. Using the theory of Chapter 3, Sections 4.1 and 4.2 present the mathematical formulation of the detector. In Section 4.3 we consider examples of detected events and discuss their potential advantages and limitations for the task of representing dynamic scenes. Finally, in Section 4.4 we compare the method to the related work in spatial domain and discuss the issue of real-time implementation.

4.1 Detection

The notion of local space-time features defined above can be formalized in several different ways. Given a signal $f(x)$, one generic approach for detecting local features consists of defining an operator $H(f)$ with desired properties at local maxima and then maximizing H over x.

For static images, the Harris operator (3.31) has shown a good performance in terms of localization (Schmid et al., 2000) as well as discriminative power (Mikolajczyk and Schmid, 2003). This operator emphasizes stable, corner-like image structures corresponding to points with high variation of image values within local neighborhoods (see Figure 3.3). Here, we extend this idea into the spatio-temporal domain and search for points that maximize the local variation of image values simultaneously over two spatial dimensions and a temporal dimension.

To maximize the local variation in a spatio-temporal function $f\colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$, consider its scale-space representation $L\colon \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}^2_+ \mapsto \mathbb{R}$ generated by the convolution of f with a separable Gaussian kernel $g(p; \Sigma)$ (3.32). The parameters σ^2 and τ^2 of the covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \tau^2 \end{pmatrix} \tag{4.1}$$

correspond to the spatial and temporal scale parameters respectively and define spatio-temporal extent of the neighborhoods.

Similar to the spatial domain (see equation (3.31)), we consider a spatio-temporal second-moment matrix defined in terms of spatio-temporal gradients and weighted with a Gaussian window function $g(\cdot;\, s\Sigma)$

$$\begin{aligned} \mu(\cdot;\, \Sigma) &= g(\cdot;\, s\Sigma) * \left(\nabla L(\cdot;\, \Sigma)(\nabla L(\cdot;\, \Sigma))^T\right) \\ &= g(\cdot;\, s\Sigma) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}. \end{aligned} \tag{4.2}$$

Here, the spatio-temporal scale-space gradient $\nabla L = (L_x, L_x, L_t)^T$ is defined according to (3.35) and the integration scale parameters $s\Sigma$ are related to the differentiation scales Σ by a constant factor $s > 1$.

The spatio-temporal second-moment matrix μ (also regarded as a structure tensor) has been considered previously (Bigün and Granlund, 1987; Knutsson, 1989; Nagel and Gehrke, 1998) (see (Jähne et al., 1999) for a review). Its interpretation in terms of eigenvalues makes it possible to distinguish image structures with variations over one, two and three dimensions. In particular, neighborhoods with two significant eigenvalues correspond to surface-like structures in space-time and represent either static image points or points with constant motion. On the contrary, three-dimensional variation of f corresponds to image points with non-constant motion. Such points can be detected by maximizing all three eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of μ over space-time.

Among different approaches to find points with high eigenvalues of μ, we here choose to extend the Harris operator (3.31) into the spatio-temporal domain and to consider the following function

$$H = \det(\mu) - k\,\mathrm{trace}^3(\mu) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3. \tag{4.3}$$

To show that positive local maxima of H correspond to points with high values of $\lambda_1, \lambda_2, \lambda_3$ ($\lambda_1 \leq \lambda_2 \leq \lambda_3$), we define the ratios $\alpha = \lambda_2/\lambda_1$ and $\beta = \lambda_3/\lambda_1$ and re-write H as

$$H = \lambda_1^3(\alpha\beta - k(1 + \alpha + \beta)^3). \tag{4.4}$$

From the requirement $H \geq 0$, we get

$$k \leq \alpha\beta/(1 + \alpha + \beta)^3 \tag{4.5}$$

and it follows that for perfectly isotropic image structures ($\alpha = \beta = 1$), k assumes its maximum possible value $k_{max} = 1/27$. For sufficiently large values of $k \leq k_{max}$, positive local maxima of H will correspond to space-time points with similar eigenvalues $\lambda_1, .., \lambda_3$. Consequently, such points indicate locations of image structures with high spatio-temporal variation and can be considered as positions of local spatio-temporal features. As k in (4.3) only controls the local shape of image structures and not their amplitude, the proposed method for local features detection will be invariant with respect to the affine variation of image brightness.

4.2 Velocity correction

Formulation of the interest operator H (4.3) in terms of eigenvalues implies its invariance with respect to 3D rotations of the spatio-temporal image pattern f. Whereas 2D rotations are common in the spatial domain, a 3D rotation in space-time does not correspond to any known physical transformation in the world. On the other hand, as argued in Section 3.2.2, the temporal domain is effected by a Galilean transformation originating from the constant motion between the camera and the observed pattern. Moreover, according to the discussion in Section 3.2.2, the separable scale-space representation L as well as its derivatives are not closed under Galilean transformation of spatio-temporal image patterns. This implies that the second-moment descriptor μ (4.2) as well as the interest operator H (4.3) will be effected by the relative camera motion.

In order to detect local space-time features independently of the camera motion, Section 5.2.1 will present a method for adapting the descriptors μ and H to the values of locally estimated velocities. Although this approach enables feature detection independently of velocity transformation, it involves adaptive filtering and is computationally expensive. In this section we note that the effect of Galilean transformation can be reduced even when using separable filters. We modify the operator H (4.3) to a velocity-corrected operator H_c and show that the new operator depends less on the velocity transformation of the pattern. Moreover, such an operator allows us to impose additional constraints on feature detection that can be useful in practice.

To derive a velocity-corrected feature detector we re-call that the second moment matrix μ(4.2) transforms under a Galilean transformation G^{-1} according to a general rule in (3.28) as

$$\mu' = \begin{pmatrix} \mu'_{xx} & \mu'_{xy} & \mu'_{xt} \\ \mu'_{xy} & \mu'_{yy} & \mu'_{yt} \\ \mu'_{xt} & \mu'_{yt} & \mu'_{tt} \end{pmatrix} = G^T \begin{pmatrix} \mu_{xx} & \mu_{xy} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix} G. \tag{4.6}$$

As argued in Section 3.4.2, the use of velocity estimates $(\tilde{v}_x, \tilde{v}_y)$

$$\begin{pmatrix} \tilde{v}_x \\ \tilde{v}_y \end{pmatrix} = \begin{pmatrix} \mu_{xx} & \mu_{xy} \\ \mu_{xy} & \mu_{yy} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{xt} \\ \mu_{yt} \end{pmatrix}, \tag{4.7}$$

in the Galilean transformation G brings μ' into a block-diagonal form according to

$$\mu' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \tilde{v}_x & \tilde{v}_y & 1 \end{pmatrix} \begin{pmatrix} \mu_{xx} & \mu_{xy} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix} \begin{pmatrix} 1 & 0 & \tilde{v}_x \\ 0 & 1 & \tilde{v}_y \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \mu_{xx} & \mu_{xy} & 0 \\ \mu_{xy} & \mu_{yy} & 0 \\ 0 & 0 & \mu'_{tt} \end{pmatrix} \tag{4.8}$$

The descriptor μ' in (4.6) can be regarded as velocity-corrected. Indeed, if estimating velocity from μ' according to (4.7), the resulting velocity values will be equal zero and it seems that μ' is invariant with respect to velocity transformations. The

truth is, however, that the velocity estimates in (4.7) depend on the shape of used filter kernels that in turn are influenced by the velocity transformation. Hence, the velocity estimates will only be correct if using velocity-adapted filter kernels with covariance matrix $\Sigma' = G\Sigma G^T$ (3.37). This situation is similar to the case of affine transformation (see Section 3.1.4) where the affine adaptation of filter kernels is required to achieve the invariance. Similarly, here the adaptation of derivative operators with respect to velocity (3.39) is required in order to achieve full invariance with respect to the relative camera motion.

When using separable filtering, μ' can be regarded as an approximation of the truly velocity-invariant second-moment descriptor. To achieve velocity correction of H, we can therefore use μ' instead of μ in the definition of H in (4.3). A slightly different approach can be taken if we note that the component μ'_{tt} in (4.6) encodes all information about the temporal variation in f. Moreover the upper-left part of μ'

$$\mu'_{spat} = \begin{pmatrix} \mu_{xx} & \mu_{xy} \\ \mu_{xy} & \mu_{yy} \end{pmatrix} \tag{4.9}$$

encodes all the spatial variations in f and it follows, that we now can treat the spatial and the temporal variations in f separately. Such a separation is achieved in the following velocity-corrected operator

$$\begin{aligned} H_c &= \det(\mu') - (k_1 \operatorname{trace}(\mu'_{spat}) + k_2 \mu'_{tt})^3 \\ &= \lambda'_1 \lambda'_2 \mu'_{tt} - (k_1(\lambda'_1 + \lambda'_2) + k_2 \mu'_{tt})^3, \end{aligned} \tag{4.10}$$

where λ'_1, λ'_2 are the eigenvalues of μ'_{spat} and the explicit expression for μ'_{tt} is

$$\mu'_{tt} = \frac{2\mu_{xt}\mu_{xy}\mu_{yt} - \mu_{xx}{\mu^2}_{yt} - \mu_{yy}{\mu^2}_{xt} - \mu_{tt}{\mu^2}_{xy} + \mu_{xx}\mu_{yy}\mu_{tt}}{\mu_{xx}\mu_{yy} - {\mu^2}_{xy}}. \tag{4.11}$$

If $k_1 = k_2 = k^{1/3}$, the operator H_c is equivalent to H (4.3) apart from the use of velocity-corrected descriptor μ'. When choosing $k_1 < k_2$, however, the positive local maxima of H will be biased to points with high *spatial* variations. On the contrary, $k_1 > k_2$ will enforce bias to points with high *temporal* variations.

4.3 Examples of detected features

In this section, we illustrate the results of detecting local space-time events in synthetic and real image sequences. For the detection, we use the interest operator H defined in (4.3) with parameter values $k = 0.005$ and $s = 2$. Synthetic sequences are considered in Figures 4.1 and 4.2 and represent typical simplified cases of local motion events: splitting, merging, collisions and spontaneous changes of velocity. For the clarity of presentation, we show image sequences as three-dimensional plots where original data is represented by a space-time threshold surface. The detected features are illustrated by ellipsoids with positions corresponding to the locations

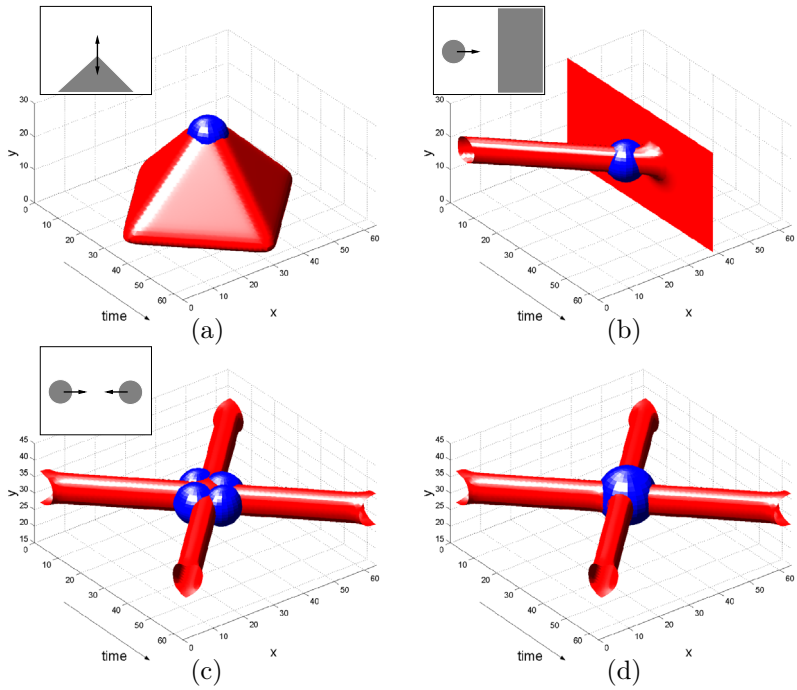


Figure 4.1: Results of detecting space-time features for synthetic image sequences. (a): a moving corner; (b) a merge of a ball and a wall; (c),(d): collision of two balls with space-time features detected for different values of scale parameters $\sigma_l^2 = 8$, $\tau_l^2 = 8$ in (c) and $\sigma_l^2 = 16$, $\tau_l^2 = 16$ in (d).

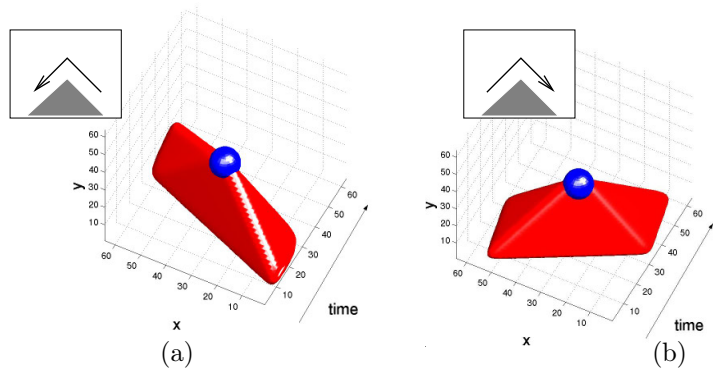
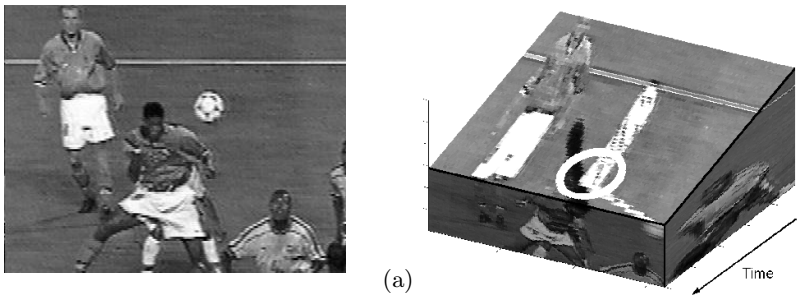


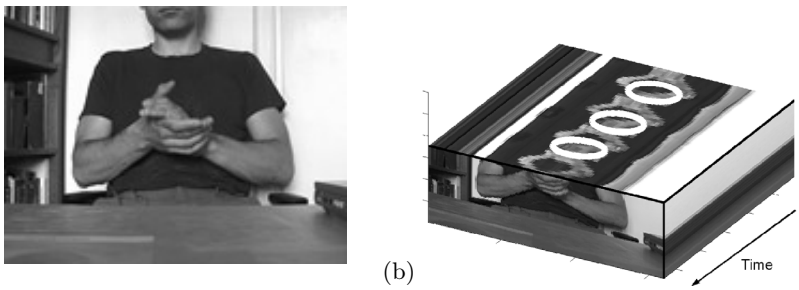
Figure 4.2: Detection of local space-time features for the event in Figure 4.1(a) influenced by the velocity transformation.

Example event: velocity discontinuity



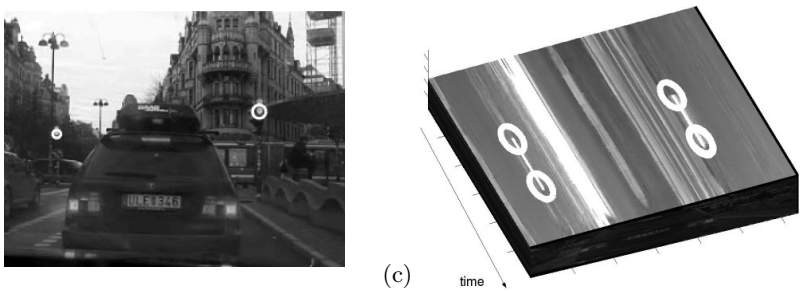
(a)

Example events: splits and unifications



(b)

Example events: appearance and disappearance



(c)

Figure 4.3: Result of detecting the most prominent (with respect to the magnitude of H) space-time events in (a): A football sequence with a player heading the ball; (b): A hand clapping sequence and (c): Change of traffic light. From the temporal slices of space-time volumes shown on the right, it is evident that the detected events correspond to image neighborhoods with high spatio-temporal variations.

of detected space-time features and with semi-axes proportional to the scales (σ, τ) that have been used to compute a second-moment descriptor.

Figure 4.1(a) illustrates a sequence with a moving corner. The event is detected at the moment in time when the motion of the corner changes its direction. This type of event occurs frequently in real image sequences with articulated motion. Note that image points with constant velocity do not give rise to the response of the detector. Other typical types of detected events correspond to splits and unifications of image structures. In Figure 4.1(b), the feature is detected at the moment and the position corresponding to the collision of a sphere and a surface. Similarly, space-time features are detected at the moment of collision and splitting of two balls as shown in Figure 4.1(c)-(d). Note, that different types of events are detected depending on the scale of observation used to compute a second-moment descriptor.

Figures 4.2(a)-(b) illustrate image sequences similar to the one in Figure 4.1(a) but effected by the relative motion between the camera and the scene. From the plots it becomes apparent that the velocity (or Galilean) transformation has a skewing effect on the spatio-temporal image pattern. Here, the maxima of H could be detected disregarding the changes in relative velocities. However, since the function H is effected by velocity transformation, the velocity-invariant detection of space-time features cannot be expected in general.

Figure 4.3 presents detection of space-time features for real image sequences. Here, the complex background makes the visualization of results more difficult, however, the most significant features can be illustrated using slices in a three-dimensional data volume. As can be seen from the slices for all three sequences, natural events such as heading a ball, clapping hands and turning on and off the traffic lights correspond to non-trivial changes in the space-time image pattern. Such changes are successfully detected by our algorithm in spite of other image structures and motion that are present in the scenes.

4.4 Discussion

When comparing the advocated approach to other methods, tracking of local spatial features (e.g. blobs, see Section 3.1) followed by the analysis of obtained trajectories could be sufficient to obtain a similar result as in the case a football sequence in Figure 4.3(a). In the the case of hand clapping (Figure 4.3(b)), the tracking would be more challenging since the appearance of hands over time undergoes significant variations. Finally, in the case of traffic lights(Figure 4.3(c)), tracking is obviously not the right method for detecting appearance and disappearance of image structures. As an alternative method here, one could think of spatial segmentation based on temporal image differencing. However, this method would give response to other motion in the scene, for example to the bus driving in the background.

The proposed method for event detection is greatly inspired by the development and applications of local features in the spatial domain (Schmid and Mohr, 1997;

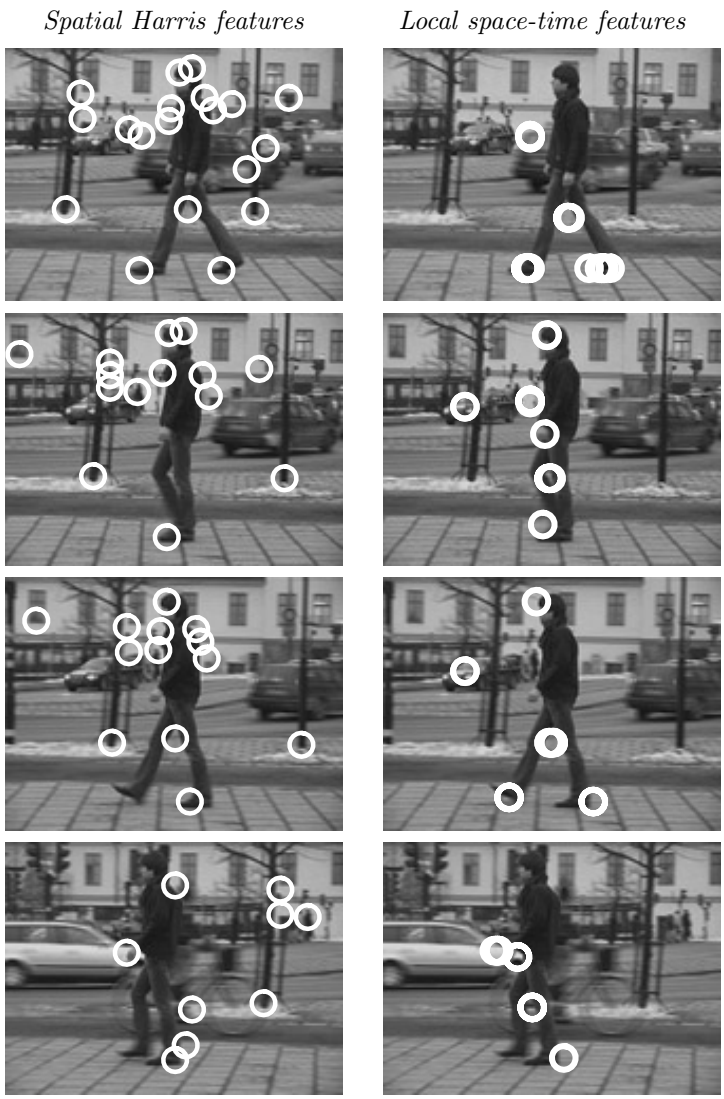


Figure 4.4: Comparison of local features detected as (left): Maxima of the spatial Harris operator (3.31); (right): Maxima of the space-time interest operator H (4.3). For a fair comparison approximately the same number of strongest features are shown for each method. As can be seen, local space-time features are more selective than the spatial ones and emphasize image structures at the moments of non-constant motion. Such points are typical for articulated motion patterns of the legs. Other points are detected at moments and positions of local occlusions.

Lowe, 1999; Weber et al., 2000; Mikolajczyk and Schmid, 2002; Sivic and Zisserman, 2003). The comparison of detected local features in space (here, Harris interest points (Förstner and Gülch, 1987; Harris and Stephens, 1988)) to the detection of local space-time events is presented in Figure 4.4. As can be seen, the space-time approach is more selective for motion and disregards most of the structures in the complex and non-stationary background. Hence, it has clear advantages over pure spatial features if patterns with complex motion are of interest.

The detected local space-time features also bear similarities to motion discontinuities (Zetzsche and Barth, 1991; Niyogi, 1995; Granlund and Knutsson, 1995). Operator H (4.3) is closely related to the curvature in space-time and its maxima correspond to the points violating the constant velocity model (Barth et al., 2002).

Based on the few presented examples, we summarize that the proposed method has a potential of detecting relevant events in real image sequences. Concerning the implementation, space-time filtering is still rather slow operation on modern computers. The constantly growing computing power, however, should make the presented approach suitable for real-time applications within a few years. Another important issue within this context concerns the causality of the temporal domain that should be taken into account in on-line scenario. As discussed in Section 3.3, Gaussian filtering is not well suited for on-line implementations due to the infinite support of the Gaussian kernel in all directions. To address this problem, Gaussian filters could be substituted by the recursive temporal filters discussed in Section 3.3. As such filters can be seen as an approximation to Gaussian filters (see Figure 3.5), the approach presented here could be implemented using recursive filters and their derivatives. We have made such an implementation and obtained sufficiently similar local features compared to the results of Gaussian implementation. A more extensive experimental study of this issue, however, should be made in the future.

Chapter 5

Scale and velocity adaptation

Recognition of objects and events requires comparison of images and image sequences influenced by transformations with unknown parameter values. To be able to compare image data in such conditions, a first step is to construct image representations that are closed under a relevant set of transformations. Given such a representation, a pair of images or video sequences could then be in principle matched by trying out all possible values of transformation parameters. Such a search procedure, however, would be highly inefficient since the number of possible parameter combinations is exponential in the number of parameters.

One of the main challenges when constructing a computer vision system, is to reduce the search space of the matching problem without compromising the final performance of recognition.[1] When comparing image sequences, local event detection as developed in the previous chapter can be seen as an approach to reduce the search space over all possible translations in space-time to the search over positions of space-time features only. As discussed in Chapters 3 and 4, besides translations, temporal image sequences are also influenced by scaling and velocity transformations. Hence, unless exhaustive search over all possible scalings and velocities can be afforded, it is desirable to estimate characteristic values of scales and velocities in image sequences and to use them when matching one image sequence to another. The estimation procedure should, of course, be invariant with respect to transformations of the data. Hence, if image sequences f_1 and f_2 are related by a linear transformation T: $f_1(p) = f_2(Tp)$, then the estimated values of transformations T_1 and T_2 in both sequences respectively should be related as $T_1 = TT_2$.

In this chapter we develop methods for estimating scales and velocities in image sequences locally using information in spatio-temporal neighborhoods. Section 5.1 presents an approach for simultaneous estimation of spatial and temporal scales as well as an iterative scheme for adapting local motion events to such scales. Section 5.1 extends this approach to estimation of local velocities and simultaneous

[1] The search space frequently depends on the data and, hence, can be reduced if such dependencies are taken into account.

adaptation of events with respect to scales and velocities. Finally, in Section 5.3 we demonstrate that velocity adaptation can be applied not only to a sparse set of points in space-time but to all points in the sequence in order to obtain dense, velocity-invariant descriptors.

5.1 Spatio-temporal scale selection

Responses of spatio-temporal Gaussian derivatives depend, in general, on the values of the scale parameters (σ^2, τ^2) as well as on spatio-temporal scale transformations of image sequences. One example of this dependency has already been illustrated in Figures 4.1(c),(d), where the result of event detection depends on the choice of scale parameters (σ^2, τ^2). This section presents an approach for eliminating scale-dependency of space-time image descriptors as well as for detecting local motion events independently of scale transformations of the data.

5.1.1 Scale selection in space-time

During recent years, the problem of automatic scale selection has been addressed in several different ways based on the maximization of normalized derivative expressions over scale (Lindeberg, 1993; Lindeberg, 1994; Lindeberg, 1998b), or the behavior of entropy measures or error measures over scales, e.g. (Jägersand, 1995; Elder and Zucker, 1998; Hadjidemetriou et al., 2002). To estimate the spatio-temporal extent of an event in space-time, we follow works on local scale selection proposed in the spatial domain by Lindeberg (1998b) as well as in the temporal domain (Lindeberg, 1997b). The idea is to define a differential operator that assumes simultaneous extrema over such spatial and temporal scales that correspond to spatio-temporal extents of considered events in space-time.

For the purpose of analysis, consider a prototype event represented by a spatio-temporal Gaussian blob

$$f(x, y, t; \Sigma_0) = \frac{1}{\sqrt{(2\pi)^3 \det(\Sigma_0)}} e^{-(p^T \Sigma_0^{-1} p)/2}, \quad \Sigma_0 = \begin{pmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_0^2 & 0 \\ 0 & 0 & \tau_0^2 \end{pmatrix}. \quad (5.1)$$

with spatial variance σ_0^2 and temporal variance τ_0^2 (see Figure 5.1(a)). Using the semi-group property of the Gaussian kernel (see Section 3.1), it follows that the separable scale-space representation of f is

$$L(\cdot; \Sigma) = g(\cdot; \Sigma) * f(\cdot; \Sigma_0) = g(\cdot; \Sigma_0 + \Sigma) \quad (5.2)$$

where Σ is defined by (4.1). To recover the spatio-temporal extent (σ_0, τ_0) of f, we consider second-order derivatives of L normalized by the scale parameters as follows

$$L_{xx,norm} = \sigma^{2a} \tau^{2b} L_{xx}, \quad L_{yy,norm} = \sigma^{2a} \tau^{2b} L_{yy}, \quad L_{tt,norm} = \sigma^{2c} \tau^{2d} L_{tt}. \quad (5.3)$$

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

The idea of scale selection we follow here is to determine the parameters a, b, c, d such that $L_{xx,norm}$, $L_{yy,norm}$ and $L_{tt,norm}$ assume extrema at scales $\tilde{\sigma}^2 = \sigma_0^2$ and $\tilde{\tau}^2 = \tau_0^2$. To find such extrema, we differentiate the expressions in (5.3) with respect to σ^2 and τ^2. For the spatial derivatives at the center of the blob we obtain the following expressions

$$\frac{\partial}{\partial \sigma^2}\Big[L_{xx,norm}(0,0,0;\,\Sigma)\Big] = -\frac{a\sigma^2 - 2\sigma^2 + a\sigma_0^2}{\sqrt{(2\pi)^3(\sigma_0^2+\sigma^2)^6(\tau_0^2+\tau^2)}}\,\sigma^{2(a-1)}\tau^{2b} \tag{5.4}$$

$$\frac{\partial}{\partial \tau^2}\Big[L_{xx,norm}(0,0,0;\,\Sigma)\Big] = -\frac{2b\tau_0^2 + 2b\tau^2 - \tau^2}{\sqrt{2^5\pi^3(\sigma_0^2+\sigma^2)^4(\tau_0^2+\tau^2)^3}}\,\tau^{2(b-1)}\sigma^{2a}. \tag{5.5}$$

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

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

which after substitutions $\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$ lead to $a = 1$ and $b = 1/4$.

Similarly, differentiating the second-order temporal derivative

$$\frac{\partial}{\partial \sigma^2}\Big[L_{tt,norm}(0,0,0;\,\Sigma)\Big] = -\frac{c\sigma^2 - \sigma^2 + c\sigma_0^2}{\sqrt{(2\pi)^3(\sigma_0^2+\sigma^2)^4(\tau_0^2+\tau^2)^3}}\,\sigma^{2(c-1)}\tau^{2d} \tag{5.6}$$

$$\frac{\partial}{\partial \tau^2}\Big[L_{tt,norm}(0,0,0;\,\Sigma)\Big] = -\frac{2d\tau_0^2 + 2d\tau^2 - 3\tau^2}{\sqrt{2^5\pi^3(\sigma_0^2+\sigma^2)^2(\tau_0^2+\tau^2)^5}}\,\tau^{2(d-1)}\sigma^{2c} \tag{5.7}$$

leads to the expressions

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

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

The normalization of partial derivatives in (5.3) guarantees that these derivative expression assume local extrema over scales at the center of the blob f for $\sigma = \sigma_0$ and $\tau = \tau_0$. From the sum of these partial derivatives, we define a normalized spatio-temporal Laplace operator according to

$$\begin{aligned} \nabla^2_{norm} L &= L_{xx,norm} + L_{yy,norm} + L_{tt,norm} \\ &= \sigma^2\tau^{1/2}(L_{xx} + L_{yy}) + \sigma\tau^{3/2}L_{tt}. \end{aligned} \tag{5.8}$$

Figures 5.1(b)-(c) show the evolution of the first-order derivatives of $\nabla^2_{norm} L$ with respect to the scale parameters when evaluated at the center of a spatio-temporal blob with spatial variance $\sigma_0^2 = 4$ and temporal variance $\tau_0^2 = 16$. The zero-crossings of the curves verify that $\nabla^2_{norm} L$ assumes extrema values at the scales

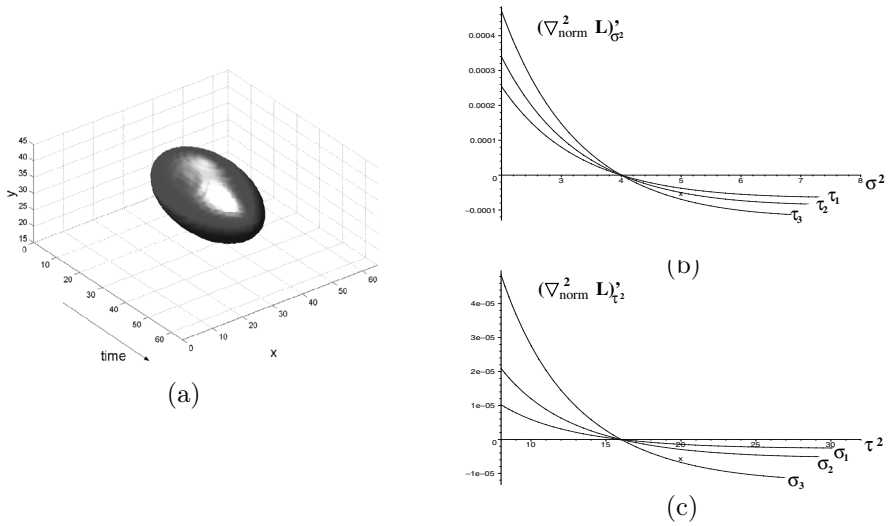


Figure 5.1: (a): A spatio-temporal Gaussian blob with spatial variance $\sigma_0^2 = 4$ and temporal variance $\tau_0^2 = 16$; (b)-(c): derivatives of $\nabla^2_{norm} L$ with respect to scales. Zero-crossings of $(\nabla^2_{norm} L)'_{\sigma^2}$ and $(\nabla^2_{norm} L)'_{\tau^2}$ indicate extrema of $\nabla^2_{norm} L$ at scales corresponding to the spatial and the temporal extent of the blob.

$\sigma^2 = \sigma_0^2$ and $\tau^2 = \tau_0^2$. Hence, the spatio-temporal extent of the Gaussian blob can be estimated by finding the extrema of $\nabla^2_{norm} L$ over spatial and temporal scales.

In the spatial domain, a similar derivation has led to the normalized Laplacian operator (3.19) which has shown a good generalization when estimating scales of image structures in real images (see Figure 3.2). In the next section, we follow this approach and use the normalized spatio-temporal Laplace operator $\nabla^2_{norm} L$ (5.8) for estimating scales of local motion events in image sequences.

5.1.2 Scale-adapted local space-time features

Local scale estimation has been successfully applied in the spatial domain by (Lindeberg, 1998b; Almansa and Lindeberg, 2000; Chomat, de Verdiere, Hall and Crowley, 2000) and others. In particular, Mikolajczyk and Schmid (2001) combined the Harris interest point operator with the normalized Laplace operator and derived a scale-invariant Harris-Laplace interest point detector (see Section 3.1.4).

Here, we extend this idea and detect local space-time features that are simultaneous maxima of the spatio-temporal interest operator H in (4.3) or (4.10) over space-time (x, y, t) as well as extrema of the normalized spatio-temporal Laplace operator $\nabla^2_{norm} L$ (5.8) over scales (σ^2, τ^2). One way of detecting such points is to

compute space-time maxima of H for each spatio-temporal scale level and then to select points that maximize $(\nabla^2_{norm} L)^2$ at the corresponding scale. This approach, however, requires dense sampling over the scale parameters and is therefore computationally expensive.

An alternative we follow here, is to detect local features for a set of sparsely distributed scale values and then to track these points in the spatio-temporal scale-time-space toward the extrema of $\nabla^2_{norm} L$. We do this by iteratively updating the scales and the position of each feature by (i) selecting the spatio-temporal scales σ_{i+1} and τ_{i+1} that maximize $(\nabla^2_{norm} L)^2$ over the scale neighborhood $(\sigma_i - \delta\sigma, \sigma_i + \delta\sigma)$, $(\tau_i - \delta\tau, \tau_i + \delta\tau)$ and (ii) re-detecting the space-time location of the feature at a new scale. Thus, instead of performing a simultaneous maximization of H and $\nabla^2_{norm} L$ over five dimensions $(x, y, t, \sigma^2, \tau^2)$, we implement the detection of local maxima by splitting the space-time dimensions (x, y, t) and scale dimensions (σ^2, τ^2) and iteratively optimizing over the subspaces until the convergence has been reached. The corresponding algorithm is presented in Figure 5.2.

---

1. Detect local space-time features $p_j = (x_j, y_j, t_j, \sigma_j^2, \tau_j^2)$, $j = 1..N$ as positive maxima of H in (4.3) or (4.10) over space-time and at sparsely selected combinations of initial spatial scales $(\sigma_1^2, .., \sigma_n^2)$ and temporal scales $(\tau_1^2, .., \tau_m^2)$.

2. **for** each local space-time feature p_j **do**

3. Compute $\nabla^2_{norm} L$ at the position (x_j, y_j, t_j) and for all combinations of scales $(\tilde{\sigma}_j^2, \tilde{\tau}_j^2)$ where $\tilde{\sigma}_j^2 = s 2^k \sigma_j^2$, $\tilde{\tau}_j^2 = s 2^k \tau_j^2$, for $k = -0.25, 0, 0.25$ and with s corresponding to a constant factor defining the relation between integration and differentiation scales in (4.2).

4. Choose the combination of scales $(\tilde{\sigma}^2_{j,max}, \tilde{\tau}^2_{j,max})$ maximizing $(\nabla^2_{norm} L)^2$

5. **if** $\tilde{\sigma}_j^2 \neq s\sigma_j^2$ or $\tilde{\tau}_j^2 \neq s\tau_j^2$
Re-detect feature p_j for a scale combination $(s^{-1}\tilde{\sigma}^2_{j,max}, s^{-1}\tilde{\tau}^2_{j,max})$ such that the new position of a feature $(\tilde{x}_j, \tilde{y}_j, \tilde{t}_j)$ is maximally close to the old position (x_j, y_j, t_j); set $p_j := (\tilde{x}_j, \tilde{y}_j, \tilde{t}_j, s^{-1}\tilde{\sigma}^2_{j,max}, s^{-1}\tilde{\tau}^2_{j,max})$;
goto 3.

6. **end**

Figure 5.2: An outline for the algorithm for scale adaption of local space-time features.

5.1.3 Experiments

This section presents results of detecting scale-adapted motion events according to the algorithm presented in the previous section. For the experiments we use synthetic and real image sequences with spatial resolution 160×120 pixels and temporal sampling frequency 25Hz (for real image sequences). When initializing the scale adaptation algorithm in Figure 5.2, we detect initial features (step 1) at combinations of spatial scales $\sigma^2 = [2, 4, 8]$ and temporal scales $\tau^2 = [2, 4, 8]$. The relation between integration and differentiation scales is defined by the constant factor $s = 2$.

Figure 5.3 illustrates a synthetic spatio-temporal pattern with oscillations over space and time. The pattern is defined by the function $f(x, y, t) = \mathrm{sgn}(y - \sin(x^4)\sin(t^4))$ and contains structures with different extents over space-time. The result of event detection without scale adaptation according to the original approach in Chapter 4 is illustrated in Figures 5.3(a)-(d). As can be seen, depending on the scales of observation, the method emphasizes different space-time structures (only the most significant features are shown here). In comparison, as illustrated in Figure 5.2(e), the algorithm for scale-adaptive feature detection described in the previous section converges to features whose scales are in correspondence with the spatio-temporal extents of the underlying image structures. This result visually confirms the invariance of the presented approach with respect to changes in spatial and temporal scales of the image pattern.

When applying the method to real image sequences, the example in Figure 5.4 illustrates the result of event detection with scale adaptation for an image sequence of a walking person. From the space-time plot in Figure 5.4(a), we can observe how the selected spatial and temporal scales of the detected features roughly match the spatio-temporal extents of the corresponding structures in the leg pattern. Moreover, given the repeating structure of the gait, we can confirm that the method consistently detects the same type of features for different cycles of the gait.

To illustrate the invariance of the approach with respect to changes in spatial scale, Figure 5.5 shows results obtained for a sequence recorded with a zooming camera. As can be seen, the method for scale-adapted detection of space-time features consistently detects similar local events independently of the size of the person in the image. Moreover, the estimated spatial size of features is consistent with the spatial size of corresponding image structures.

The last example illustrates the performance of the method for variations in the temporal scale. Figure 5.6 shows a person making hand-waving gestures with a high frequency on the left and a low frequency on the right. Distinct features are detected at positions in space-time where the palm of the hand changes its direction of motion. Whereas the spatial scales of the detected features remain constant, the adapted temporal scales depend on the frequency of the waving pattern. High frequency of the hand movements result in short events and give rise to features with small temporal extent (see Figure 5.6(a)). Hand motions with low frequency result in features with long temporal extent as shown in Figure 5.6(b).

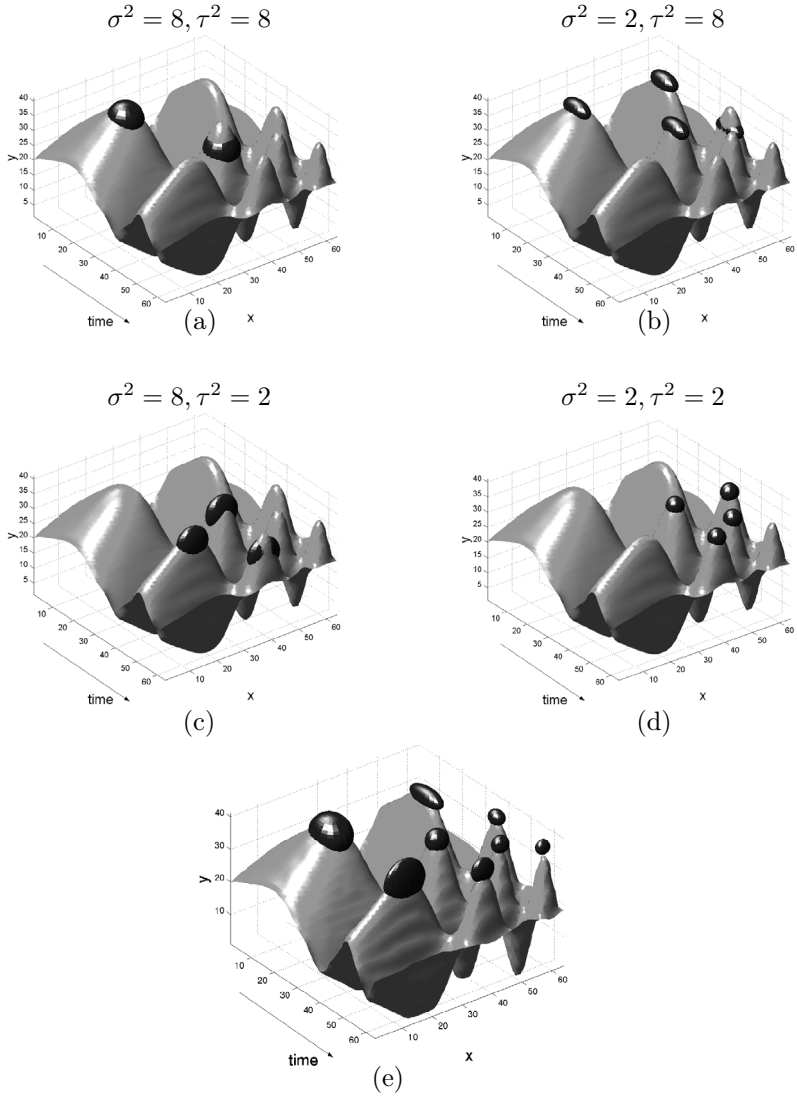


Figure 5.3: Results of detecting local space-time features with and without scale adaptation for a synthetic sequence $f(x, y, t) = \mathrm{sgn}(y - \sin(x^4)\sin(t^4))$ illustrated by a threshold surface. (a)-(d): Local features without scale adaptation computed for different values of the spatial and temporal scale parameters. (e): Local features with scale adaptation detected according to the algorithm in Figure 5.2. The features are illustrated as ellipsoids with the length of the semi-axes proportional to the values of selected scales (σ, τ). Observe that the scales of adapted features match with the spatio-temporal extents of the corresponding image structures.

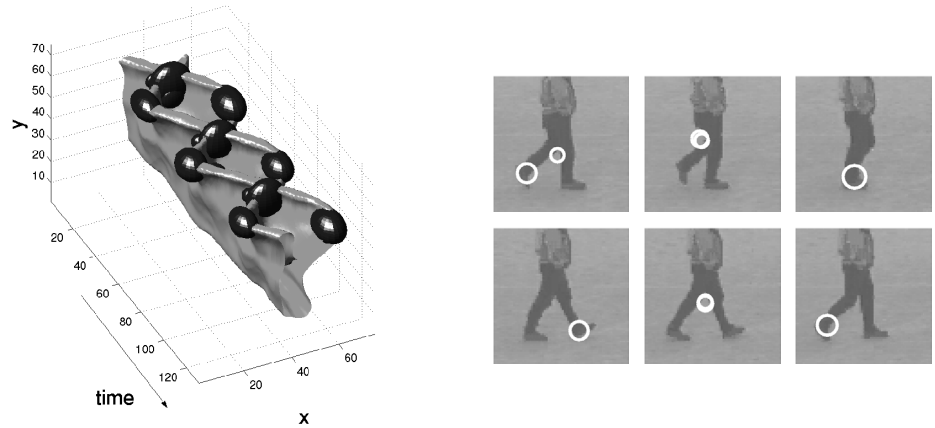


Figure 5.4: Results of detecting scale-adapted local space-time features for a walking person. (left): The pattern of legs is illustrated by a three-dimensional threshold surface (up-side-down). The detected features are illustrated by the ellipsoids with the length of the semi-axes corresponding to the estimated spatio-temporal extent of underlying image structures. (right): Local space-time features overlaid on single frames of the original sequence.

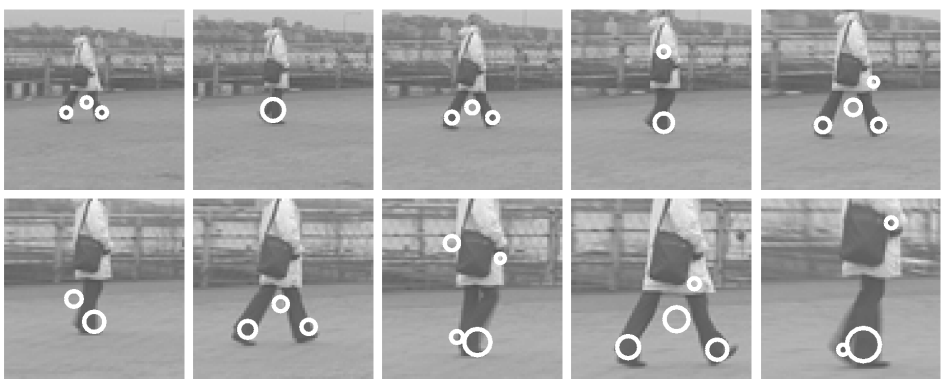


Figure 5.5: Results of detecting scale-adapted local space-time features in a zoom-in sequence with a walking person. The spatial scale of the detected features (proportional to the size of circles) matches the increasing spatial extent of the image structures and verifies the invariance of local features with respect to changes in spatial scale.

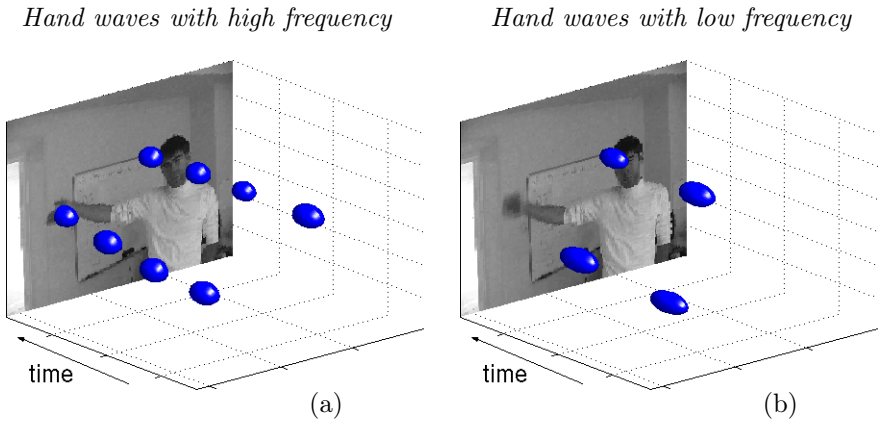


Figure 5.6: Result of detecting local space-time features for a sequence with waving hand gestures. (a): Feature detection for hand movements with high frequency. (b): Feature detection for hand movements with low frequency (only the strongest features are shown here).

5.2 Velocity adaptation

Constant motion between the camera and the observed pattern effects the spatio-temporal image data by a Galilean transformation. This fact has already been discussed in Sections 3.2.2 and 4.2. In particular, it has been shown that a separable spatio-temporal scale-space can be extended to a non-separable Galilean scale-space (3.37) in order to obtain representations that are closed under velocity transformations.

In this section we develop a method for detecting local space-time features independently of Galilean transformations. For this purpose we re-call that under a Galilean transformation G^{-1} the second-moment descriptor μ (4.2) transforms according to a general rule in (3.28) as $\mu'(p'; \Sigma') = G^T \mu(p; \Sigma) G$ where $p' = G^{-1}p$ and $\Sigma' = G^{-1}\Sigma G^{-T}$.

Assume now that μ' is computed using a diagonal covariance matrix Σ' and has a block-diagonal form with elements $\mu'_{xt} = 0$ and $\mu'_{yt} = 0$.[2] The Galilean transformation of μ' with respect to G defined by velocities (v_x, v_y) (3.36) can be

[2] Such a block-diagonal second moment descriptor μ' will correspond to space-time image structures with stationary points. Note, however, that the block-diagonal form of μ' will also correspond to other "stabilized" spatio-temporal image structures whose motion cannot be described by a constant velocity model. Such structures typically correspond to local motion events introduced in the last Chapter. See also space-time image structures marked with circles in Figure 3.7(b).

written as

$$\mu = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -v_x & -v_y & 1 \end{pmatrix} \begin{pmatrix} \mu'_{xx} & \mu'_{xy} & 0 \\ \mu'_{xy} & \mu'_{yy} & 0 \\ 0 & 0 & \mu'_{tt} \end{pmatrix} \begin{pmatrix} 1 & 0 & -v_x \\ 0 & 1 & -v_y \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \mu_{xx} & \mu_{xy} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix} \tag{5.9}$$

After expansion it follows that

$$\begin{pmatrix} \mu_{xt} \\ \mu_{yt} \end{pmatrix} = \begin{pmatrix} \mu'_{xx} & \mu'_{xy} \\ \mu'_{xy} & \mu'_{yy} \end{pmatrix} \begin{pmatrix} v_x \\ v_y \end{pmatrix} \tag{5.10}$$

and

$$\begin{pmatrix} \mu_{xx} & \mu_{xy} \\ \mu_{xy} & \mu_{yy} \end{pmatrix} = \begin{pmatrix} \mu'_{xx} & \mu'_{xy} \\ \mu'_{xy} & \mu'_{yy} \end{pmatrix}. \tag{5.11}$$

Consider now an inverse problem, given any image structure with a corresponding second-moment descriptor μ, we want to find G that brings μ into a block-diagonal form of μ'. By combining Equations (5.10) and (5.11) it is straightforward to get estimates of the velocities as

$$\begin{pmatrix} \tilde{v}_x \\ \tilde{v}_y \end{pmatrix} = \begin{pmatrix} \mu_{xx} & \mu_{xy} \\ \mu_{xy} & \mu_{yy} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{xt} \\ \mu_{yt} \end{pmatrix}. \tag{5.12}$$

However, we note that Equation (5.11) holds only if the covariance matrix Σ of μ is related to the diagonal matrix Σ' of μ' by $\Sigma' = G^{-1}\Sigma G^{-T}$. Since G is the transformation we are trying to estimate, it follows that the equality in (5.11), in general, cannot be assumed. Hence the velocity estimates in (5.12) can only be regarded as approximate $\tilde{v}_x \approx v_x$ and $\tilde{v}_y \approx v_y$. Note, that these estimates correspond exactly to the solution for the optical flow by Local Least Squares of Lucas and Kanade (1981) presented in Section 3.4.

Since the approximate velocity-invariance of the descriptor $\tilde{\mu}' = \tilde{G}^T \mu \tilde{G}$, with $\tilde{G}$ defined by $\tilde{v}_x$ and $\tilde{v}_y$, may not be sufficient in practice, we improve the obtained velocity estimates using an iterative scheme. For this purpose, we use the framework of Galilean scale-space and adapt filter kernels to the recent values of estimated velocities. We start an iterative estimation process by computing μ using a diagonal covariance matrix Σ_0. Velocity values $\tilde{v}_{x,i}$, $\tilde{v}_{x,i}$ obtained in each iteration i are then used to update the covariance matrix according to $\Sigma_{i+1} = \tilde{G}_i^T \Sigma_0 \tilde{G}_i$, where G_i is defined by $\tilde{v}_{x,i}$, $\tilde{v}_{y,i}$. The convergence of velocity estimates to stable values indicates that (5.9) is satisfied and that the velocity-invariant second-moment descriptor $\mu' = G^T \mu G$ has been found.

We note the similarity of this iterative procedure with the procedure of estimating affine transformations in Section 3.1.4. In both cases, the second-moment matrix gives initial estimates of the transformations that are required to transform the underlying image pattern to a standard form, – to an isotropic pattern in the case of the affine transformations and to a stationary pattern in the case of Galilean transformations. However, the initial estimates in both cases depend on the shape of the filter kernels that have to be adapted to the required transformation in order to obtain true invariance.

5.2.1 Velocity and scale adaptation of events

To detect local space-time features independently of scaling and velocity transformations in image sequences, we propose to combine the iterative scheme for scale adaptation presented in Section 5.1.2 with the iterative scheme for velocity adaptation developed above. Hence, we start the estimation process using filter kernels with the initial covariance matrix $\Sigma_0 = G_0 S_0^2 G_0^T$ defined by a scale transformation matrix $S_0(\sigma_0, \tau_0)$ (3.33) and a Galilean transformation matrix $G_0(v_{x,0}, v_{y,0})$ (3.36). To obtain new estimates of scales $S_i(\sigma_i, \tau_i)$ at iteration i, we estimate (σ_i, τ_i) by maximizing the normalized Laplacian $(\nabla^2_{norm} L)^2$ (5.8) according to the iterative step of the algorithm in Figure 5.2. To update the Galilean transformation $G_i(v_{x,i}, v_{y,i})$, we use velocity estimates $(v_{x,i}, v_{y,i})$ according to (5.12). Then, at each iteration, the local space-time features are re-detected using the new estimate of the covariance matrix $\Sigma_i = G_i S_i^2 G_i^T$. The procedure stops when both the scales and velocities converge to stable values.

The result of event detection with iterative velocity estimation for a synthetic image sequence is presented in Figure 5.7(a). Here, a sequence with a moving corner in Figure 4.1(a) has been transformed by a Galilean transformation with velocities $v_x^a = 1.4$ in Figure 5.7(a) and $v_x^b = -0.8$ in Figure 5.7(b). The result of event detection is shown by ellipsoids with positions corresponding to the positions of the detected events and the shapes corresponding to the iteratively estimated covariance matrices Σ such that ellipsoids are described by $p^T \Sigma^{-1} p = 1$. The obtained estimates of velocities correspond exactly to the velocities v_x^a and v_x^b. We can also visually confirm that the shapes of estimated neighborhoods are adapted to the "skewing" transformation of the spatio-temporal image pattern.

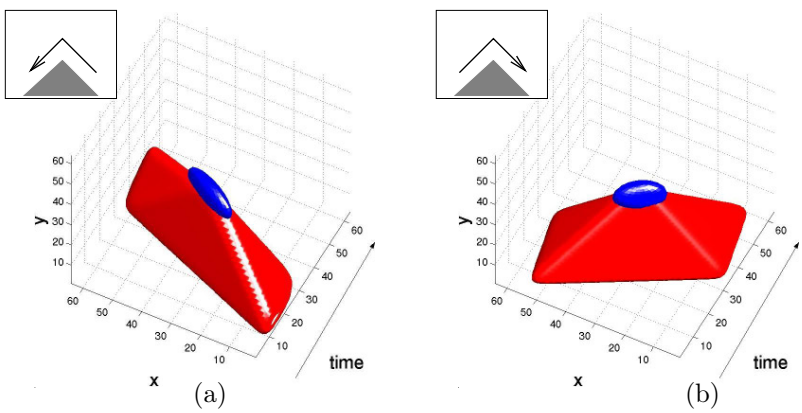


Figure 5.7: Detection of local space-time features for synthetic image sequences with moving corners that have been transformed by Galilean transformations with (a): $v_x = 1.4$ and (b): $v_x = 1.4$ with respect to the original pattern in Figure 4.1(a).

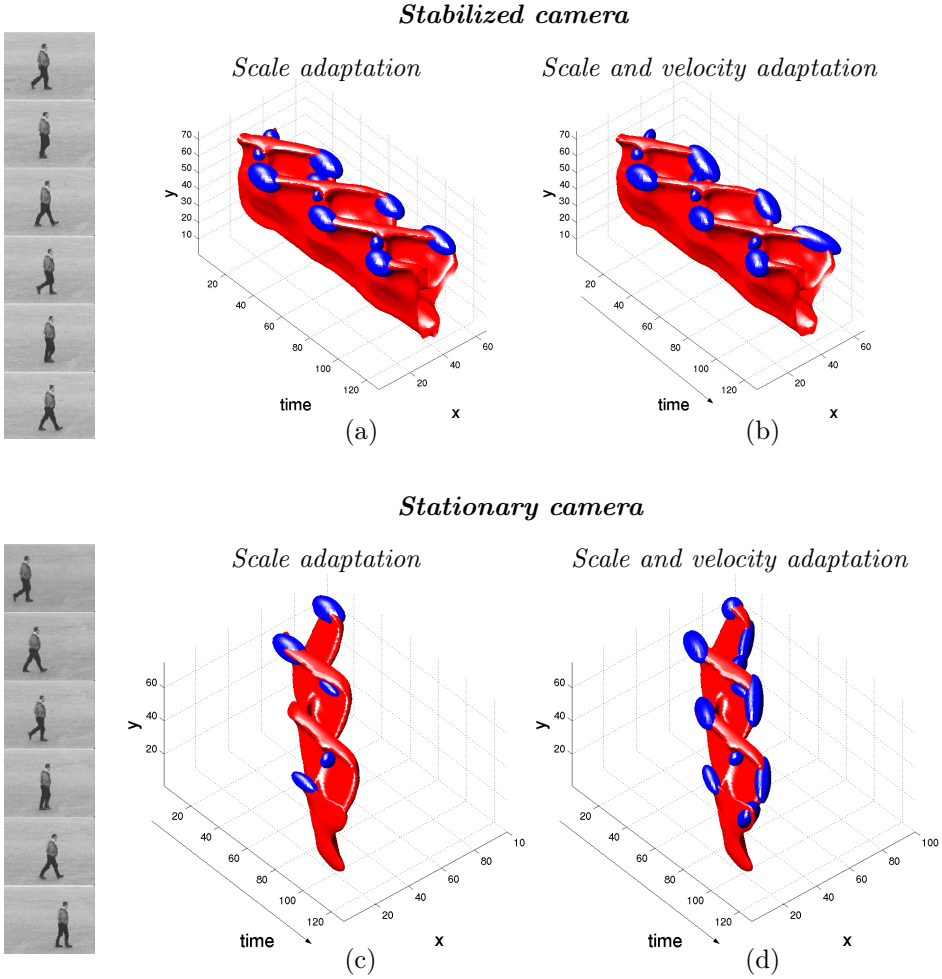


Figure 5.8: Detection of locally adapted space-time features for walking sequences acquired with a (a),(b): manually stabilized camera; (c),(d): stationary camera. Spatio-temporal patterns of the legs are illustrated by threshold surfaces in space-time (up-side-down). Space-time features are shown by ellipsoids with positions and shapes corresponding to the positions and covariance matrices of detected and adapted features respectively. (a),(c): Local features with adaptation to scale. (b),(d): Local features with adaptation to scale and velocity. By comparing (a) with (c) and (b) with (d) it follows that velocity adaptation is an essential mechanism if the same type of events are to be detected independently of the relative camera motion.

Results of detecting adapted space-time features for real image sequences are illustrated in Figure 5.8. Here, two sequences of a walking person were taken with a hand-stabilized camera (top) and a stationary camera (bottom). As illustrated by corresponding plots with three-dimensional threshold surfaces, the resulting spatio-temporal patterns relate to each other by a skew transformation in time originating from the relative Galilean transformation between the image sequences.

The result of detecting features with simultaneous adaptation to scales and velocities (see Figures 5.8(b),(d)) is compared to the feature detection with scale adaptation only (see Figures 5.8(a),(c)). As can be seen, the approach with scale adaptation only fails to detect similar events in both sequences. Moreover, the estimated shape of a few corresponding features differs substantially depending on the motion of the camera and it follows that scale adaptation alone is not sufficient for comparing image sequences effected by different velocity transformations. On the contrary, local feature detection with simultaneous adaptation to scales and velocities results in the similar types of events disregarding the motion of the camera (see Figures 5.8(b),(d)).

5.2.2 Discussion

When formulating a theory for detecting local space-time features in Sections 4.1 and 4.2, we introduced a velocity-corrected operator H (4.10) defined from the velocity-corrected second-moment descriptor (4.8). By comparing this approach to the iterative approach above, it becomes apparent that velocity correction in Section 4.2 corresponds to the first step of the iterative velocity adaptation presented in this section. Since the velocity correction is less computationally expensive at the loss of precision, the choice between these two methods should be made depending on the available computational power. Note also that the presented iterative approach for invariant velocity adaptation is not restricted to the adaptation of a *sparse* set of points such as motion events, but can also be applied to the *dense* adaptation of all points in the sequence.

The aim of velocity adaptation so far has been to develop a method for velocity-invariant detection of space-time features. Another advantage of adaptation is that it enables invariant estimation of neighborhoods of the detected events. This advantage will be explored in the next chapter when formulating descriptors for comparing and matching local events in different image sequences.

Finally, we note that the presented iterative adaptation scheme is unlikely to be suited for real-time applications due to its high computational complexity. In this work, we currently disregard this problem in order to investigate the implications of the full invariance with respect to scales and velocities. For real-time applications, however, the analysis presented here could constitute a basis for developing related methods with approximate invariance. One interesting direction toward real-time implementations would be to investigate possibilities for a sparse sampling of the Galilean scale-space and its interpolation, as well as, non-iterative methods for invariant feature detection from interpolated representations.

5.3 Dense velocity adaptation

So far we have investigated methods for obtaining scale and velocity invariant local space-time features. The use of such features has been motivated in Chapter 4 by the idea of dividing continuous motion into a set of structural elements (events) for the purpose of subsequent interpretation. A similar idea for image representations in terms of local features has shown to be very successful for matching and recognition in the spatial domain (Schmid and Mohr, 1997; Lowe, 1999; Weber et al., 2000; Fergus et al., 2003; Sivic and Zisserman, 2003). The advantages of local features include their stability with respect to photometric and geometric transformations in the image as well as the stability to the presence of occlusions and variations in the background.

Local features, however, capture only partial information in images within local neighborhoods of *sparse* locations. In the context of the present work, this implies that local space-time features will not be sensitive to image neighborhoods with constant motion or neighborhoods with one-dimensional spatial variations. Whereas information in such neighborhoods can be crucial in some situations, it is interesting to consider alternative *dense* representations that are capable of capturing all the local information present in the scene. Among the many alternatives for such methods, global histograms of dense filter responses have given good performance, in some situations, for the tasks of spatial and spatio-temporal recognition (Swain and Ballard, 1991; Schiele and Crowley, 2000; Chomat, de Verdiere, Hall and Crowley, 2000; Chomat, Martin and Crowley, 2000; Zelnik-Manor and Irani, 2001; Linde and Lindeberg, 2004).

Similar to local features, dense filter responses depend on the common image transformations such as scalings and rotations. In particular, as discussed in Section 3.2, spatio-temporal filtering depends on the velocity-transformations of space-time image patterns (see Figure 3.4). Previous work has addressed this problem by first stabilizing patterns of interest in the field of view (Irani et al., 1995), and then computing spatio-temporal descriptors using a fixed set of filters. Camera stabilization, however, may not always be available, for example, in situations with multiple moving objects, moving backgrounds or in cases where initial segmentation of the patterns of interest cannot be achieved without (preliminary) recognition (see Figure 3.5 for an illustration).

In contrast to the previous work on *global* velocity adaptation, this section considers *local* velocity adaptation of dense spatio-temporal filter responses. The main idea is to obtain information about motion within local spatio-temporal neighborhoods and to use this information for velocity adaptation of motion descriptors in the same neighborhoods. In other words, we aim to cancel out the effect of constant motion on the space-time descriptors and consider higher order descriptors in order to describe non-constant motion independently of the relative (constant) velocity of the camera. With respect to previous methods for motion estimation, we do not enforce a correct estimation of optical flow but will require the invariance of velocity estimation with respect to Galilean transformations.

5.3.1 Mechanism for dense velocity adaptation

To achieve Galilei-invariant velocity estimation at each point in space-time, we could, in principle, use an iterative approach of velocity adaptation described in Section 5.2. An iterative approach applied to each space-time point, however, would be computationally expensive and not suited in practice. An alternative we follow here, is to sample the Galilean scale-space using a sparse set of velocities and to measure how well does each velocity describes the local motion in the neighborhood of each space-time point.

To accomplish this task, one approach would be to compute a second-moment matrix μ (4.2) for each space-time-velocity point (x, y, t, v_x, x_y) and to measure how close this matrix is to a block-diagonal form according to (4.8). Here we use an alternative approach inspired by the related work on automatic scale selection (Lindeberg, 1998b) as well as by motion energy approaches for computing optic flow (Adelson and Bergen, 1985; Heeger, 1988). Given a set of image velocities $V_x = \{-v_x^N, ..., 0, ..., v_x^N\}$, and $V_y = \{-v_y^M, ..., 0, ..., v_y^M\}$, the responses of a Laplacian operator in space

$$\nabla^2 L_{spat}(\cdot;\, \Sigma) = L_{xx}(\cdot;\, \Sigma) + L_{yy}(\cdot;\, \Sigma) \tag{5.13}$$

are computed for each image velocity v_x^i, v_y^j using $\Sigma = GSG^T$ with a diagonal scaling matrix $S(\sigma^2, \tau^2)$ according to (3.32) and a Galilean matrix $G(v_x^i, v_y^j)$ according to (3.36). Then, a motion estimate at each space-time point $(\cdot)$ is obtained by maximizing the responses of $(\nabla^2 L_{spat}(\cdot;\, \Sigma))^2$ over velocities according to

$$(\tilde{v}_x, \tilde{v}_y) = \underset{v_x \in V_x, v_y \in V_y}{\operatorname{argmax}} \left(\nabla^2_{spat} L(\cdot;\, \Sigma) \right)^2 \tag{5.14}$$

This approach corresponds to the application of a set of velocity-adapted Laplacian operators (see Figure 5.9) and selecting velocities corresponding to the filter with the maximum absolute response. To obtain invariant velocity estimates $(\tilde{v}_x, \tilde{v}_y)$ at different scales, we maximize (5.14) separately at each scale level (σ^2, τ^2).

Figure 5.10 illustrates the results of local velocity adaptation for a synthetic spatio-temporal pattern in Figure 5.10(a) and its velocity-transformed version in Figure 5.10(d) (for the clarity of presentation, only one spatial dimension is considered here). From the responses of velocity-adapted filters and from the ellipses displaying the selected orientations of filters in space-time, it is apparent that the proposed filtering scheme adapts to the local motion and enhances structures both in the moving pattern and in the static background. This result is in contrast to the global velocity adaptation illustrated previously for the same pattern in Figure 3.5. Moreover, by comparing the results in Figures 5.10(e) and 5.10(f), we can visually confirm the invariance of locally adapted filter responses with respect to the Galilean transformation of the pattern or, equivalently, to the relative motion between the pattern and the camera.

Application of the local velocity adaptation to a real sequence with a walking person is illustrated in Figure 5.11. Note, that filtering here has been done in

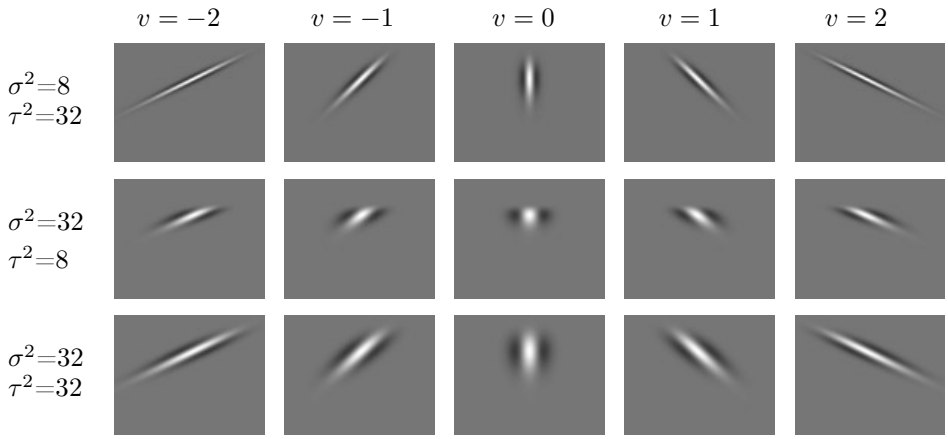


Figure 5.9: Spatio-temporal filters L_{xx} computed from a velocity-adapted spatio-temporal scale-space for a 1+1–D image pattern, for different values of the velocity parameter v, the spatial scale σ^2 and the temporal scale τ^2.

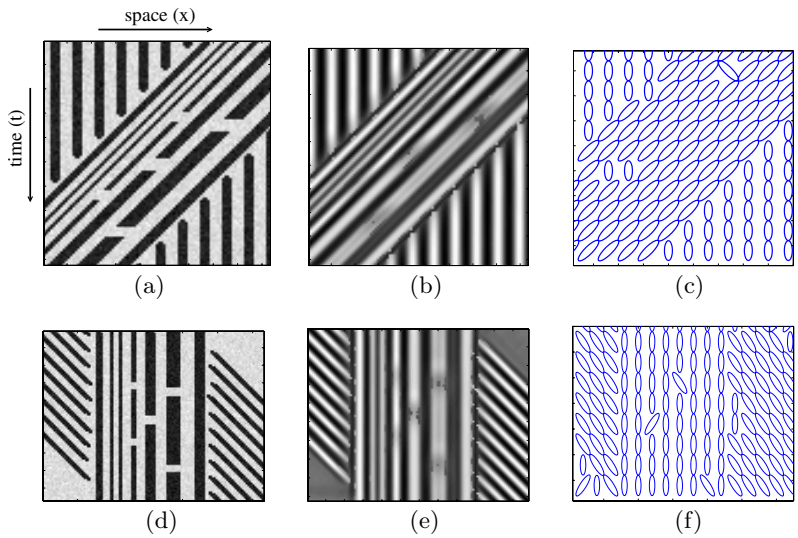


Figure 5.10: Results of filtering original patterns in (a) and (d) using the proposed *local velocity adaptation* are illustrated in (b) and (e) respectively. The orientation of the ellipses in (c) and (f) show the chosen velocity at each point of the pattern. Note that filtering with local velocity adaptation preserves the details of the moving and stationary pattern. The similarity of the filter responses in (b) and (e) also illustrates the independence of the filtering results with respect to the constant motion of the camera.

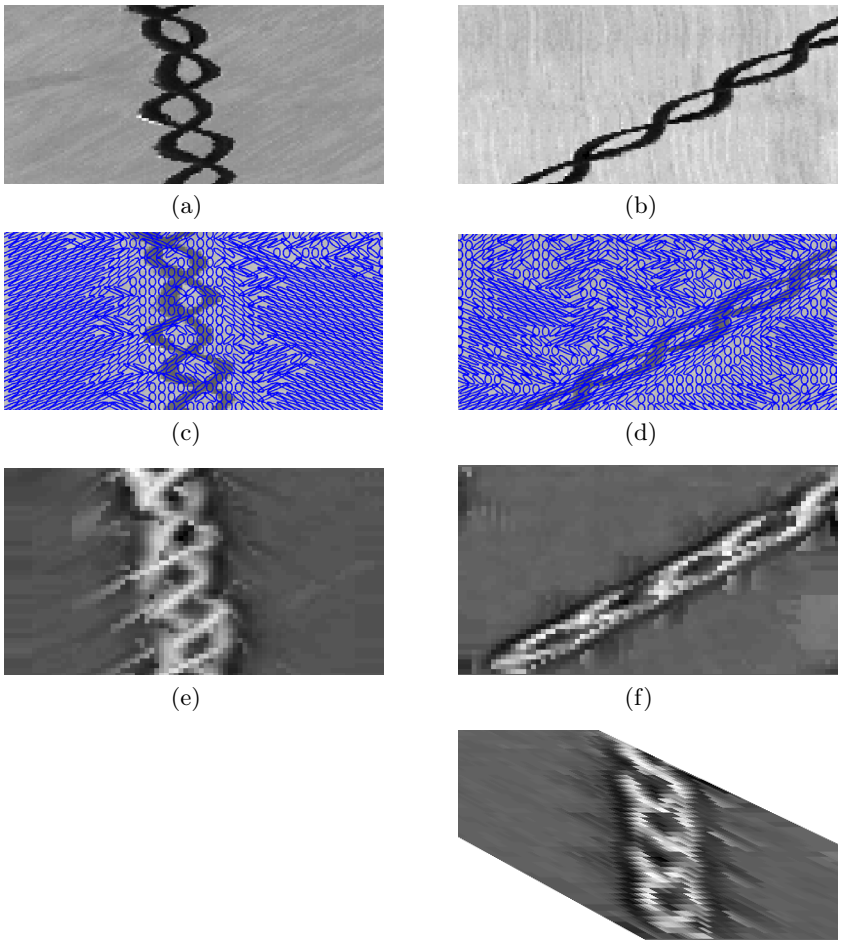


Figure 5.11: Spatio-temporal filtering with local velocity adaptation applied to an image sequence of a walking person recorded with a (a): stabilized camera and (b): stationary camera (see Figure 3.4 for comparison). (c)-(d): velocity adapted shape of filter kernels; (e)-(f): results of filtering with a second-order derivative operator; (g): velocity-warped version of (f) showing high similarity with the results in (e).

three dimensions while for the purpose of demonstration, the results are shown only for one x-t–slice of a spatio-temporal cube presented in Figure 3.4. As for the synthetic pattern in Figure 5.10, we observe successful adaptation of filter kernels to the motion structure of a gait pattern (see Figures 5.11(c),(d)). The results in Figures 5.11(e) and (g) also demonstrate approximative invariance of filter responses with respect to camera motion. The desired effect of the proposed local velocity adaptation is especially evident when these results are compared to the results of separable filtering illustrated in Figures 3.4(d)-(f).

Given estimates of the velocity-invariant covariance matrices $\Sigma = GSG^T$, it is now straightforward to compute velocity-adapted spatio-temporal derivative responses $L_{x^m,y^n,t^k}(\cdot;\, \Sigma)$ (3.35) at every point $(\cdot)$ in space-time. The implications of such an adaptation will be evaluated in Chapter 7 on the problem of action recognition using histograms of spatio-temporal filter responses.

In relation to the previous work, the proposed adaptive filtering procedure bares similarities with anisotropic filtering that has been investigated in the spatial domain (Perona and Malik, 1990; Alvarez et al., 1992; Florack et al., 1995; Black et al., 1998; Weickert, 1998; Almansa and Lindeberg, 2000) and in space-time (Nagel and Gehrke, 1998; Guichard, 1998). Here, instead of using local image gradient to adapt the shape of spatial filters, we use velocity estimates in order to adapt the shape of spatio-temporal filters with respect to a Galilean transformation.

Chapter 6

Motion descriptors

The aim of this work is to explore the role of motion events when representing and recognizing complex patterns of motion. The idea is to use local motion events as primitives for describing the characteristic parts of motion patterns and to construct a part-based representation of video data. The advantages of representing image data by parts have recently been explored in the field of spatial recognition. While being local, parts, in general, are more stable than global representations with respect to geometric and photometric transformations (Schmid et al., 2000; Mikolajczyk and Schmid, 2003) as well as to possible non-rigid transformations of the object. Moreover, part-based representations have been demonstrated to be stable with respect to within-class variations of the appearance of certain object classes. For example, in (Weber et al., 2000) and following work (Fergus et al., 2003; Fei-Fei et al., 2003) it has been shown that some characteristic image parts can be reliably detected and can be used for recognizing a number of visual categories.

Recognition by parts involves matching between parts in the training set and in test samples. This poses natural requirements on the parts in terms of *stability* and *discriminability*. Stability implies that the same parts should be detected for the same objects or activities in different images or video sequences. Moreover, to reduce the ambiguity of matching, it is also desirable that the parts are supplemented with descriptors that allow for matching of similar parts and discrimination between different types of parts.

Figure 6.1 shows examples of detected local space-time features for image sequences with human actions. From these results and other examples in Chapters 4 and 5, we can visually confirm the consistency of detected features with respect to repeated events in the data. Moreover, by analyzing image information within spatio-temporal neighborhoods of detected features (see Figure 6.2), we observe that different actions give rise to different types of motion events. Given the stability of space-time features and the possibility to discriminate between these, it follows that such features provide promising candidates for *parts* when constructing part-based representations of motion patterns.

The detection of motion events according to the methods in Chapters 4 and 5 provides information about the position and the shape of such events in space-time. To discriminate different events and to enable reliable matching of similar events in different sequences, the next two sections present a number of local space-time descriptors as well as different metrics for measuring similarities between the events. At the end of this chapter, Section 6.3 presents two methods for constructing and comparing part-based motion representations in terms of local space-time features. The next chapter evaluates both the stability of space-time features and the discriminative power of corresponding descriptors.

6.1 Local space-time descriptors

An ideal descriptor for motion events should capture characteristic properties of the image data while disregarding irrelevant transformations. Transformations of the spatio-temporal image data may depend on different factors including scale and velocity transformations, photometric transformations, inaccuracy of the feature detector as well as individual variations in appearance and motion of particular events. Some of these factors, such as scale and velocity transformations, have been addressed in the previous chapters and are rather well understood. Other factors, however, such as individual variations, depend on the classes of considered events and are harder to formalize. In such a situation, the criteria of optimality depend on the task and may be difficult to derive analytically. To deal with this problem, we here take an experimental approach and define a number of alternative descriptors whose performance will then be evaluated in practice. The design of these descriptors is inspired by related work in the spatial domain (Koenderink and van Doorn, 1987; Koenderink and van Doorn, 1992; Lowe, 1999; Mikolajczyk and Schmid, 2003) and in the spatio-temporal domain (Yacoob and Black, 1999; Hoey and Little, 2000; Zelnik-Manor and Irani, 2001).

To capture variations of image values in space and time, it is common to use differential measurements such as image derivatives. In this work, all local descriptors are defined in terms of combinations of Gaussian spatio-temporal derivatives. Moreover, to compensate for scale and velocity transformations in image sequences, we use scale-normalized and velocity-adapted derivative operators according to equations (3.35) and (3.38). Given the estimates of scales (σ, τ) and velocities (v_x, v_y), the responses of such operators are computed as

$$L_{x^m y^n t^k} = \sigma^{m+n} \tau^k \partial_{(x')^m (y')^n (t')^k} L'(\cdot;\, \Sigma') \tag{6.1}$$

using Galilei-transformed derivative operators (3.38) and adapted filter kernels with a covariance matrix $\Sigma' = GS^2G^T$ defined by $S = S(\sigma, \tau)$ (3.33) and $G = G(v_x, v_y)$ (3.36). Assuming invariant estimates of velocities and scales according to the previous chapter, the transformation properties of Gaussian derivatives in Section 3.2 guarantee that $L_{x^m y^n t^k}$ (6.1) is invariant with respect to scale and velocity transformations in image sequences. Hence, it follows, that all descriptors derived

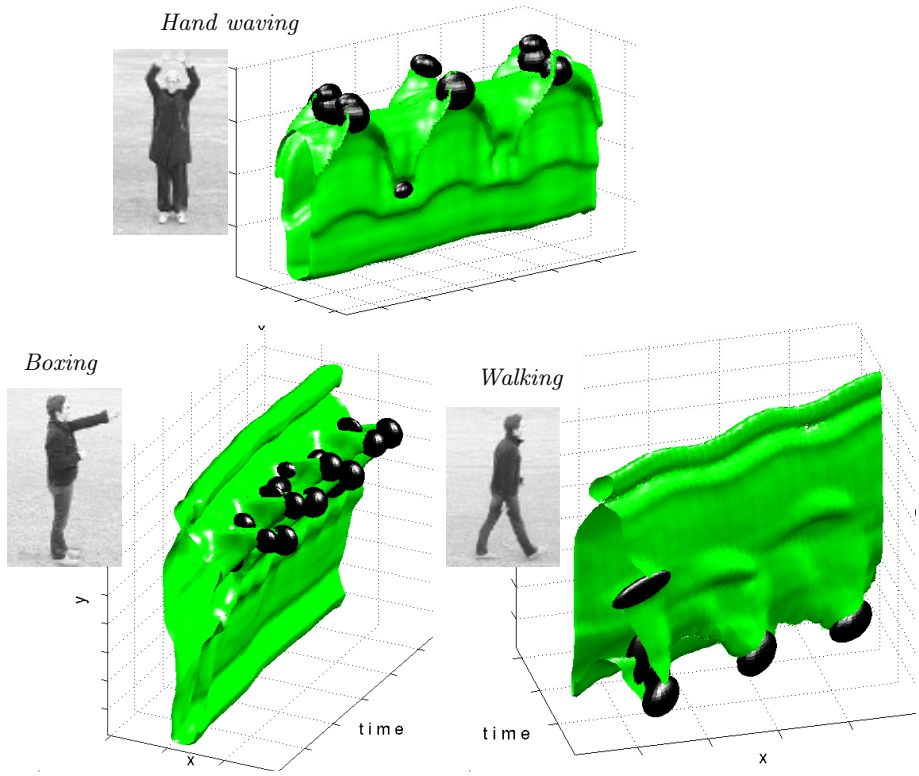


Figure 6.1: Examples of scale and Galilean adapted spatio-temporal interest points. The illustrations show one image from the image sequence and a level surface of image brightness over space-time with the space-time interest points illustrated as dark ellipsoids.

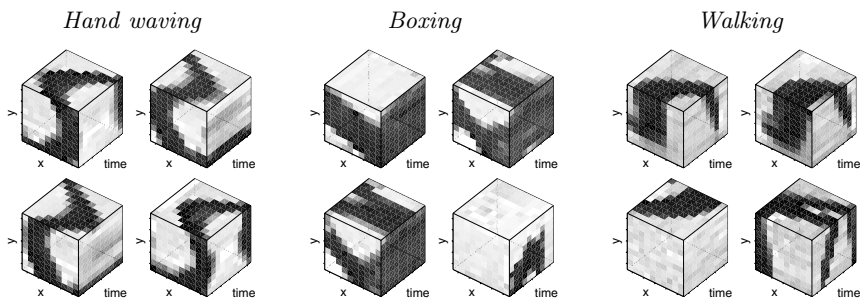


Figure 6.2: Examples of spatio-temporal patches corresponding to neighborhoods of local space-time features detected for different actions in Figure 6.1.

from (6.1) are also invariant with respect to scale and velocity transformations in the data.

To define descriptors for the neighborhoods of local space-time features, we use derivative responses evaluated either at the center-point of the feature or at all points within a locally adapted neighborhood. The first type of measurement corresponds to a local spatio-temporal N-jet descriptor and is described in Section 6.1.1. For dense measurements within local neighborhoods, we either compute spatio-temporal gradients or optic flow. Furthermore, to combine dense measurements of both types into descriptors, we either use Principle Component Analysis (PCA) as described in Section 6.1.2 or marginalized histograms of vector fields as described in Section 6.1.3.

While some of these descriptors are defined by derivative responses and others in terms of optic flow (computed from spatio-temporal derivatives), there are a number of similarities and differences between these two approaches. Measurements in terms of pure derivatives contain an encoding of the complete space-time image structure around the motion event, including an implicit encoding of the optic flow. By explicitly computing the optic flow, we obtain representations which only emphasize the motion and suppress information about the local contrast and orientation of the image structures in the spatial domain. Such invariance can 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 both types of image measurements.

With respect to different combinations of measurements into descriptors, representations in terms of N-jets, PCA or histograms can also be expected to have different advantages and limitations due to a number of specific properties. N-jet descriptors have a relatively low dimensionality and few parameters while describing the data using a standard set of basis functions (partial derivatives) that do not take advantage of any specific properties of events in space-time. PCA-descriptors, on the other hand, are defined in terms of bases learned from local features detected in training sequences. Such an approach can be expected to perform better when representing specific events but also to be more sensitive to possible variations between events in the training and the testing sets. Finally, histogram-based representations partly or fully disregard relative positions between measurements within neighborhoods. While this implies a loss of information, on the other hand, such descriptors can be expected to gain stability with respect to certain individual variations of events as well as to possible errors in their estimated positions.

Beside invariance to scale and velocity, all descriptors, in one way or another, are normalized with respect to the amplitude of derivatives to provide invariance with respect to changes in the illumination. It should be noted, however, that each type of invariance increases the stability of descriptors at the cost of decreasing their discriminative power. Determining this trade-off is a task of learning and experimental evaluation.

6.1.1 Spatio-temporal N-Jets

Local N-jet descriptors were proposed as visual operators by Koenderink and van Doorn (1987) and contain partial derivatives of an image function evaluated at the same point. Such a representation corresponds to a truncated Taylor expansion of a function about a given point. Local spatio-temporal N-jets are defined as vectors containing responses of partial spatio-temporal derivative operators up to order N:

$$\begin{aligned}\mathcal{J}(\cdot; S, G) \quad = \quad & (L_x, L_y, L_t, \\ & L_{xx}, L_{xy}, L_{xt}, L_{yy}, L_{yt}, L_{tt}, \\ & L_{xxx}, L_{xxy}, L_{xxt}, L_{xyy}, L_{xyt}, \ldots, L_{ttt}, \\ & L_{xxxx}, L_{xxxy}, L_{xxxt}, L_{xxyy}, L_{xxyt}, \ldots, L_{tttt}, \ldots)\end{aligned} \tag{6.2}$$

To obtain invariance with respect to scalings and Galilean transformations, we here use scale-normalized and velocity-adapted derivatives as defined in (6.1). The notation $\mathcal{J}(\cdot; S, G)$ in (6.2) indicates that all derivatives in the jet-vector are computed at the center of the detected event $(\cdot)$ and at scales $S(\sigma, \tau)$ and velocities $G(v_x, v_y)$ estimated by the scale and velocity adaptation mechanism described in Section 5.2.1. In this work, we use N-jets of order two and four. Such descriptors will be denoted as **2Jets** and **4Jets** and will contain 9 and 34 elements respectively.

The evolution of image derivatives over scales depends on the underlying image function and encodes the local properties of the image. To increase the information content of the descriptor, we can extend N-jets computed at a single scale to *multi-scale N-jets* computed at several scales. For this purpose and in order to preserve invariance to scale transformations, we compute multi-scale N-jets with respect to scale matrices $S^{ij}(s_x^i \sigma, s_t^j \tau)$, with factors s_x and s_t applied to the spatial and temporal scales respectively. In our experiments, we use combinations of three spatial and three temporal scales defined by factors $s_x = \{0.5, 1, 2\}$ and $s_t = \{0.5, 1, 2\}$. Hence, the resulting multi-scale N-jets are defined in terms of single-scale N-jets in (6.2) at nine spatio-temporal scales S^{ij} as

$$\mathcal{J}^{ms}(\cdot; S, G) = \big(\mathcal{J}(\cdot; S^{11}, G), \mathcal{J}(\cdot; S^{12}, G), \ldots, \mathcal{J}(\cdot; S^{33}, G)\big) \tag{6.3}$$

Multi-scale jets of order two contain 81 elements and will be denoted as **MS2Jets**. Similarly, the abbreviation **MS4Jets** will correspond to multi-scale jets of order four, containing 306 elements. Filter shapes for the selected components of multi-scale N-jets are illustrated in Figure 6.3. To compensate for illumination variations, we normalize all types of N-jets to unit l_2-norm.

The computation of the Galilei-adapted N-jets in (6.2) and (6.3) involves non-separable filtering (or, equivalently, velocity-warping). Whereas this operation is more computationally expensive than separable filtering, the need for non-separable filtering is evident from Figure 6.4. Here, the original impulse-like signal with velocity $v_x = 2$ was filtered using a velocity adapted filter kernel in Figure 6.4(a) and separable filter kernels in Figure 6.4(b),(c). By comparing these results with the corresponding responses in the third column of Figure 6.3, we note that a correct

shape of the second-order derivative in space and the first-order derivative in time is only obtained for the case of non-separable filtering. From the incorrect shape of a filter response in Figure 6.4(b), we can also conclude that the adaptation of derivative operators alone according to (3.38) without adapting filter kernels is not sufficient for computing velocity-invariant derivative responses. Hence, in order to obtain velocity-invariant N-jet descriptors, both the derivative operators and filter kernels have to be adapted to velocity transformations according to (6.1). This statement is further supported by experimental evaluation in Chapter 7.

6.1.2 Principal Component Analysis

Principle Component Analysis (PCA) is a standard technique for dimensionality reduction of data (Duda et al., 2001). Given samples of n-dimensional data vectors d with a corresponding covariance matrix C and a mean m, the eigenvectors e_i and the corresponding eigenvalues λ_i of C are computed by solving the equations $Ce_i = \lambda_i e_i$. Then, by choosing the $k \ll n$ eigenvectors corresponding to the largest eigenvalues, the k-dimensional subspace spanned by the chosen eigenvectors is regarded to correspond to the "signal" while the rest of $n - k$ dimensions are assumed to correspond to noise. By forming an $n \times k$ matrix E whose columns consist of k chosen eigenvectors, dimensionality reduction of a data set d is achieved by projecting d to the subspace E as $d' = E^T(d - m)$.

Besides reducing the dimensionality, PCA also disregards variations in the data that are not present in the initial (training) data set. This can be a valuable property in order to suppress irrelevant variations in the test set by considering only distances between training and test samples in the projected space E.

For the purpose of using PCA to represent image variation in the neighborhoods of motion events, we consider data vectors d as either components of optic flow (OF) vectors or components of spatio-temporal gradient (STG) vectors as follows: Given a neighborhood of an event with estimated scales (σ, τ) and velocities (v_x, v_y), we velocity-warp the spatio-temporal image patch around the position p_0 of the detected feature using the Galilean transform $p' = G(-v_x, -v_y)(p - p_0)$. Then, either OF or STG vectors are computed for each point $p'(x_i, y_i, t_i)$ in the warped image patch according to (3.45) or (3.35) respectively. Vector fields obtained in a space-time window $(\pm\alpha\sigma, \pm\alpha\sigma, \pm\beta\tau)$ are then re-sampled to a patch with a standard size of $9 \times 9 \times 9$ pixels using trilinear interpolation. The re-sampled flow fields are finally represented as vectors

$$\begin{array}{rcl} d^{of} & = & (v_{x,1}, v_{y,1}, v_{x,2}, v_{y,2}, \ldots, v_{x,279}, v_{y,279}) \\ d^{stg} & = & (L_{x,1}, L_{y,1}, L_{t,1}, L_{x,2}, L_{y,2}, L_{t,2}, \ldots, L_{x,279}, L_{y,279}, L_{t,279}) \end{array} \tag{6.4}$$

for OF and STG with the dimensionality 1458 and 2187 respectively. To compensate for illumination variations, the vectors d^{stg} are, in addition, normalized to unit l_2-norm.

To obtain eigenvectors, we randomly select about 10000 space-time features detected from training sequences, compute d^{of} and d^{stg} for each feature and solve the

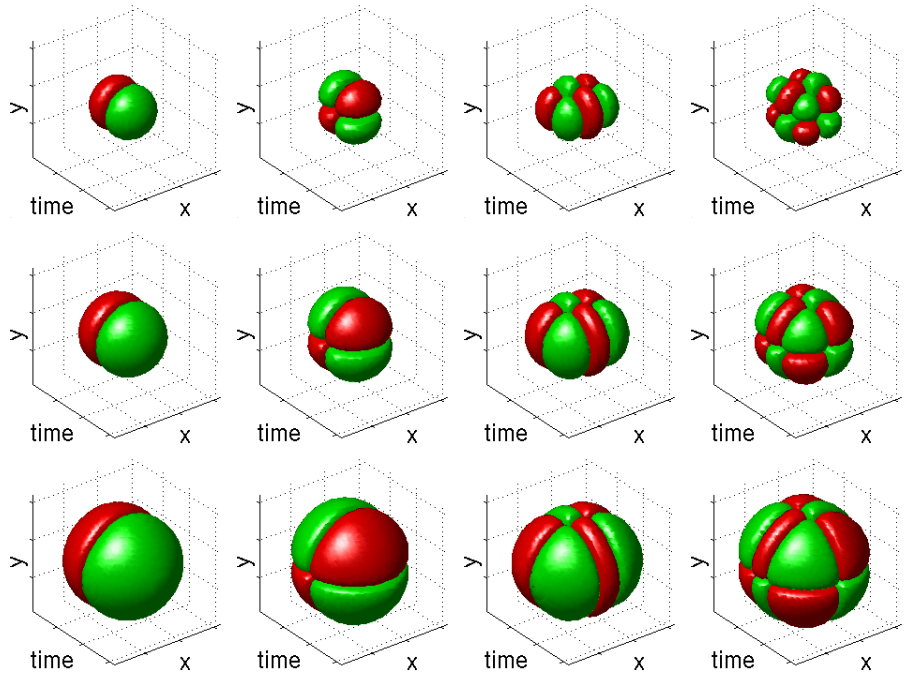


Figure 6.3: Examples of impulse responses of spatio-temporal derivative operators at different scales. The responses are illustrated by threshold surfaces with the color corresponding to the sign of responses. From left to right: L_t, L_{yt}, L_{xxt}, L_{xytt}.

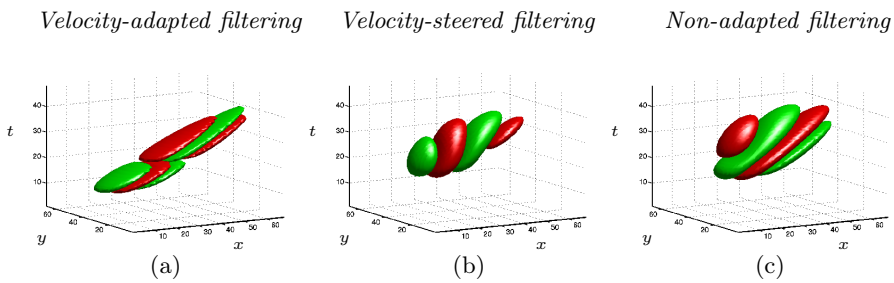


Figure 6.4: Responses of the ∂_{xxt}-derivative operator to a prototype impulse with velocity $v_x = 2$. (a): Velocity-adaptation of the derivative operator and the filter kernel according to (6.1); (b): separable filtering with velocity-adaptation of the derivative operator alone according to (3.38); (c): separable filtering without adaptation of the derivative operator. A correct shape of the filter response is obtained only for the case of velocity-adapted filtering in (a).

eigenvalue problem for covariance matrices obtained from d^{of} and d^{stg} respectively. We then select the $k = 100$ most significant eigenvectors and obtain subspaces E^{of} for OF and E^{stg} for STG. To reveal which variations in the data are represented by different components, we can reshape the obtained eigenvectors back into the form of three-dimensional vector fields. Such vector fields, corresponding to the first twelve STG eigenvectors, are illustrated in Figure 6.5.

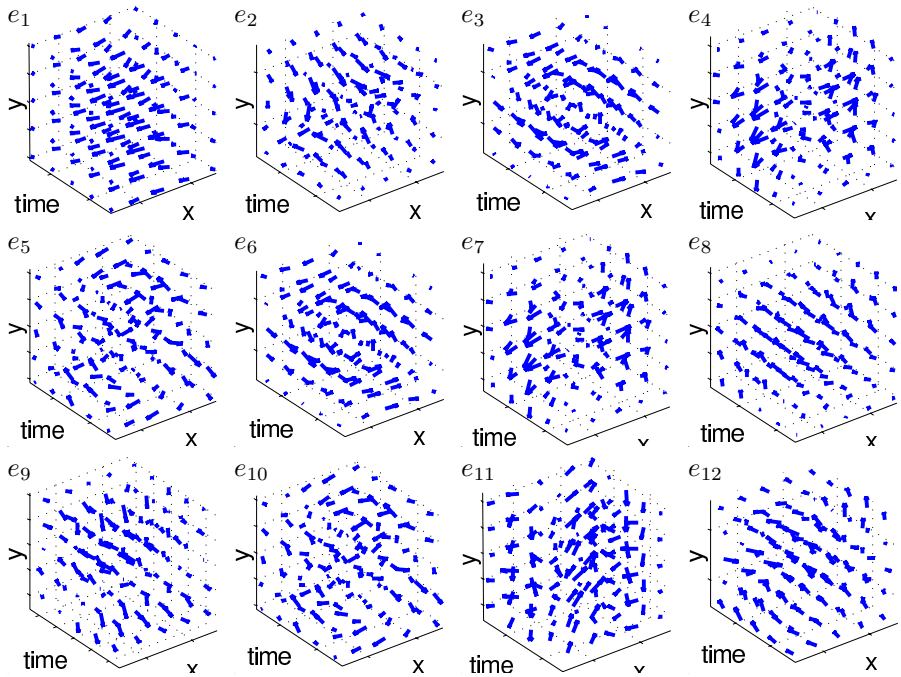


Figure 6.5: The twelve most significant eigenvectors obtained by performing PCA on spatio-temporal gradient fields computed at the neighborhoods of motion events. Although the interpretation of three-dimensional vector fields is somewhat difficult, we can observe increasing levels of details for eigenvectors with lower eigenvalues.

Given any neighborhood of a motion event, its PCA representation is computed by evaluating either d^{of} or d^{stg} according to (6.4). Then, to form an optic flow descriptor (**OF-PCA**), the vector d^{of} is projected to the OF basis according to $d^{ofpca} = (E^{of})^T d^{of}$. Similarly, PCA descriptors for spatio-temporal gradients (**STG-PCA**) are obtained by projecting d^{stg} according to $d^{stgpca} = (E^{stg})^T d^{stg}$. The dimensionality of both types of descriptors equals $k = 100$.

We note, that OF-PCA shares similarities with the PCA-based representation of optic flow proposed earlier by (Black and Jepson, 1998a; Rui and Anandan, 2000). Here, however, we encode variations of optic flow over both space and time and compute descriptors for local space-time patches rather than for the whole scene.

6.1.3 Histogram-based descriptors

Both N-jets and PCA-based descriptors encode image variations depending on their positions inside neighborhoods. Hence, inaccuracies in the estimation of neighborhoods due to possible errors in feature detection or adaptation may lead to instabilities of such descriptors. A simple but effective way to eliminate positional dependencies in the image data consists of computing histograms of local filter responses. Such histograms encode only the frequency of local measurements and disregard their positions in the image. Histogram-based representations were used by (Swain and Ballard, 1991; Schiele and Crowley, 1996) for recognizing objects in still images. Extensions of this approach to the spatio-temporal domain were later presented by (Chomat and Crowley, 1999; Zelnik-Manor and Irani, 2001). With respect to local representations, Lowe (1999) proposed a histogram-based SIFT descriptor which has shown the best performance when compared to other types of local descriptors in the context of matching (Mikolajczyk and Schmid, 2003).

In this section we present several histogram-based descriptors for representing information in the neighborhoods of motion events. As for PCA-based descriptors, we compute vectors of spatio-temporal gradients $\nabla L = (L_x, L_y, L_t)^T$ or vectors of optic flow $V = (v_x, v_y)^T$ for each point in the velocity-warped neighborhood of a detected space-time feature. Similar to multi-scale N-jets in Section 6.1.1, both the gradients and the optic flow vectors are computed at combinations of three spatial and three temporal scales $S^{ij}(s_x^i\sigma, s_t^j\tau)$ for $s_x = \{0.5, 1, 2\}$ and $s_t = \{0.5, 1, 2\}$. Each one of the gradient vectors is additionally normalized by its magnitude, hence, this representation is closely related to normal flow.

To combine the resulting local measurements into a descriptor, one approach we follow consists of computing histograms of either ∇L or V in the *whole* neighborhood of a local feature. For simplicity, we accumulate histograms separately for each component of the gradient L_x, L_y, L_t or each component of the velocity v_x, v_y. Moreover, separate histograms are computed for each component at each one of the nine combinations of spatio-temporal scales. To emphasize measurements at the center of the neighborhood and to reduce the dependency of descriptors with respect to translations, we compute weighted histograms where the contribution of each local measurement is weighted with a Gaussian window function in space-time with the center at the position of a space-time feature and with a covariance matrix $\Sigma = (3S)^2$ defined by the scale of a feature $S(\sigma, \tau)$. As result of this approach, for the gradient measurements we obtain 27 histograms $h^{stg}_{k,i,j}$ where $k = \{1, 2, 3\}$ stands for different components L_x, L_y, L_t and $i, j = \{1, 2, 3\}$ indicate different spatio-temporal scales. Similarly, for the case of optic flow we obtain 18 histograms $h^{of}_{k,i,j}$ where $k = \{1, 2\}$ indicates components v_x, v_y and $i, j = \{1, 2, 3\}$ stands for the scales. By combining the histograms into a vector, we obtain two histogram-based descriptors, one for spatio-temporal gradients (**STG-HIST**) and one for optic flow (**OF-HIST**):

$$\begin{aligned} H^{stg} &= (h^{stg}_{1,1,1}, h^{stg}_{2,1,1}, h^{stg}_{3,1,1}, \ldots, h^{stg}_{3,3,3}) \\ H^{of} &= (h^{of}_{1,1,1}, h^{of}_{2,1,1}, h^{of}_{3,1,1}, \ldots, h^{of}_{3,3,3}). \end{aligned} \tag{6.5}$$

Using 32 bins for representing either $h^{stg}_{k,i,j}$ or $h^{of}_{k,i,j}$ histograms, results in a 864-dimensional STG-HIST descriptor and a 576-dimensional OF-HIST descriptor.

When comparing STG-HIST to previous work, we recognize its similarity to a histogram-based representation by (Zelnik-Manor and Irani, 2001). The differences here are that (i) in addition to different temporal scales, we compute histograms for different spatial scales, (ii) we encode the sign of L_x, L_y, L_t and (iii) we compute histograms from all points in local spatio-temporal neighborhoods rather than from moving points in the whole image sequence.

Position-dependent histograms. Positional dependencies of local measurements are clearly important for visual interpretation. In the search for an appropriate trade-off between invariance and discriminative power of local descriptors, Lowe (1999) proposed to subdivide local neighborhoods into parts and compute histograms for each of these parts separately. As different parts are defined in terms of relative positions with respect to the center of local features, the coarse positional information is preserved in the descriptor. At the same time, histogram-based representation of parts provides sufficient invariance with respect to the limited variations of the neighborhoods.

The idea of position-dependent histograms is extended here for representing motion events. For this purpose, we subdivide spatio-temporal neighborhoods of space-time features into $M \times M \times M$ position-dependent parts as illustrated in Figure 6.6(left). The centers for parts with indexes (i_x, i_y, i_t) are uniformly distributed in the velocity-warped neighborhoods of features and are defined by coordinates $p^{i_x,i_y,i_t} = (x_c, y_c, t_c)$ relative to the center of a feature with $x_c = \alpha\sigma(2i_x - 1 - M)/M$, $y_c = \alpha\sigma(2i_y - 1 - M)/M$ and $t_c = \beta\tau(2i_t - 1 - M)/M$. We choose $\alpha = 3$, $\beta = 3$ and compute local histograms $h^{stg}_{k,i,j}$ and $h^{of}_{k,i,j}$ as above for each of the parts using a Gaussian window functions with centers at p^{i_x,i_y,i_t} and with covariance matrices $\Sigma = (1.5S)^2$. By combining all these histograms into descriptors and using 16 bins

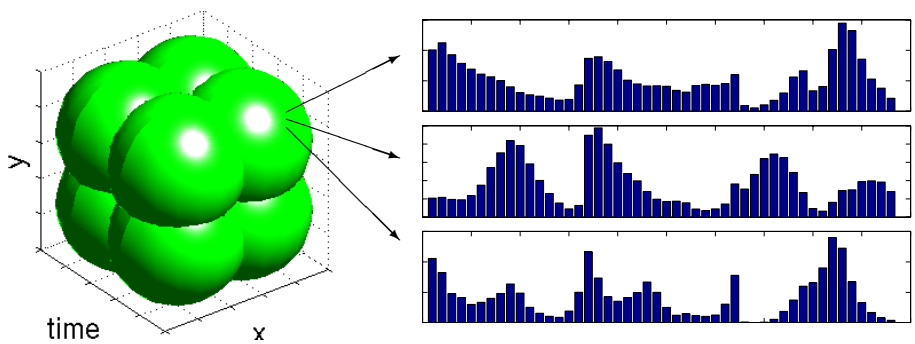


Figure 6.6: Examples of position dependent histograms (right) computed for overlapping Gaussian window functions (left).

per histogram for $M = 2$ and 4 bins per histogram for $M = 3$, we obtain

STG-PD2HIST: STG-based descriptors with $M = 2$ (3456 dimensions)
STG-PD3HIST: STG-based descriptors with $M = 3$ (2916 dimensions)
OF-PD2HIST: OF-based descriptors with $M = 2$ (2304 dimensions)
OF-PD3HIST: OF-based descriptors with $M = 3$ (1944 dimensions)

In spite of the high dimensionality of these descriptors, we see in the next chapter that such position-dependent histograms result in the highest performance compared to other proposed space-time descriptors when evaluated on the problem of recognizing human actions.

6.1.4 Summary

In this section, we have presented twelve different types of local representations for describing information in the scale and velocity invariant neighborhoods of local space-time features. Four of these descriptors are in terms of local N-jets: 2Jets, 4Jets, MS2Jets and MS4Jets; two descriptors are based on PCA: STG-PCA and OF-PCA and six descriptors are defined by local histograms: STG-HIST, OF-HIST, STG-PD2HIST OF-PD2HIST, STG-PD3HIST and OF-PD3HIST. Each of these descriptors will be evaluated against the others experimentally in Chapter 7.

6.2 Dissimilarity measures

Local descriptors have to be complemented with a method for their comparison. In this work, we use three alternative metrics for comparing the descriptors introduced in the previous section. Given the representations of two motion events by vectors d_1 and d_2, we consider the following dissimilarity measures:

- The normalized scalar product:

$$S(d_1, d_2) = 1 - \frac{\sum_i d_1(i) d_2(i)}{\sqrt{\sum_i d_1^2(i)} \sqrt{\sum_i d_2^2(i)}} \tag{6.6}$$

- The Euclidean distance:

$$E(d_1, d_2) = \sum_i (d_1(i) - d_2(i))^2 \tag{6.7}$$

- The χ^2-measure:

$$\chi^2(d_1, d_2) = \sum_i \frac{(d_1(i) - d_2(i))^2}{d_1(i) + d_2(i)} \tag{6.8}$$

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 will be used to compare histogram-based descriptors only: STG-HIST, OF-HIST, STG-PD2HIST, OF-PD2HIST, STG-PD3HIST and OF-PD3HIST.

Using the proposed descriptors and dissimilarity measures, we can now match local events in different image sequences by searching for pairs of features with the lowest dissimilarity of corresponding descriptors. Figure 6.7 presents matched features for sequences with human actions. To generate matches, here we used the STG-PD2HIST descriptor in combination with the normalized scalar product. As can bee seen, matches are found for similar parts (legs, arms and hands) at moments of similar motion. The locality of the descriptors enables us to match similar events in spite of variations in clothing, lighting and backgrounds that substantially change the global appearance of sequences. Due to the local nature of these descriptors, however, some of the matched features correspond to different parts of (different) actions which are difficult to distinguish based on local information only.

Figure 6.7: Examples of matched local space-time features in sequences with human actions. The matches are generated by minimizing the dissimilarity measure (6.6) between STG-PD2HIST descriptors.

6.3 Motion representations

Until now we concerned with the task of representing and comparing individual local events. Given the problem of recognizing a set of motion patterns, such as human actions, reliable recognition may not be possible by matching single pairs of features only. On the other hand, beside the local descriptors, motion events may also have joint properties within a sequence. Such properties could then be explored and used for representing motion patterns as a whole.

For the purpose of recognition, the representations of the data should be as stable and as distinctive as possible. The space-time features in Figure 6.1 would suggest that their relative spatio-temporal arrangement provides distinctive information about the type of underlying actions. Stable modeling of such arrangements, however, is not trivial due to many factors, including the presence of outliers, variations of motion and the appearance of individual patterns and occasional failures of the feature detector. Moreover, the number of partial constellations of n features grows exponentially with n, which creates problems for both learning and recognition. Due to these problems, currently, there exists no general solution for optimal selection of feature arrangements although many interesting ideas and methods in this direction have been developed (Schmid and Mohr, 1997; Amit and Geman, 1999; Mel and Fiser, 2000; Piater, 2001; Lowe, 2001; Rothganger et al., 2003; Leibe and Schiele, 2003; Fergus et al., 2003; Sivic and Zisserman, 2003).

To avoid the issues of stability, we here introduce two rather simple representations that capture only weak joint properties of features in a sequence. In both cases, we represent the sequence by a disordered set of motion events. In the first case, the sets of features are compared using a greedy matching approach as described in Section 6.3.1. In the second case, in Section 6.3.2 we quantize the descriptors of motion events into a finite vocabulary of labels and consider the relative frequency of such labels within a sequence.

6.3.1 Greedy matching

Greedy matching of a set of points, with some distance function, is obtained by repeatedly selecting and removing the pair of points with minimum distance. Greedy matching can be seen as an approximation of minimum weight matching (Reingold and Tarjan, 1981). When applying greedy matching to image sequences, the idea is to enforce as many local matches between feature pairs as possible without enforcing non-trivial constraints on the position of features within a sequence.

Given two sets of space-time features detected in two sequences, the dissimilarity measure is evaluated for each pair of features using the descriptors and metrics as introduced in Sections 6.1 and 6.2 respectively. The pair with minimum dissimilarity is matched and the corresponding features are removed from both sequences. The procedure is repeated until no more feature pairs can be matched, either due to a threshold on dissimilarity or lack of data.

When matching an entire image sequence, the dissimilarity measure is then defined as the sum of the individual dissimilarity measures of the N strongest feature matches. One could also consider adding these measures transformed by a monotonically increasing function. This matching procedure could also be extended in the future to incorporate similarities between *stable* relative properties of local features in image sequences.

6.3.2 Quantization of local events

Whereas local space-time descriptors represent motion events as vectors in a continuous descriptor space, events in image sequences often have a discrete semantic interpretation such as the step of a foot or the waving of a hand. Hence, it is natural to associate events discrete labels within a finite vocabulary of motion events. Such an assignment can be done in an unsupervised manner by standard vector quantization techniques such as k-means clustering (see e.g. (Duda et al., 2001)). In computer vision, vector quantization has been applied previously for representing image textures (Malik et al., 1999) and for local feature-based schemes for image recognition (Weber et al., 2000) and video indexing (Sivic and Zisserman, 2003).

To obtain a vocabulary of motion events, we take all the space-time features detected in a set of training sequences and perform k-means clustering in the space of the associated image descriptors. The number of clusters K is chosen manually and the clustering is repeated several times to obtain a better solution. The resulting clusters with corresponding cluster centers c_i, are then regarded as prototype events and together constitute a vocabulary of K feature labels. Given a new sequence, each of its events with an associated descriptor d_i is assigned with a label l_i by minimizing the distance between d_i and all of the clusters centers c_i, $i = 1, .., K$

$$l_i = \underset{j}{\operatorname{argmin}}\big(D(c_j, d_i)\big), \tag{6.9}$$

where the distance D corresponds to one of the measures in (6.6), (6.7) or (6.8). One example of performing the clustering of motion events represented by 4Jets-descriptors is shown in Figure 6.8. Displayed in the top row are all the features detected in the image sequence which are used as input for k-means clustering. Interestingly, the four most populated clusters, produced by the k-means, correspond to distinct motion events as illustrated by the white circles in Figure 6.8 with labels $c_1, \ldots, c_4$. This probably can be explained by the repetitive nature of the gait pattern that results in many similar events. Spatio-temporal neighborhoods of events corresponding to clusters $c_1, \ldots, c_4$ are shown in Figure 6.9, where we can visually confirm the similarity of neighborhoods within clusters and their dissimilarities between clusters. A classification of events on a test sequence is shown on the bottom row of Figure 6.8. Here, besides a correct classification of events, we can also confirm the invariance of feature detection and descriptors with respect to variations in the spatial scale. Moreover, by using a quantization approach here,

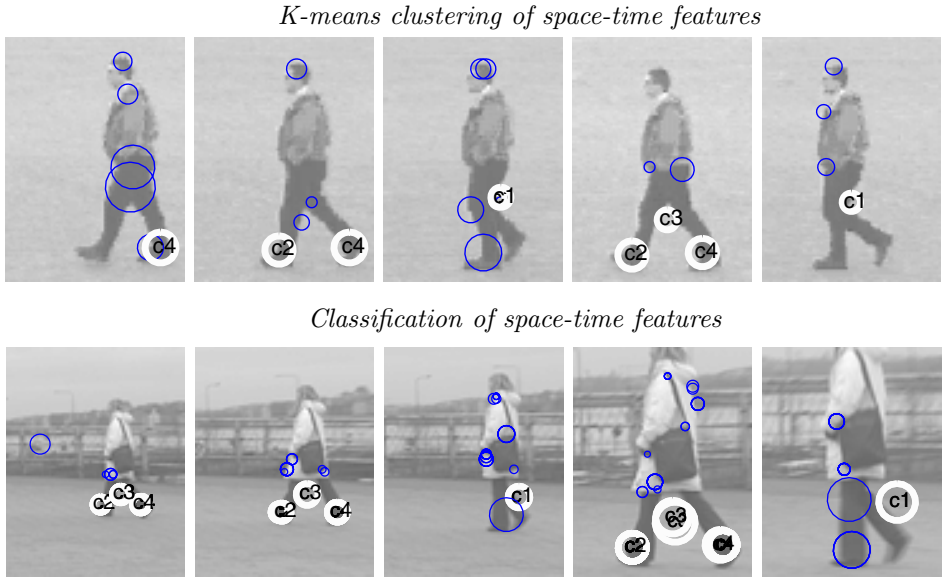


Figure 6.8: Local space-time features detected for sequences of walking people. Top row: the result of clustering space-time features. The labeled points correspond to the four most populated clusters; Bottom row: the result of classifying motion events with respect to the clusters found in the first sequence.

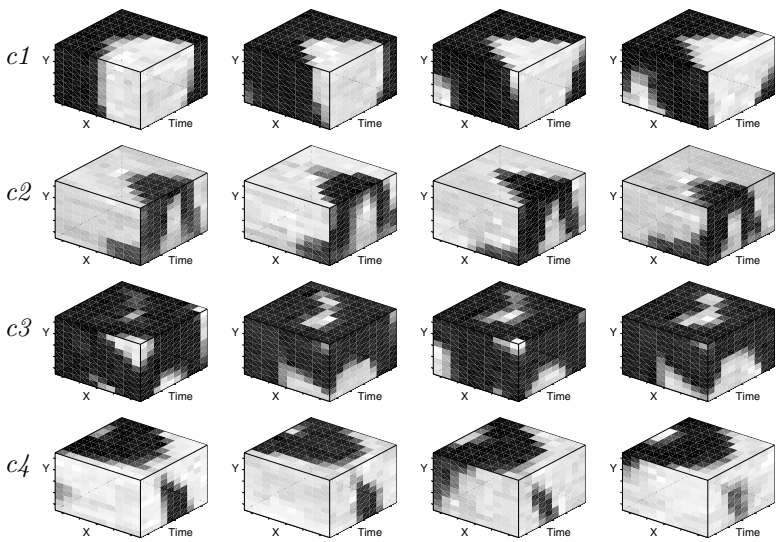


Figure 6.9: Local spatio-temporal neighborhoods of motion events corresponding to the first four most populated clusters obtained from a sequence of a walking person.

most of the noise features can be separated from the motion events originating from the gait pattern.

Given motion events represented by labels $l_1, \ldots, l_K$, we can now compute the frequency of each label for any given sequence. Such histograms of feature labels can be used as a representation of image sequences. The comparison of any two sequences can be achieved by evaluating the distance measure between their histograms according to (6.6), (6.7) or (6.8).

Chapter 7

Evaluation

In previous chapters, we investigated methods for detecting and adapting local space-time features in video sequences as well as methods for describing such features in terms of local spatio-temporal neighborhoods. The intuition behind these methods has been supported by several examples, where we demonstrated the repeatability of the detected features and their stability with respect to scale and velocity transformations (see Figures 5.4, 5.5, 5.6 and 5.8) as well as the possibility of matching similar events using the associated image descriptors (see Figure 6.7).

As discussed earlier, the stability and the discriminative power of local motion events is crucial when using such features for the purpose of matching and motion recognition. Hence, besides a qualitative illustration of the properties of space-time features, it is of interest to perform a quantitative evaluation of local features and to investigate their stability under variations in the data. Such an evaluation with respect to scale and velocity transformations of image sequences will be presented in Sections 7.1 and 7.2. Here we will use a controlled set of scaling and velocity transformations and will analyze their influence on (i) the repeatability of space-time features, (ii) the stability of space-time descriptors and (iii) the recognition performance obtained using feature-based representations. In particular, we will analyze the influence of the mechanisms for scale and velocity adaptation presented in Chapter 5 and will quantify the importance of such mechanisms for both the reliable feature detection and for the subsequent task of motion recognition.

To analyze the discriminative power of space-time features, Section 7.3 will perform a relative evaluation of the local space-time descriptors that were introduced in Chapter 6. All the descriptors will be evaluated on the common task of motion recognition and will be compared with respect to recognition performance. Moreover, as velocity adaptation will turn out to have a strong influence on the recognition performance, the performance of all of the descriptors will be analyzed separately for the cases of velocity-adapted features and for features without velocity adaptation. Finally, as a complement to local feature methods, Section 7.4 will also evaluate the impact of dense velocity adaptation presented earlier in Section 5.3.

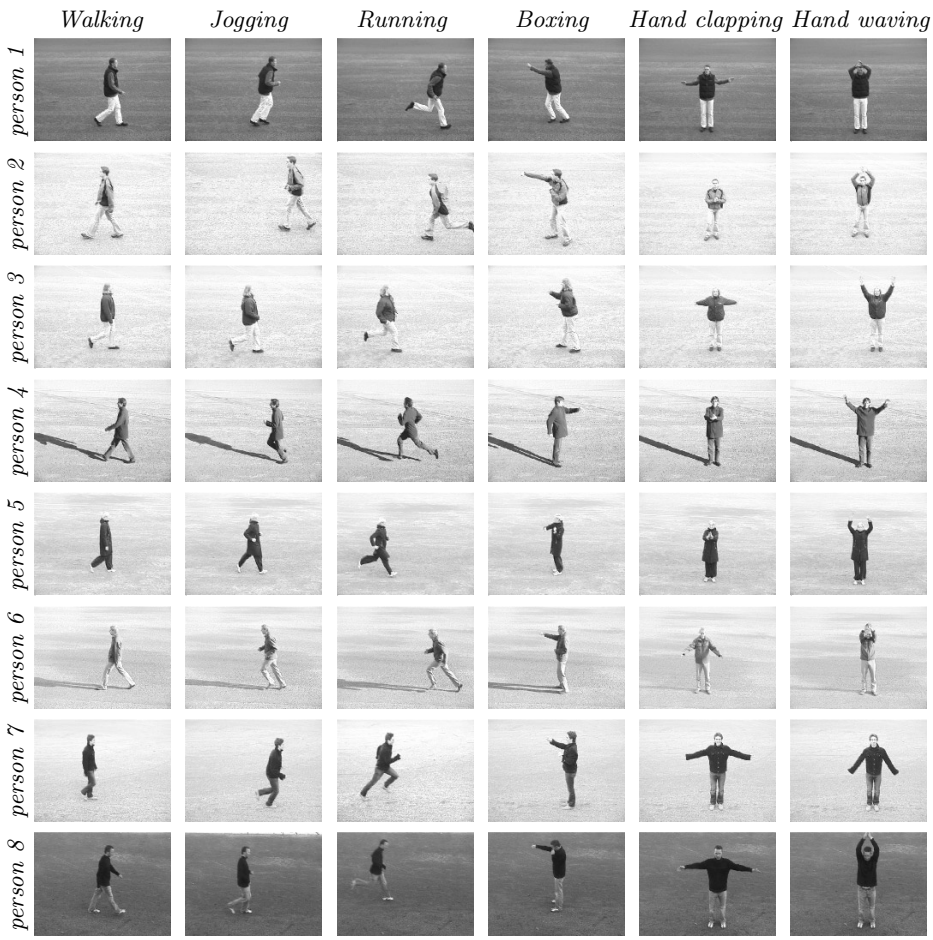


Figure 7.1: Example sequences from a dataset with six human actions performed by eight different persons. Each action in the dataset is repeated four times by each person, where for walking, jogging and running, each action is repeated twice in each directions, i.e. from the left to the right and from the right to the left. The whole dataset contains $6 \times 8 \times 4 = 192$ sequences and is a subset of a larger database presented in (Schüldt et.al., 2004). All sequences were obtained with a stationary camera with the frame rate of 25fps and with subsampling of the spatial resolution to 160×120 pixels.

For the purpose of evaluation, we will in this chapter use real image sequences with six types of human actions performed by eight different people as illustrated in Figure 7.1. To make the evaluation tractable, all the sequences were acquired

in scenes with homogeneous backgrounds and contain only one person performing a single type of action at a time. Given a number of different people and different weather conditions, however, we will consider variations in image sequences caused by changes in the lightning as well as due to individual variations in the cloth and the motion of different subjects.

When performing recognition, the aim will be to recognize the type of actions performed by people in test sequences. For this purpose, the whole dataset will be divided into a training set and a test set with respect to the subjects, such that the same person will not appear in the test and the training sequences simultaneously. As the size of the whole dataset is relatively small (192 sequences), random permutations of the training and the test sets will be considered and the recognition rates will be averaged.

7.1 Stability under scale variations

Variations in the spatial and the temporal extents of motion patterns in image sequences affect the computation of differential image descriptors as well as the subsequent process of motion interpretation. To compensate for such variations when detecting local space-time features, Chapter 5 presented a mechanism for adapting the size of detected image features to the spatio-temporal extents of the underlying image structures. Whereas this mechanism has been proved to be scale-invariant for ideal image signals (see Section 5.1), the performance of the discrete implementation on real image sequences has to be evaluated experimentally. The aim of this section is to perform such an evaluation and to investigate the stability of the detected features as well as the stability of subsequent recognition with respect to scale changes in test sequences. To simplify the evaluation, we will here consider separate variations in the spatial scale in Sections 7.1.1 and in the temporal scale in Section 7.1.2.

7.1.1 Spatial scale

To evaluate the stability of our methods under variations in the spatial scale, we used six image sequences with human actions (walking, boxing and hand-clapping) similar to the sequences in Figure 7.1 but with a higher original spatial resolution. The spatial resolution in these sequences was then gradually reduced by factors $s_i = 2^{-i/2}$, $i = 0, .., 4$ using bilinear interpolation. To evaluate different methods for feature detection, local space-time features were detected as

Hcorr: maxima of the velocity-corrected operator H_c (4.10) for a single, fixed scale level $\sigma^2 = 8$;

HcorrMulti: maxima of H_c for a set of fixed scale levels $\sigma^2 = \{2, 4, 8, 16\}$;

HcorrScVel: maxima of H_c in combination with the iterative scale and velocity adaptation according to the adaptation method presented in Sections 5.1 and 5.2.

The features with space-time positions $p_i = (x_i, y_i, t_i)$ were detected for all sequences and for all five levels of resolution i obtained from subsampling.

Repeatability. When comparing motion representations in terms local features, one necessary requirement is that similar features are detected for corresponding events in different image sequences. Hence, in the case of scale transformations, the features should be repeatedly detected at different levels of resolution. To evaluate the repeatability, we search for corresponding points in the original image sequence and in the corresponding sequences with a reduced spatial resolution. For this purpose, we warp the spatial coordinates of detected features according to $\tilde{p}_i = (\frac{1}{s_i}x_i, \frac{1}{s_i}y_i, t_i)$ and search for pairs of *matching* points $(p_0, \tilde{p}_i)$ with similar positions in the original coordinate frame. The similarity is evaluated by thresholding the Euclidean distance between positions in space-time. Note, that here, p_0 corresponds to features detected in the original image sequence while $\tilde{p}_i$ corresponds to features detected in the transformed sequence with the spatial resolution reduces by a factor s_i. The repeatability is then evaluated as the ratio between the number of matched points and the total number of points in both image sequences.

Figure 7.2(a) illustrates the repeatability evaluated over different scale factors and for different methods of feature detection. As can be seen, the repeatability drops with the decrease of the spatial resolution for all three methods while it decreases faster for features detected at a single and fixed spatial scale (*Hcorr*). Hence, we observe that the results of feature detection depend on the spatial resolution as predicted from the theory in Chapter 3. Moreover, we can observe that the use of scale adaptation results in significantly more stable features (*HcorrScVel*) compared to the features detected without scale adaptation (*Hcorr*). Finally, an exhaustive approach of detecting non-adapted features for a large set of scales (*HcorrMulti*) also results in a high repeatability. This is related to the fact that image representations with descriptors evaluated at all levels of scale are closed under scale transformations. Hence, the high repeatability of *HcorrMulti* is not surprising, however, such a representation can be expected to be redundant and problematic for the purpose of matching. This expectation will indeed be confirmed by the experiments below.

Stability of descriptors. The repeatability of local space-time features is a necessary but not sufficient condition for reliable matching of corresponding events in image sequences. When using local image descriptors for the purpose of matching, such descriptors should also be stable with respect to transformations in the data. To evaluate the stability of local space-time descriptors with respect to changes in the spatial scale, we computed *4Jet*-descriptors (see Section 6.1.1) and evaluated the distance between the descriptors of corresponding features. The pairs of corresponding features were obtained as above by warping and comparing the positions of features in the original and in the scaled version of the same sequence. An average Euclidean distance between corresponding *Jet*-descriptors was then evaluated for different levels of resolution and for different methods of feature detection.

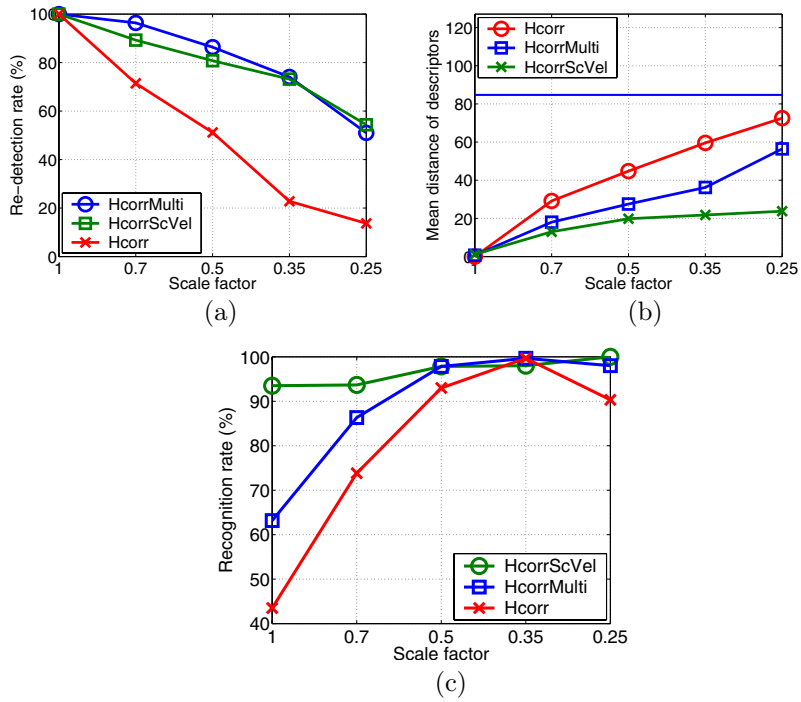


Figure 7.2: Evaluation of space-time features under changes in spatial scale. (a): Repeatability of local features for different spatial resolutions and for different methods of feature detection (see text for explanations); (b): mean Euclidean distance between *4Jet*-descriptors of corresponding features at different levels of resolution. The horizontal line in the graph corresponds to the mean distance between all the descriptors in the sequences; (c): average performance of action recognition subject to scale variations in the data and different methods of feature detection.

As can be seen from Figure 7.2(b), the lowest distance is achieved for descriptors corresponding to adapted features (*HcorrScVel*). The stability of descriptors for non-adapted features is lower while single-scale features (*Hcorr*) have the lowest performance. From these results, we can confirm that scale-adaptation of space-time features contributes strongly to the stability of the corresponding descriptors. Moreover, the relatively poor stability of (*HcorrMulti*) might be caused by the presence of multiple responses of such features at multiple scales, which increases the ambiguity of matching.

Recognition performance. Finally, we can note that reliable matching of local space-time features cannot be guaranteed unless the the corresponding local image descriptors are *discriminative*. To evaluate the discriminative power of local

descriptors, one can either maximize the distance between non-corresponding descriptors or evaluate the performance of motion recognition based on these descriptors. Whereas the first approach would require manual labeling of non-corresponding features, we here choose the second approach and consider the problem of recognizing human actions with each action represented in terms of local space-time features as described in Section 6.3.1. After all, recognizing motion patterns in terms of local space-time features is our general goal and it is interesting to evaluate the performance of such recognition scheme under scale variations in the data.

To perform recognition, we use a simple nearest neighbor strategy and estimate the *type* of an action in the test sequence f^{test} by the type of an action in a training sequence f_j^{train} where f_j^{train} is most similar to f^{test} among all sequences $j = 1, .., m$ in the training set. The similarity of two image sequences is evaluated with greedy matching of local space-time features (see Section 6.3.1) using a combination of *4Jet*-descriptors and a Euclidean distance function. Moreover, for the test set, we use the same image sequences as in the two previous evaluations above, while for training we use subsets of sequences corresponding to five random subjects in Figure 7.1. The results of recognition are then averaged over all sequences in the test set and for 100 recognition experiments using different training sets.

The results of recognition for different scale transformations and different methods of feature detection are illustrated in Figure 7.2(c). Notably, the recognition performance for adapted features is almost independent of the scale transformations in the data. On the contrary, the performance for non-adapted features drops significantly with variations in scale and it follows that neither the *Hcorr* nor the *HcorrMulti* methods are stable enough under variations of the spatial scale. The peak in recognition performance is obtained for $s_i = 2^{-3/2}$, since this level of resolution in test sequences roughly corresponded to the level of resolution in training sequences. Hence, we conclude that motion recognition using non-adapted local space-time features is possible if the relation between the spatial size of image patterns in the test sequences and in the training sequences is known in advance. In other situations, local scale adaptation of space-time features is crucial in order to obtain stable results of recognition.

Remarks. In the present evaluation, we analyzed the scale dependency of various methods using artificially subsampled image sequences. This method is suboptimal since the subsampling constitutes only an approximation to real scale variations in image sequences and may introduce artifacts. An alternative approach would be to record image sequences using multiple cameras simultaneously with different external and/or internal camera parameters. This approach was not chosen here, due to its practical difficulty and since it would also require additional estimation of spatio-temporal correspondences between image sequences.

Due to the computational complexity, the evaluation was here restricted to local descriptors in terms of *4Jets* only. Since the methods for scale and velocity adaptation do not depend on the choice of descriptors, we believe that similar results would be obtained using the other types of image descriptors defined in Chapter 6.

7.1.2 Temporal scale

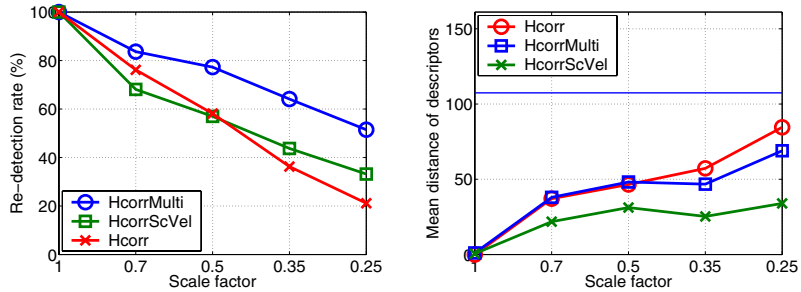
The notion of temporal scale is related to the duration (or temporal extent) of events in time. Variations of temporal extents in image sequences may occur due to within-class variations of motion patterns (e.g. slow vs. fast walking) or due to different sampling rates of the camera. The frame rate of video cameras, however, is usually standardized (25fps. for PAL and 30fps. for NTSC). Moreover, in the recorded set of human actions used in this work, we did not find any significant variations of actions in terms of the frequency, although no specific instruction were given to the subjects performing the actions in the database. Hence, variations in the temporal scale appear to be more restricted than variations in the spatial scale and it follows that the need of automatic scale adaptation in time is less important at least for the set of actions concerned here. However, as adaptation of the temporal scale might be important in other situations, we will here evaluate the repeatability of space-time features as well as the stability of the corresponding descriptors subject to the variations in the temporal resolution of the data.

For the purpose of evaluation, we recorded image sequences of walking people where we advised the subjects to walk as slow as possible. Then, we subsampled the sequences in the temporal dimension using linear interpolation and scale factors $s_i = 2^{-i/2}$, $i = 0, .., 4$. As in the case for spatial scale, we detected space-time features for all levels of temporal resolutions i and warped the temporal coordinates of the features back to the original coordinate frame according to $\tilde{p}_i = (x_i, y_i, \frac{1}{s_i} t_i)$. The features were detected either for a single temporal scale $\tau^2 = 4$ without adaptation (*Hcorr*), for multiple temporal scales $\tau^2 = \{2, 4, 8, 16\}$ without adaptation (*HcorrMulti*) or when using iterative adaptation with respect to both scales and velocities (*HcorrScVel*). The repeatability of the features and the stability of the corresponding *4Jet*-descriptors were evaluated as in the previous section.

The results of the evaluation are illustrated in Figure 7.3. As can be seen, the use of *HcorrMulti* features resulted in the best re-detection rate. The use of scale and velocity adapted features, on the other hand, resulted in a relatively poor re-detection. This behavior was rather unexpected. Low repeatability of adapted features could possibly be explained by convergence problems of the adaptation process. Hence, this issue might require more investigation in the future. Despite the low repeatability, the descriptors of the adapted features appear to be the most stable ones as illustrated in Figure 7.3(b). Hence, the performance of matching under variations in the temporal scale is still likely to be better for the *HcorrScVel* features than for the features without scale adaptation (*HcorrMulti*, *Hcorr*).

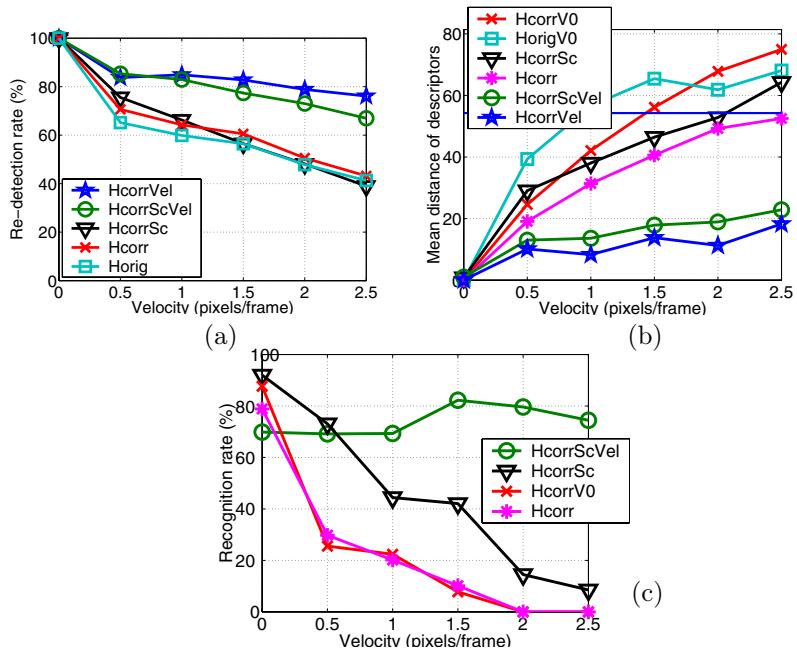
7.2 Stability under velocity variations

Constant motion between the camera and the observed patterns results in a Galilean (or velocity) transformation of the image sequence which affects computation of image descriptors. In this section, we analyze this dependency when detecting local space-time features and when using such features and the corresponding



(a) (b)

Figure 7.3: Evaluation of space-time features under changes in temporal scale. (a): Repeatability of features for different temporal resolutions and for different methods of feature detection; (b): mean Euclidean distance between *4Jet*-descriptors of corresponding features at different levels of resolution. The horizontal line in the graph corresponds to the mean distance between all descriptors in the sequences;



(a) (b) (c)

Figure 7.4: Evaluation of space-time features under Galilean transformations. (a): Repeatability of local space-time features for different values of velocity; (b): mean Euclidean distance between *4Jet*-descriptors of corresponding features in the original and in the velocity-warped sequences. The horizontal line in the graph corresponds to the mean distance between all descriptors in the sequences; (c): average performance of action recognition subject to velocity transformations in the data.

image descriptors for the task of motion recognition. Similar to the analysis of spatial scale, we here evaluate the repeatability of features, the stability of local space-time descriptors and the performance of recognition under different Galilean transformations of image sequences and for different methods of feature detection.

For the purpose of evaluation, we used thirteen sequences with human actions taken with a stationary camera and transformed according to Galilean transformations $G(v_x, v_y)$ (3.36) with $v_x = \{.5, 1.0, 1.5, 2.0, 2.5\}$ and $v_y = 0$. In practice this was achieved by warping the original image sequences using bilinear interpolation and in this way simulating different velocities of the camera. To evaluate different methods of feature detection, we detected local space-time features using:

Horig: maxima of the space-time operator H (4.3) without neither scale nor velocity adaptation;

Hcorr: maxima of the velocity-corrected operator H_c (4.10) without adaptation;

HcorrSc: maxima of H_c with iterative adaptation of the spatio-temporal scales only according to Section 5.1;

HcorrVel: maxima of H_c with iterative adaptation of the velocity only according to Section 5.2;

HcorrScVel: maxima of H_c in combination with iterative scale and velocity adaptation.

The purpose of using these methods was to analyze the influence of the iterative velocity adaptation and velocity correction on the detection of space-time features. To compare the features detected for different values of velocity transformation, we warped the positions of the features $p = (x, y, t)$ detected for $G(v_x, v_y)$ back to the original coordinate frame according to $\tilde{p} = G^{-1}p = (x - v_x t, y - v_y t, t)$.

Repeatability. To evaluate the repeatability, we matched features at similar positions detected in the original image sequence and in the corresponding Galilei-transformed sequence. The repeatability rate was computed as a ratio between the number of matched features and the total number of features in both sequences. Figure 7.4 illustrates the repeatability averaged over all sequences in the test set and computed for different velocity transformations and for different methods of feature detection. As can be seen, the curves cluster into two groups corresponding to high re-detection rates for features *with* iterative velocity adaptation and to lower re-detection rates for features *without* velocity adaptation. Hence, we confirm that iterative velocity adaptation is an essential mechanism for stable detection of features under velocity transformations in the data. By comparing the results of *Horig* and *Hcorr*, we also observe a slightly better repeatability of features detected using the velocity-corrected operator (*Hcorr*). To restrict the number of evaluated detectors, we will in the following mainly use the velocity-corrected features when evaluating the stability of image descriptors and the performance of recognition.

Stability of descriptors. Similar to the case of scale transformations, velocity transformations affect the computation of local space-time descriptors and, hence, subsequent matching and recognition of motion patterns. To compensate for velocity transformations, filter kernels of image descriptors can be adapted to estimated velocity values using either iterative velocity adaptation as proposed in Section 5.2 or "one-step" adaptation corresponding to the non-iterative optic flow estimation according to Section 3.4.1. The first approach is truly invariant under velocity transformations and is natural when computing image descriptors for velocity-adapted features (*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 features detected without iterative velocity adaptation.

Here, we evaluate the effect of velocity adaptation of image descriptors in combination with different methods for feature detection and compare the stability of concerned descriptors under velocity transformations. In particular, we compute 4Jet-descriptors using

(i) filter kernels with iterative velocity adaptation for velocity-adapted features *HcorrVel, HcorrScVel*;

(ii) filter kernels with one-step velocity adaptation for features *HcorrSc, Hcorr*;

(iii) separable filter kernels without velocity adaptation for non-adapted features *Horig, Hcorr* here denoted as *HorigV0, HcorrV0*.

The stability of the descriptors is evaluated by computing the average Euclidean distance between descriptors of corresponding features. Pairs of corresponding features are determined as for the repeatability test above, by comparing positions of features in the original sequence and in the corresponding velocity-transformed sequences. The results of such an evaluation are illustrated in Figure 7.4(b). As can be seen, the best performance in stability is achieved for features and descriptors with iterative velocity adaptation (*HcorrVel, HcorrScVel*). The performance of descriptors with approximative velocity adaptation (*HcorrSc, Hcorr*) is better than for descriptors without velocity adaptation (*HorigV0, HcorrV0*), however, it is significantly worse than for the case involving iterative velocity adaptation. Hence, we conclude that the iterative velocity adaptation is crucial for obtaining stability under velocity transformations.

Recognition performance. As mentioned previously, besides stability of image descriptors and repeatability of feature detection, reliable matching and motion recognition also requires space-time features to be discriminative. Here, we evaluate the discriminative power of velocity-adapted features and the stability of recognition performance under Galilean transformations. To evaluate recognition performance, we use a similar method as used previously in Section 7.1.1. Hence, 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. For the test set, we use the same sequences as in the evaluation of repeatability and stability above, while for training we use subsets of sequences corresponding to five random subjects in Figure 7.1. The recognition rates are then averaged over all sequences in the test set and for 100 recognition experiments using different training subsets with six actions performed by five different subjects each.

Figure 7.4(c) illustrates the results of recognition for different velocity transformations and for different types of space-time features. As can be seen, the only method that is stable under velocity transformations is the one with iterative velocity adaptation of the detection and descriptors (*HcorrScVel*). Notably, for features detected without iterative adaptation, the use of approximate velocity-adapted descriptors (*Hcorr*) does not result in the better performance than the use of non-adapted descriptors (*HcorrV0*).

Another interesting observation is that 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 test and training sequences coincide. 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.

Finally, we can note that the use of the scale-adaptive detection of space-time features in combination with approximative velocity adaptation of descriptors *HcorrSc* results in a similar recognition performance as for *HcorrScVel* at $v_x = .5$. Given that the typical velocity of a person in our sequences is $v_x \approx 1$ (one pixel per frame), the velocity $v_x = .5$ can be considered as a reasonable error bound on the velocity estimate with a hand-held camera when trying for example to follow a walking person. Hence, since *HcorrSc* gives better performance than *HcorrScVel* in the interval $v_x \in (-.5, .5)$, we can conclude that in cases when the relative velocity between the motion pattern and the camera is approximately known, it might be an advantage of *not* using velocity adaptation. However, if the relative velocity of the motion pattern with respect to the camera is not known, the use of invariant velocity adaptation for motion recognition is essential.

Remarks. The results and the conclusions in this section might depend on the particular types of human actions chosen for the experiments. Since these results, however, are well explained by the theory of the used methods, we believe that similar results would be obtained in other situations as well.

7.3 Evaluation of local motion descriptors

In Chapter 6, we presented a number of image descriptors for describing local information in spatio-temporal neighborhoods of local space-time features. The purpose of constructing these descriptors was to enable the discrimination of different motion events in image sequences as well as to enable the matching of similar events in different image sequences. The descriptors were formulated either in terms of optic flow or spatio-temporal derivatives and, hence, capture different aspects of local image structures. Moreover, local measurements were combined into different types of descriptors using N-jets, histograms or principal component analysis. Although some predictions about the performance of these descriptors could be made in advance (e.g. lower order N-jets can be expected to be less discriminative than higher order N-jets), the performance of descriptors is likely to be dependent on many other factors whose influence may be difficult to predict (e.g. the precision of the estimated positions of space-time features). In this context, it is important to perform an experimental evaluation and to compare the performance of different descriptors on a common benchmark problem.

To evaluate the performance of local motion descriptors, we performed a set of recognition experiments where the aim was to recognize the types of human actions in the example sequences in Figure 7.1. We performed leave-X-out experiments, where all image sequences corresponding to X random subjects were removed from the database to be used as testing data, while the remaining image sequences were used as training data. Then, for each image sequence in the test set, a best match was determined among all the image sequences in the training set. A match was regarded as correct if the activity of the training sequence agreed with the activity of the image sequence in the test set. To compare different sequences, we used greedy matching as described in Section 6.3.1. Moreover, as velocity adaptation was found to influence the performance of recognition in previous section, we performed separate experiments where we used either scale-adapted features or features detected with the iterative adaptation of scales and velocities.

The results of these experiments when using different types of local motion descriptors as well as different dissimilarity measures are shown in Figure 7.5. Due to the large number of tested descriptors, we here only show the descriptor within each descriptor class that maximizes the recognition performance within its class. The first observation from the results is that the recognition rate is relatively high for most of the descriptors, while it is highest (96.5%) for OF-PD2HIST descriptor in combination with the Euclidean distance measure. Moreover, independently of the dissimilarity measure and of the type of local measurements (STG or OF), the position-dependent histograms result in the best performance when using scale-adapted features (see Figure 7.5(left)). This result coincides with a similar result in the spatial domain, where the $SIFT$-descriptor (Lowe, 1999), which is conceptually similar to the position-dependent histograms used here, was found to outperform other local image descriptors when evaluated on the task of matching local features in static images (Mikolajczyk and Schmid, 2003).

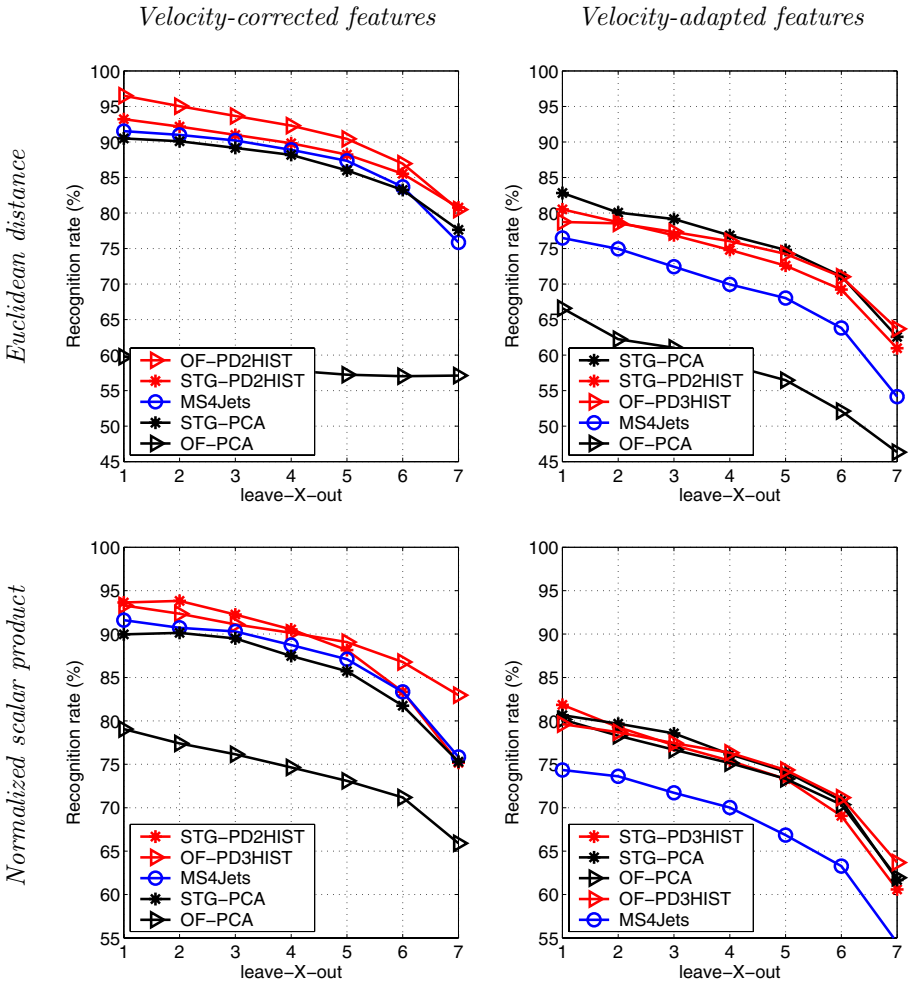


Figure 7.5: Results of human actions recognition using the dataset in Figure 7.1. Recognition is reported in terms of leave-X-out experiments (see the text) when using either (top) The Euclidean distance or (bottom) The normalized scalar product for feature comparison. (Left column): Recognition rates obtained for scale-adapted features with complementary velocity correction; (Right column): Recognition rates obtained for scale and velocity adapted features. All recognition results are averaged over a large number (500) of random perturbations of the database. The results are shown only for the descriptor with best performance within its descriptor class (e.g. MS4Jets is chosen among MS4Jets, MS2Jets, 4Jets and 2Jets).

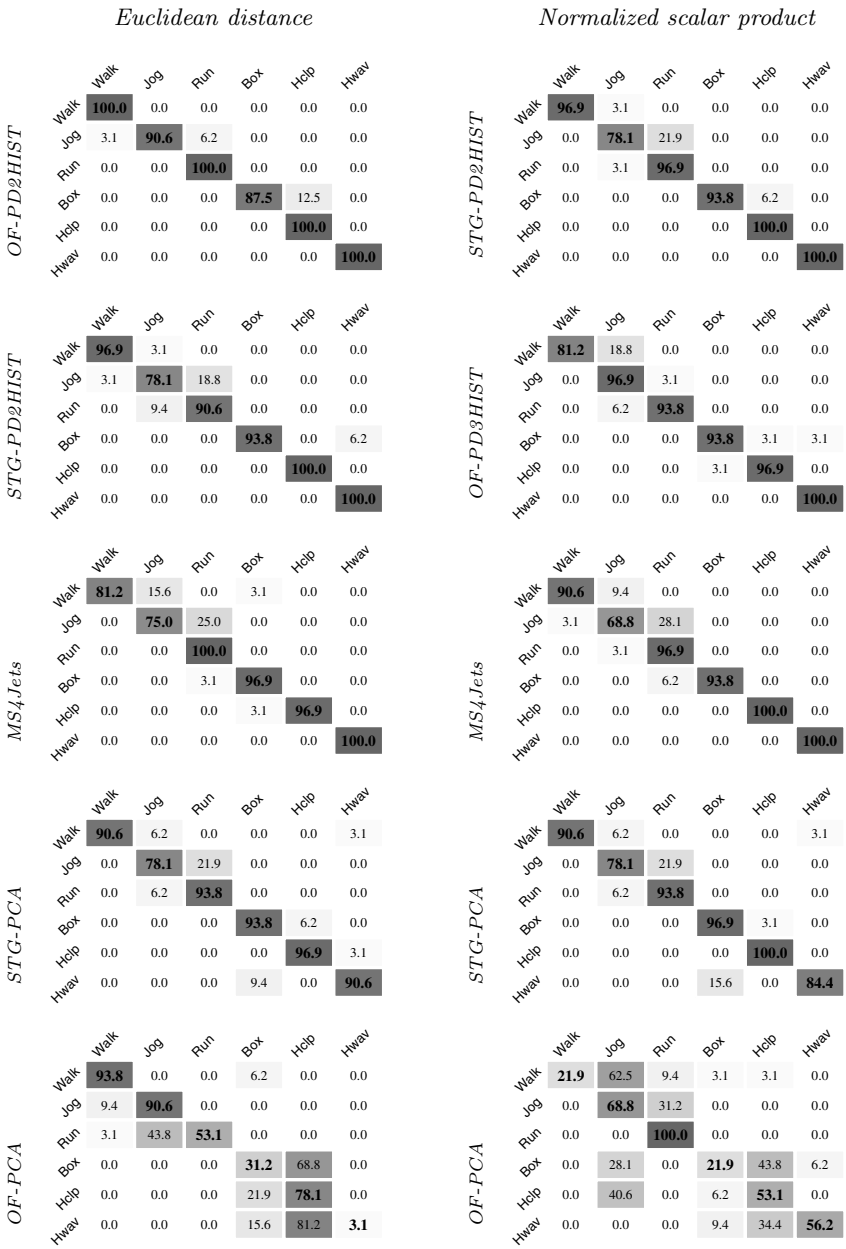


Figure 7.6: Confusion matrices when recognizing human actions (static camera) in leave-one-out experiments using scale-adapted features. The matrices represent the results obtained for the different types of descriptors in Figure 7.5(left). Value $(i, j) = x$ means that the action i was recognized as the action j in x% of all cases.

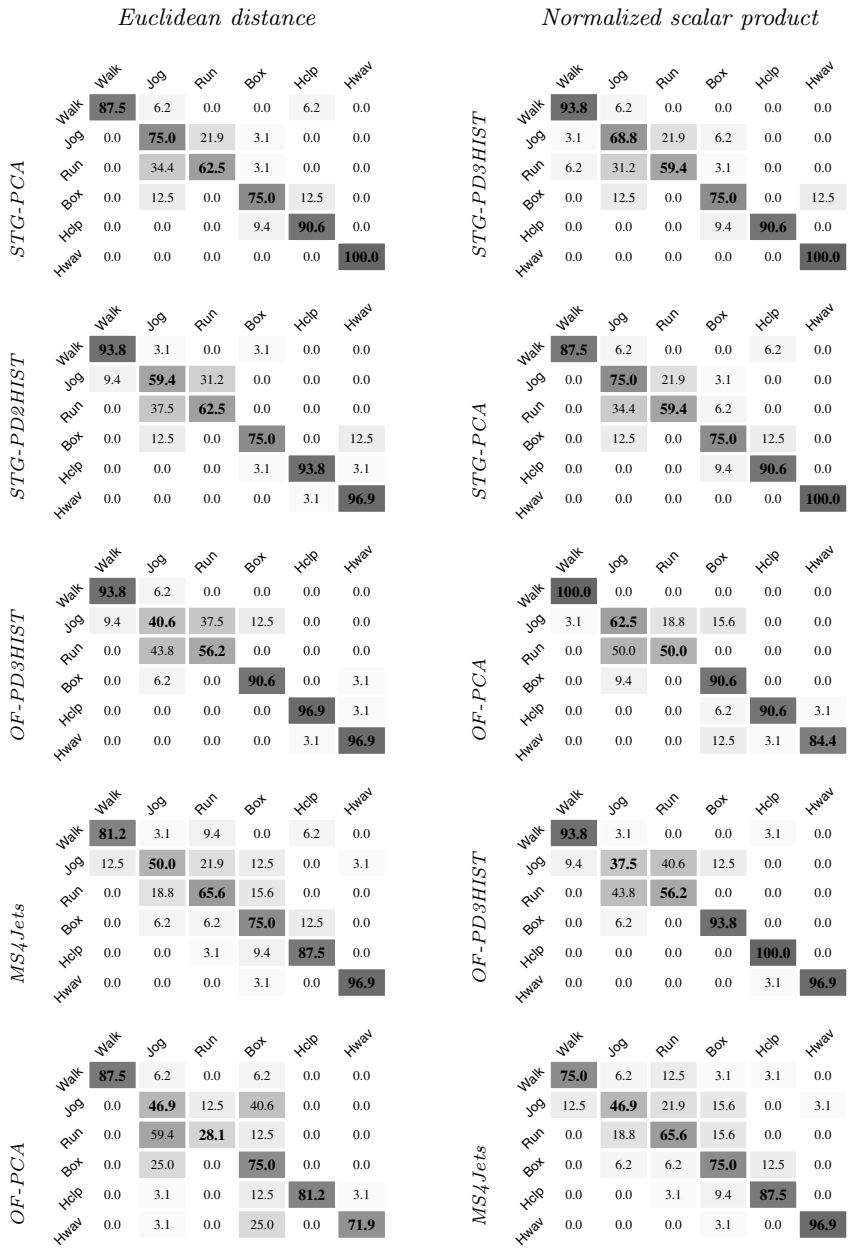


Figure 7.7: Confusion matrices when recognizing human actions (static camera) in leave-one-out experiments using scale and velocity adapted features. The matrices represent the results obtained for the descriptors in Figure 7.5(right). Value $(i, j) = x$ means that the action i was recognized as the action j in $x\%$ of all cases.

Since the camera was stationary in both the training sequences and in the test sequences, the recognition performance for velocity and scale adapted features was slightly worse than for features with only scale adaptation. Note, that this behavior is consistent with the velocity-adaptation experiments in Section 7.2.

Concerning the other descriptors, we can observe that optic flow (OF) in combination with PCA does not perform well in most of the experiments. Moreover, descriptors based on spatio-temporal gradients (STG) outperforms descriptors based on optic flow (OF) for the case of velocity-adapted features (see Figure 7.5(right)). This can be explained by the fact that STG-based descriptors capture both the motion and the shape of local image structures while the OF-descriptors are only sensitive to the motion which discriminative power is reduced when using velocity-adapted descriptors. Finally, we can note that the performance of recognition for all methods is rather stable for different subdivisions of the dataset into X test subjects and $(8-X)$ training subjects. This can be regarded as an interesting positive result, considering that the dataset in Figure 7.1 contains quite strong variations of lightning conditions as well as individual cloth and motion variations.

To investigate the errors of the presented methods, Figures 7.6 and 7.7 show confusion matrices corresponding to each one of the methods presented in Figure 7.5. From most of the confusion matrices, we can observe an almost perfect separation between the hand and the leg actions. Moreover, independently of the used descriptors, the dissimilarity measures and the methods used for feature detection, most of the errors occur due to confusion between jogging and running actions as well as between walking and jogging actions. This result is rather intuitive since these combinations of actions can be subjectively regarded as most similar ones when considering all combinations of the six studied actions. As expected, the confusion between jogging and running actions increases when using velocity-adapted features (see Figure 7.7). Finally, we can note that the jogging actions of some of the people in our dataset were found quite similar to the running actions of other people. This gives further explanations to the confusions in Figures 7.6 and 7.7.

7.3.1 Jet-based descriptors

To analyze the performance of N-Jet descriptors, Figure 7.8 presents recognition results for N-Jets of different orders and for jets computed either at a single scale or at multiple scales according to Section 6.1.1. From the results, we observe a consistent relative performance of different N-Jet descriptors for both types of dissimilarity measures as well as for both feature detectors. The use of multiple scales and the use of higher order derivatives seem to give advantages for recognition. The use of multiple scales, however, appears to be more important than the use of higher order derivatives. The relatively good performance of the N-Jets in comparison to other descriptors in Figure 7.5, as well as the low dimensionality and a small number of parameters associated with these descriptors make the N-Jets an attractive choice for local motion descriptors.

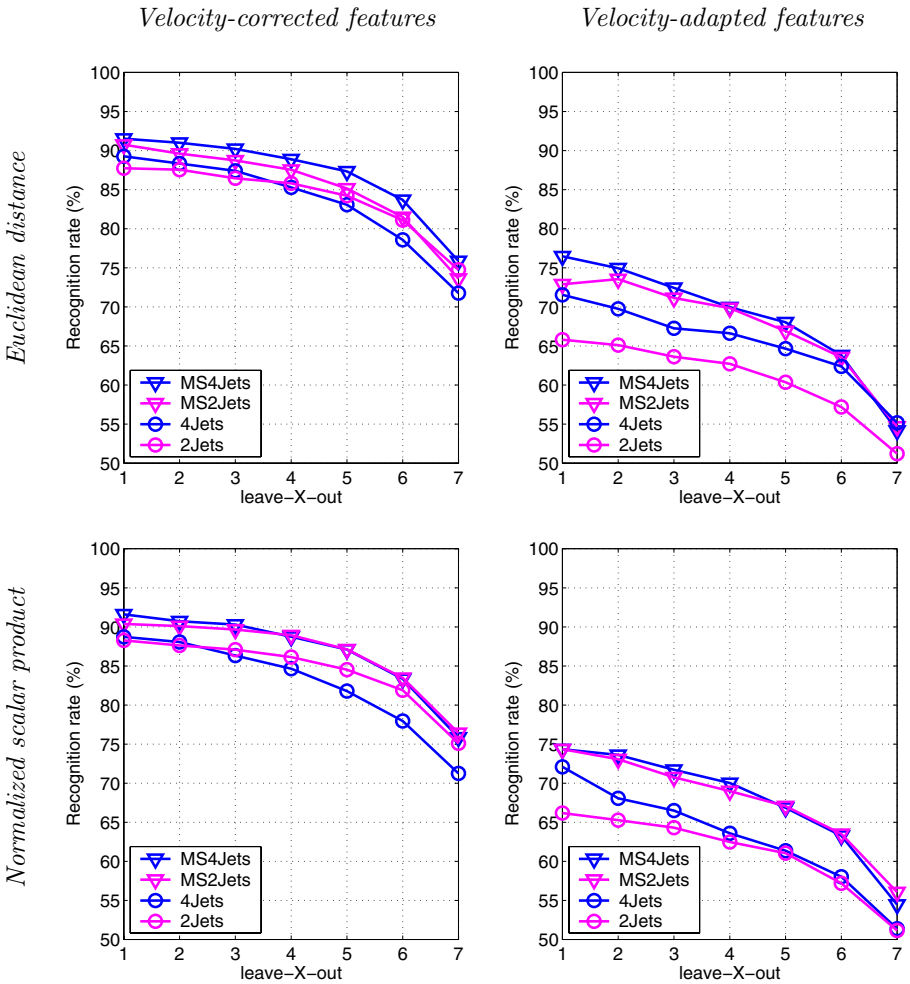


Figure 7.8: Results of recognizing human actions in leave-X-out experiments using different types of N-Jet descriptors. The results are shown for N-Jets of order two and four as well as for multi-scale and single-scale jets. All results are averaged over 500 random perturbations of the database.

7.3.2 Histogram-based descriptors

In Figure 7.9, we analyze the relative performance of different histogram-based descriptors. The histograms were accumulated either from the components of spatio-temporal gradients or from the components of optic flow according to Sec-

tion 6.1.3. The results for position-independent histograms are also compared to the results of position-dependent histograms with the number of position-dependent bins $M = \{2, 3\}$ along each dimension in space-time. Besides the Euclidean distance and the normalized scalar product measure, the results are also evaluated for χ^2-measure which has been successfully used in histogram-based recognition schemes previously (Schiele and Crowley, 2000). As can be seen from the results, position-dependent histograms clearly outperform position-independent histograms for all three types of dissimilarity measures, for both types of feature detectors as well as for image measurements in terms of OF and STG. The comparison of descriptors with different number of position-dependent bins (PD2HIST, PD3HIST) does not indicate an advantage of using $M = 3$ compared to $M = 2$. This result, however, should be taken with caution since it might depend on the number of bins used to accumulation the responses of gradient vectors or optic flow vectors.

7.3.3 Comparison to other methods

To conclude the evaluation of local motion descriptors, we will compare the obtained results with the performance of other related methods evaluated on the same dataset. At first, we consider a method in terms of *spatial* local features detected as maxima of Harris operator (3.31) for every fourth frame in the image sequences. The obtained features are adapted with respect to the spatial scale using the approach by (Mikolajczyk and Schmid, 2001) 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 the same way as the local space-time features. Such a method is very similar to ours, except that it does not use any temporal information neither for the feature detection nor for the computation of local descriptors. The main motivation of comparing such an approach was to confirm that the temporal information captured by space-time features is actually *necessary* 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) in Figure 7.10, we can confirm that the performance of the local spatial features is close to chance and that the use of temporal information in our method is indeed justified.

Two other methods which we use 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

(i) 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-MS)

(ii) absolute values of the same components $|L_x|/\|\nabla L\|$, $|L_y|/\|\nabla L\|$, $|L_t|/\|\nabla L\|$ at multiple temporal scales only (Global-STG-HIST-ZI)

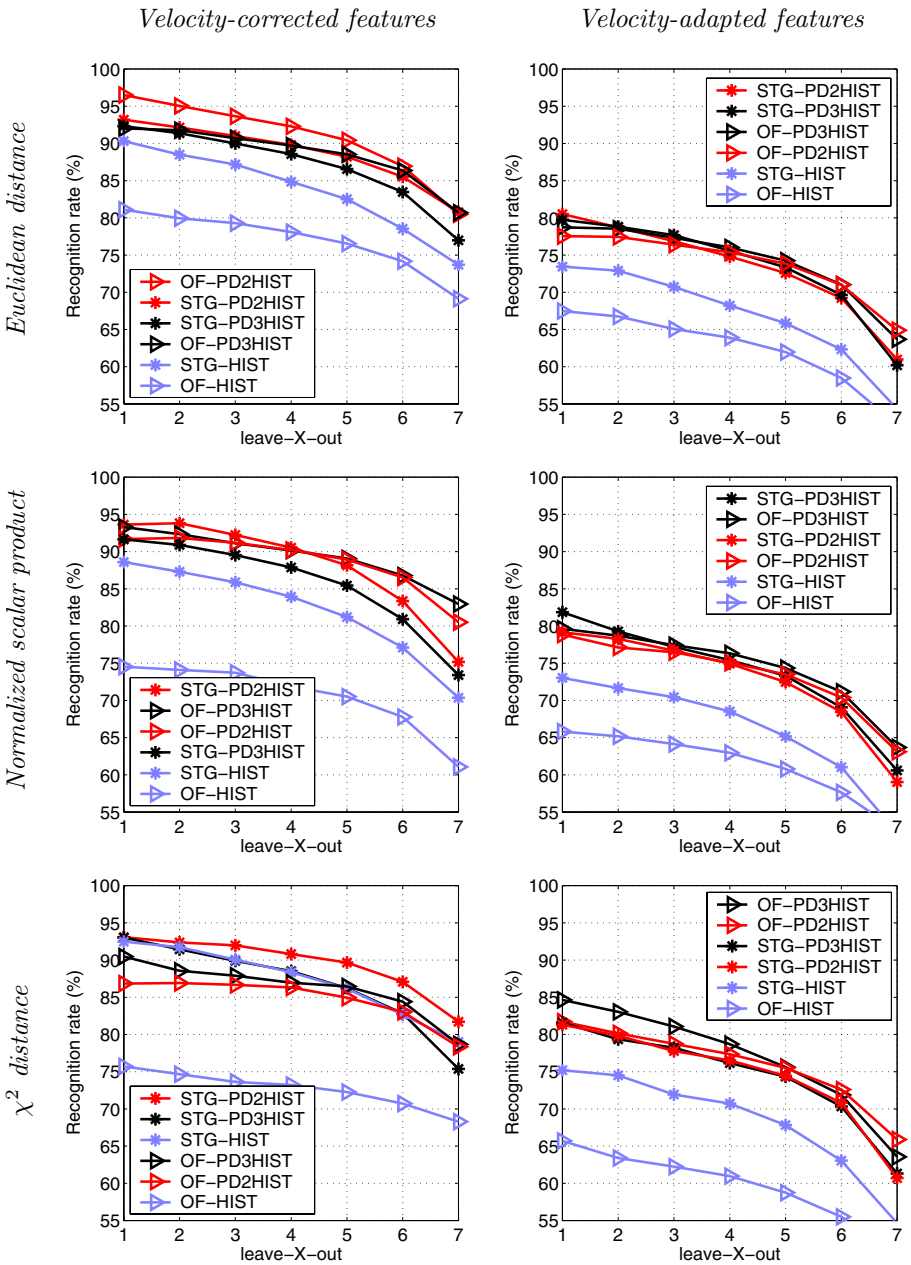


Figure 7.9: Results of recognizing human actions in leave-X-out experiments using different types of histogram-based descriptors and different dissimilarity measures. The results are averaged over 500 random perturbations of the database.

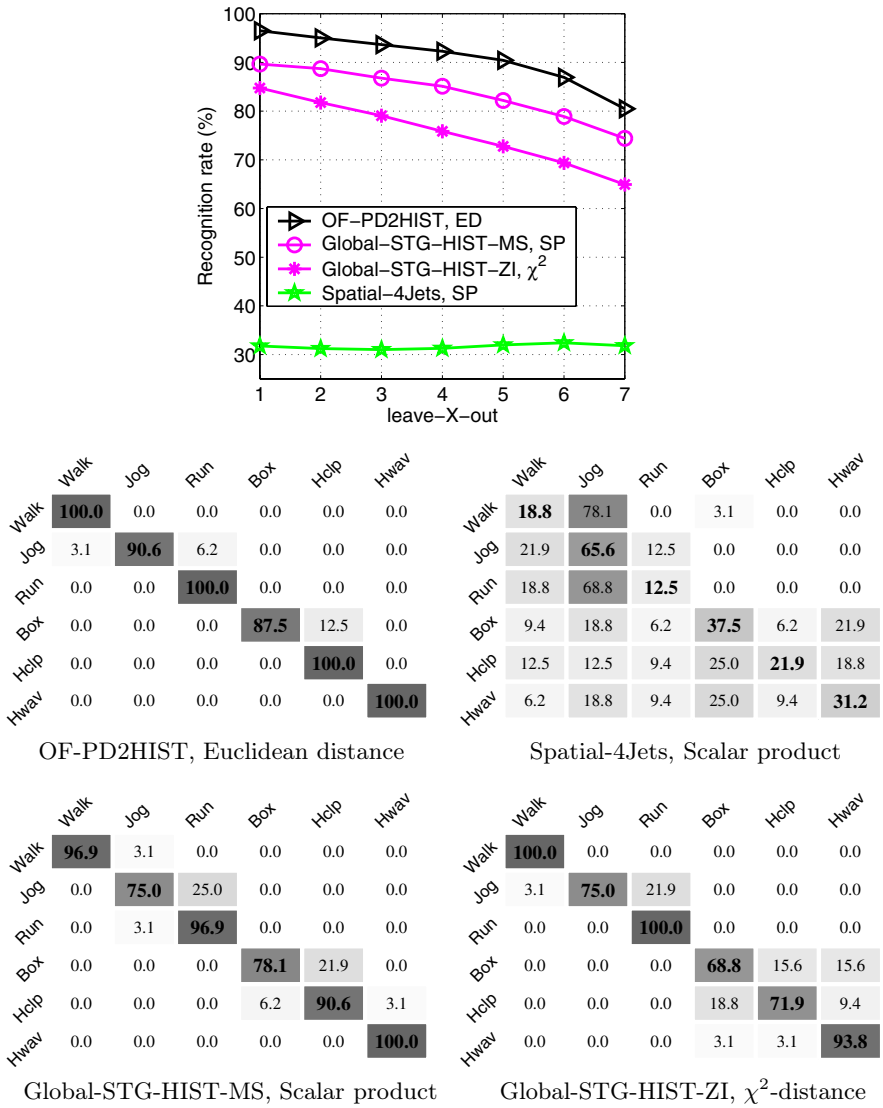


Figure 7.10: Comparison of recognition performance using local space-time features (OF-PD2HIST) and other methods in terms of (Spatial-4Jets): spatial interest points with fourth-order spatial Jets descriptors; (Global-STG-HIST-MS): Global histograms of normalized spatio-temporal gradients computed at multiple spatial and temporal scales; (Global-STG-HIST-ZI): Global histograms of normalized spatio-temporal gradients according to (Zelnik-Manor and Irani, 2001). The results are shown as leave-X-out plots and as confusion matrices corresponding to leave-one-out experiments for each one of the four methods.

The second approach corresponds to the method by (Zelnik-Manor and Irani, 2001) which was successfully applied to a similar task of action recognition in previous work. The first approach is an extension of the second one, where we additionally take the direction of spatio-temporal gradients at multiple spatial scales into account. For the purpose of recognition, we computed histograms for all sequences in the dataset and used nearest neighbor classification where the dissimilarity between the histograms was evaluated according to the measures in Section 6.2. The results for both methods optimized over three dissimilarity measures are shown in Figure 7.10. As can be seen, both methods perform rather well with the better performance for Global-STG-HIST-MS. The local feature method (OF-PD2HIST), however, results in the best performance for the methods compared here.

To conclude, the presented results confirmed that it is possible to perform spatio-temporal recognition based on local space-time features and corresponding local image descriptors. Moreover, the use of local descriptors in terms of position-dependent histograms has shown the best performance, which is consistent with related results concerning local descriptors in the spatial domain. Whereas this evaluation has been performed in simplified scenarios with homogeneous backgrounds, Section 8.1 will demonstrate that high recognition performance using local space-time features can also be achieved for scenes with complex and non-stationary backgrounds.

7.4 Evaluation of dense velocity adaptation

As argued in Section 5.3, local space-time features capture only partial information in the image data and might therefore not be insufficient for recognition in situations dominated by locally constant motion or spatial one-dimensional image structures. From the space-time plots and the corresponding features detected for human actions in Figure 6.1, we can indeed observe that many characteristic parts of the motion patterns, such as the sweeping of the hands in a hand-waving example, are not captured by local features and, hence, do not contribute to event-based motion recognition. One problem with such types of image structures is that their space-time positions are hard to localize reliably. This, in turn, increases the ambiguity of matching and decreases the performance of subsequent recognition.

One conceptually simple way to avoid the ambiguity problem and to capture all local variations in the data consists of computing global histograms of derivative responses accumulated for the whole image sequence (Chomat, Martin and Crowley, 2000; Zelnik-Manor and Irani, 2001). Although this approach is known to require segmentation when applied to complex scenes, it constitutes an interesting alternative to local methods and is potentially effective if combined with reliable segmentation. A relatively high performance of such a method for the task of action recognition has already been demonstrated above in Section 7.3.3.

When computing histograms of spatio-temporal derivatives, the derivative responses depend on the relative motion between the camera and the moving pattern.

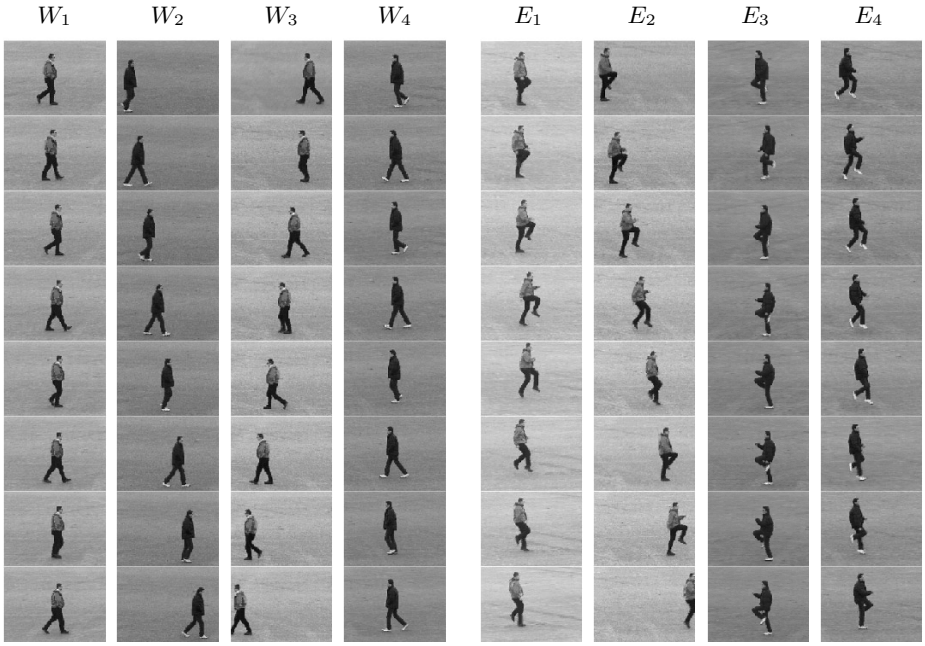


Figure 7.11: Test sequences of people walking W_1–W_4 and people performing an exercise E_1–E_4. Whereas the sequences W_1,W_4,E_1,E_3 were taken with a manually stabilized camera, the other four sequences were recorded using a stationary camera.

In this section, we investigate the effect of adapting derivatives with respect to the local motion in the pattern using the dense velocity adaptation approach presented in Section 5.3. In particular, we use velocity-adapted spatio-temporal derivative operators up to order four and collect histograms of these at different spatial and temporal scales. For simplicity, we restrict ourselves to 1-D histograms for each type of filter response. To achieve independence with respect to the direction of motion (left/right or up/down) and the sign of the spatial grey-level variations, we simplify the problem by only considering the absolute values of the filter responses. Moreover, to emphasize the parts of the histograms that correspond to stronger spatio-temporal responses, we use heuristics and weight the accumulated histograms $H(i)$ by a function $f(i) = i^2$ resulting in $h(i) = i^2 H(i)$.

7.4.1 Experimental setup

As a test problem we have chosen a data set with image sequences containing people performing actions of type *walking* $W_1 ... W_4$ and *exercise* $E_1 ... E_4$ as illustrated in Figure 7.11. As can be seen, some of the sequences were taken with a stationary camera, while the others were recorded with a manually stabilized camera. Each of

these 4 sec. long sequences were subsampled to a spatio-temporal resolution of $80 \times 60 \times 50$ pixels and convolved with a set of spatio-temporal smoothing kernels for all combinations of seven velocities $v_x = -3...3$, five spatial scales $\sigma^2 = \{2, 4, 8, 16, 32\}$ and five temporal scales $\tau^2 = \{2, 4, 8, 12, 16\}$.

For each combination of the spatial scale and the temporal scales (σ_i, σ_j), velocity adaptation was performed according to (5.14) at scale level $(\sigma_{i+1}, \sigma_{j+1})$. Since in our examples the relative camera motion was mostly horizontal, we performed maximization in (5.14) over v_x only. The result of this adaptation for the sequences W_2 and E_1 is illustrated in Figure 7.12.

To represent the patterns, we accumulated histograms of derivative responses for each combination of scales and each type of derivative. For the purpose of evaluation, separate histograms were accumulated over (i) velocity-adapted derivative responses; (ii) velocity-steered derivative responses (see Section 6.1.1) and (iii) non-adapted partial derivative responses computed at velocity $v = 0$.

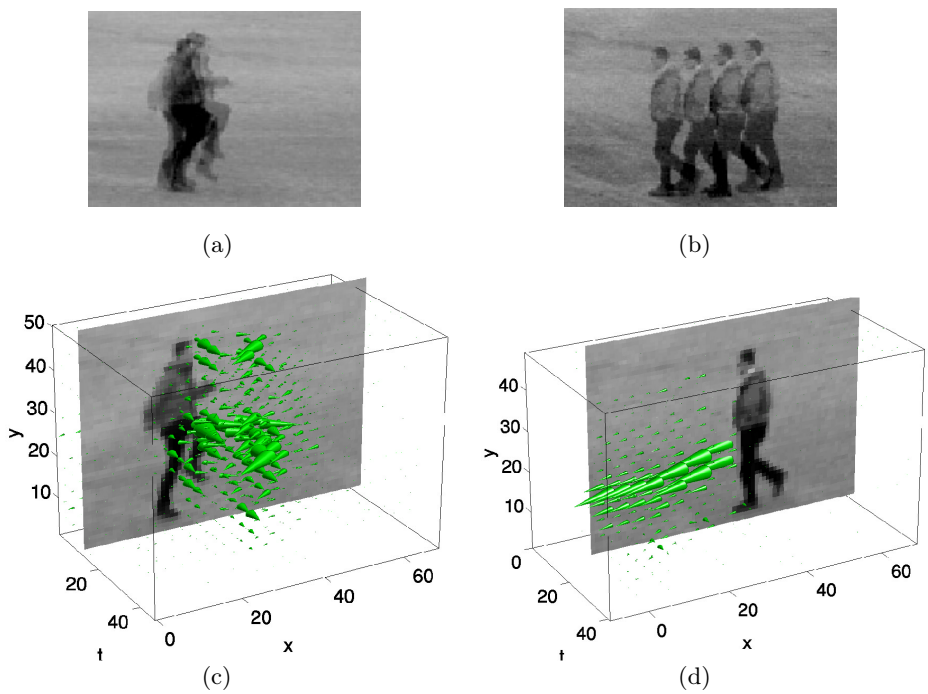


Figure 7.12: Results of local velocity adaptation for image sequences recorded with a manually stabilized camera (a), and with a stationary camera (b). Directions of cones in (c)-(d) correspond to the velocity chosen by the proposed adaptation algorithm. The size of the cones corresponds the value of the squared Laplacian $\big((\partial_{xx} + \partial_{yy})L(x, y, t; \sigma, \tau)\big)^2$ at the selected velocities.

7.4.2 Discriminability of histograms

Figure 7.13 illustrates the means and the variances of the accumulators in the histograms computed separately for both of the classes. As can be seen from Figures 7.13(a)-(c), velocity-adaptation of receptive fields results in more discriminative class histograms and lower variations between histograms computed for the same class of activities. On the contrary, the high variations in the histograms in Figures 7.13(d)-(f) and Figures 7.13(g)-(i) clearly indicate that activities are much harder to recognize when using velocity-steered or non-adapted receptive fields.

Whereas Figure 7.13 presents histograms for three types of derivatives L_{xxt}, L_{xyt} and L_{yyt} at scales $\sigma^2 = 4$, $\tau^2 = 4$ only, we have observed a similar behavior for other derivatives at most of the other scales considered.

7.4.3 Discriminability measure

To quantify these results, let us measure the dissimilarity between pairs of histograms (h_1,h_2) according to the χ^2-measure defined in (6.8). To evaluate the dissimilarity between a pair of sequences, we accumulate differences of histograms over different spatial and temporal scales as well as over different types of receptive fields according to $d(h_1, h_2) = \sum_{l,\sigma,\tau} D(h_1, h_2)$, where l denotes the type of the spatio-temporal filters, σ^2 the spatial scale and τ^2 the temporal scale.

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

Figure 7.14 shows distance ratios computed separately for different types of receptive fields. The lower values of the curve corresponding to velocity-adaptation clearly indicate the better recognition performance obtained by using velocity-adapted filters compared to velocity-steered or non-adapted filters. Computing distance ratios over all types of derivatives and scales used, results in the following distance ratios: $r_{adapt} = 0.64$ when using velocity-adapted filters, $r_{steered} = 0.81$ using velocity-steered filters, and $r_{non-adapt} = 0.92$ using non-adapted filters.

7.4.4 Dependency on scales

When analyzing discrimination performance for different types of derivatives and different scales, we have observed an interesting dependency of the distance ratio on the spatial and the temporal scales. Figures 7.15(a)-(b) show that the distance ratio has a minimum at scales $\sigma^2 = 2$, $\tau^2 = 8$ indicating that these scales give rise to the best discrimination for patterns considered here. In particular, it can be

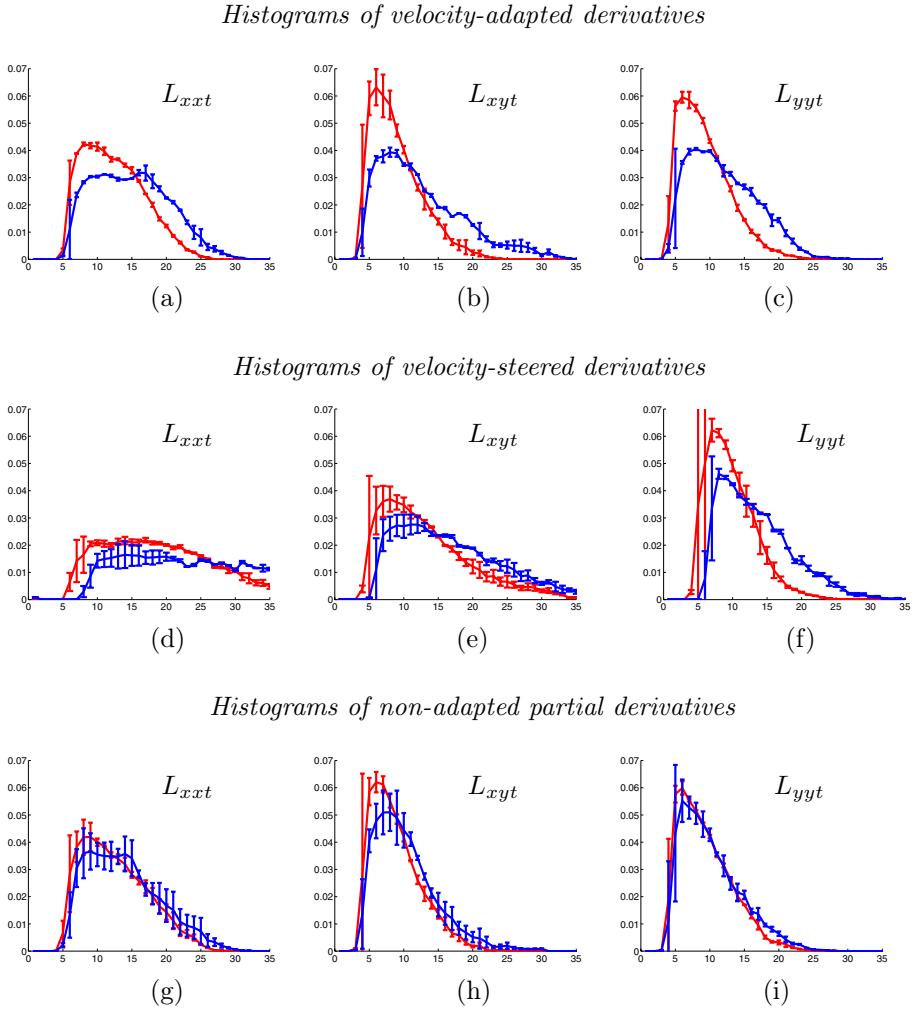


Figure 7.13: Means and variances of histograms for the activities "walking" (red) and "exercise" (blue). (a)-(c): histograms of *velocity-adapted* derivatives L_{xxt}, L_{xyt}, L_{yyt}; (d)-(f): histograms of velocity-steered derivatives L_{xxt}, L_{xyt}, L_{yyt}; (g)-(i): histograms of non-adapted partial derivatives L_{xxt}, L_{xyt}, L_{yyt}. As can be seen, the velocity-adapted filter responses give considerably better possibility to discriminate the motion patterns compared to velocity-steered or non-adapted filters.

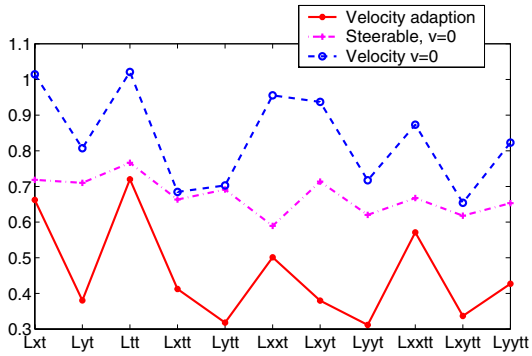


Figure 7.14: Distance ratios computed for different types of derivatives and for velocity-adapted (solid lines), velocity-steered (point-dashed lines) and non-adapted (dashed lines) filter responses. As can be seen, local velocity adaptation results in lower values of the distance ratio and therefore better recognition performance compared to steered or non-adapted filter responses.

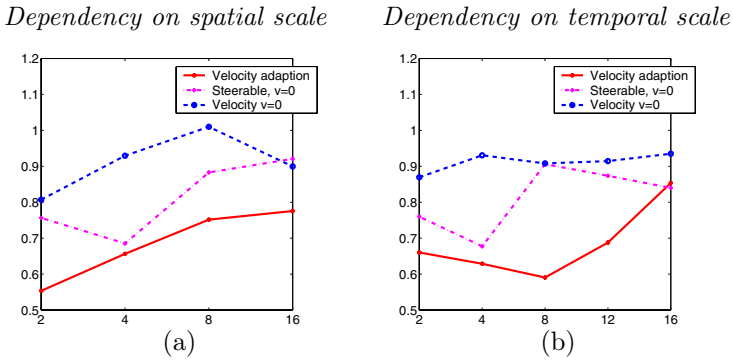


Figure 7.15: Evolution of the distance ratio r over spatial scales (a) and temporal scales (b). Minima over scales indicate scale values with the highest discrimination ability.

noted that $\tau^2 = 8$ approximately corresponds to the temporal extent of one gait cycle in our examples.

Computation of distance ratios for the selected scale values results in $r_{adapt} = 0.41$ when using velocity-adapted filters, $r_{steered} = 0.71$ using velocity-steered filters and $r_{non-adapt} = 0.79$ using non-adapted filters (see Figure 7.16). The existence of such preferred scales motivates approaches for automatic *dense* selection of both spatial (Lindeberg, 1998b) and temporal (Lindeberg, 1997b) scales.

	velocity-adapted filtering	steerable filtering	non-adapted filtering
Average over all considered scales	0.64	0.81	0.92
At (manually) selected scales $\sigma^2 = 2$, $\tau^2 = 8$	0.41	0.71	0.79

Figure 7.16: Values of distance ratios when averaged over all scales and at the manually selected scales that give best discrimination performance.

7.4.5 Summary and discussion

In this section, we have addressed the problem of dense velocity adaptation when discriminating human actions in terms of histograms of spatio-temporal derivatives. Experiments on a test problem have shown that the use of a velocity adaptation results in an improvement of the discriminability compared to using either steerable derivatives or regular partial derivatives computed from a non-adapted spatio-temporal filtering step. Whereas for the treated set of examples, recognition could also have been accomplished by using a camera stabilization approach, a major aim here has been to consider a filtering scheme that can be extended to recognition in complex scenes, where reliable camera stabilization may not be possible, i.e. scenes with complex non-static backgrounds or multiple events of interest. Full-fledged recognition in such situations, however, requires more sophisticated statistical methods for recognition complimented with the spatio-temporal segmentation.

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

Chapter 8

Towards applications

Through the coarse of previous chapters, we investigated methods for detecting, adapting and describing motion events in image sequences. One of the main motivations for this approach is to overcome the problem of spatial segmentation prior to motion recognition, which is a hard problem in complex scenes. Instead of relying on segmentation and tracking or using global measurements, we proposed to detect motion events based on the *local* image information in space-time which is independent from the *global* context of the scene. The locality of the proposed method provides potential advantages when analyzing complex scenes with multiple motions and heterogeneous, non-static backgrounds, which is a typical situation for real-life video data, such as movies and TV-programs.

Until now, however, most of the demonstrations and experiments have been performed for image sequences of simple scenes with homogeneous backgrounds. The purpose of using such simplified scenarios was to analyze basic properties of space-time features, such as the repeatability of feature detection, the stability of local descriptors and the discriminative power of the descriptors in the context of human motion recognition. Although the results of this analysis appear satisfying, the advocated power of local space-time features as being useful and robust in complex scenes has not been verified so far.

The purpose of this chapter is to show that local space-time features *can* be used for the interpretation of motion patterns in complex scenes. For the demonstration, we consider two methods for human motion analysis in outdoor street-like scenarios. In the first method presented in Section 8.1, we extend the evaluation of local descriptors in Section 7.3 and consider the task of recognizing human actions. In particular, we use the same strategy in terms of greedy matching and nearest neighbor classification and show that such a conceptually simple method, when combined with local space-time features, can be directly applied for motion interpretation in complex scenes.

As a second test problem in Section 8.2, we consider the task of detecting walking people and estimating their pose. The detection is formulated in terms of

matching pairs of image sequences, where both the model and the test sequence is represented in terms of local space-time features. Differently to the first method, we here use both local descriptors of features and positions of features in space-time to achieve spatio-temporal alignment of two sequences. The functionality of the method is demonstrated on the detection of walking people in image sequences with substantial variations in the background.

8.1 Recognition of human actions

To analyze the applicability of motion events in complex scenes, we recorded image sequences of human actions in city environments as illustrated in Figure 8.1. The type of actions was the same as in the previous evaluation in Section 7.3 except jogging. As can be seen, all the sequences (except sequences 15 and 16) contain heterogeneous background while most of the sequences also contain background motion. Moreover, about the half of all 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 variations in the spatial scale (sequences 1–3, 17, 22–27, 37), occlusions (sequences 5, 35, 13, 36, 38) and three-dimensional view variations (sequences 17, 20, 22-24).

To recognize the type of actions in this test set, we used a training set with 192 sequences of corresponding actions performed in scenes with homogeneous backgrounds and taken with a stationary camera as illustrated in Figure 7.1. In all sequences, we detected local space-time features using iterative adaptation with respect to scale and velocity according to Sections 5.1 and 5.2. For each feature, we then computed scale and velocity adapted local image descriptors according to Section 6.1. To compare a pair of image sequences, we used a greedy matching method as described in Section 6.3.1 and classified an action in each test sequence by the type of action in the most similar training sequence. 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. The recognition rate was separately computed for all (valid) combinations of twelve local descriptors and three dissimilarity measures according to Section 6.2.

The recognition rates for different types of local descriptors are presented in Figure 8.2 where for each descriptor the result is optimized over different dissimilarity measures (ED, SP, χ^2). As can be seen, the highest recognition rate is obtained for STG-PCA and STG-PD2HIST descriptors. Notably, the same type of descriptors (in the same order) gave the best performance when evaluated on action recognition in simple scenes using the same type of scale and velocity adapted features (see Figure 7.5(right)). Given the high number of all tested alternatives, the consistency of these results indicates the *stability* of all the involved steps in the current recognition scheme *independently* of the variations in the background. Furthermore, the value of the recognition rate for STG-PCA descriptor (84.3%) is close to the corresponding value in the leave-one-out experiments (82.8%) in Figure 7.5(right).

Figure 8.1: 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 and occlusions.

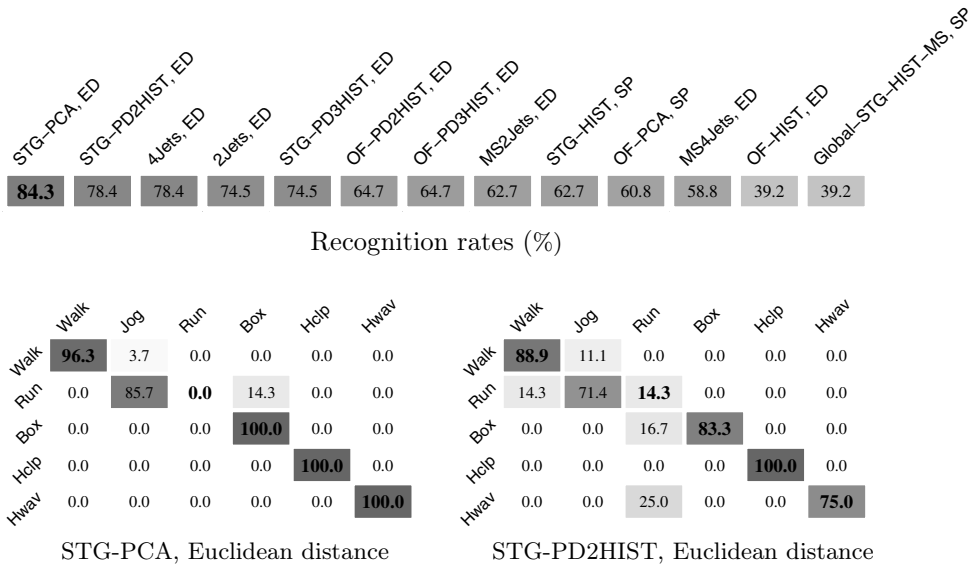


Figure 8.2: Results of recognizing human actions in complex scenes for the sequences in Figure 8.1. (Top): Recognition rates for twelve different types of local descriptors and one global descriptor maximized over different dissimilarity measures: Euclidean distance (ED), Normalized scalar product-based measure (SP) and χ^2-measure as defined in Section 6.2. (Bottom): Confusion matrices for the two best methods.

Confusion matrices for the two best descriptors are illustrated in Figure 8.2(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. Given the somewhat diffuse definition of the boundary between these two classes of actions, the recognition results can be considered as good and *stable* with respect to variations in non-static, cluttered backgrounds and other distracting factors such as scale variations, variations in the motion of the camera and occlusions. If "running" and "jogging" actions are merged into one class, the performance of STG-PCA increases to 96% while the performance of other descriptors increases with (almost) exact preservation of their relative order.

When analyzing the performance of the other descriptors, we can note that the third and the fourth best descriptors are the single-scale jets (4Jets, 2Jets) whose performance is significantly better than of the corresponding multi-scale jets (MS4Jets, MS2Jets). This result contradicts with the results of evaluation in Section 7.3.1, where multi-scale jets consistently outperformed single-scale jets. One possible explanation to this qualitative change in performance might be the larger spatio-temporal support of multi-scale descriptors which may be more influenced in complex scenes by variations in the background. We can also 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 other types of descriptors might be extremely redundant. Among the histogram-based descriptors, we can note that position-dependent histograms perform significantly better than position-independent histograms, which is consistent with the previous results in Section 7.3.2. When comparing 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 can also compare the performance of local methods to the performance of the two methods in terms of global histograms of spatio-temporal gradients as described in Section 7.3.3. From Figure 8.2, 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 global descriptors. The low performance of global histograms with the best performance for Global-STG-HIST-MS (39.2%) is not very surprising, since such descriptors depend on the motion of the camera and the motion in the background.

Given the results, we conclude that local methods significantly outperform global methods on the given problem of action recognition. Moreover, the stability of the performance for STG-PCA and STG-PD2HIST descriptors in complex scenes confirms the predicted stability of local space-time features with respect to common variations in realistic environments. We should also note that in the current recognition scheme we did not take the positions of the detected features into account. The relative positions of events in space-time are very likely to contain discriminative information for the purpose of recognition, hence, there is a straightforward direction for improvement of the current recognition scheme. Concerning other possibilities for improvement, we can note that the substitution of the nearest neighbor (NN) module in the current method with a more sophisticated classification scheme, such as Support Vector Machines (SVM), can be used and has been used (Schüldt et al., 2004) to further improve the results presented here.

Concerning limitations of the present investigation, its weakest point is probably the relatively small number of concerned types of actions (six) and the limited number of test sequences (51). Such a setting is probably not sufficient to draw general conclusions about the performance of local space-time features for the general task of representing and recognizing human actions and other classes of motion patterns in image sequences. Whereas general conclusions would require much more more experiments, the present work can be regarded as a pilot investigation with a strong theoretical and experimental support for advantages of local space-time features.

The current recognition scheme contains an implicit and limiting assumption about the single type of action per image sequence. This assumption, however, is not imposed by the use of local features nor greedy matching but by the nearest neighbor classification only. Hence, there is a straightforward extension of the current scheme to handle multiple actions within the same image sequence. Such an extension, however, would probably benefit from combining the classification and the localization of actions, since different actions usually occupy different locations

in space and possibly in time. In the next section, we investigate one possible approach for action localization, where we take both the image descriptors and the positions of local features in space-time into account.

8.2 Sequence matching

In this section, we illustrate how a sparse representation of video sequences in terms of motion events can be used for the localization of specific patterns of motion in image sequences. We consider the problem of detecting walking people and estimating their poses from side views in outdoor scenes. Such a task is complicated, since the variations in appearance of people together with the variations in the background may lead to ambiguous interpretations. Human motion has been used to resolve this ambiguity in a number of previous works. Some of the works rely on pure spatial image features while using sophisticated body models and tracking schemes to constrain the interpretation (Baumberg and Hogg, 1996; Bregler and Malik, 1998; Sidenbladh et al., 2000). Other approaches use spatio-temporal image cues such as optical flow (Black et al., 1997) or motion templates (Baumberg and Hogg, 1996). The work of Niyogi and Adelson (1994) concerns the spatio-temporal structure of the gait pattern in image sequences and is closer to ours. A related pure image-based approach for matching video sequences of the same spatio-temporal scene was presented in (Caspi and Irani, 2002).

The idea we follow here is to represent both the model and the data sequences using local space-time features and to search for a match between two sequences by matching the corresponding local features in a space-time window (see Figure 8.3). To reduce the ambiguity of matching, here, we classify the detected features into a finite vocabulary of events and assign each feature with a label c_i as described in Section 6.3.2. The match between any two image sequences is then evaluated by matching the labels and the space-time positions of corresponding features across sequences.

8.2.1 Walking model

As a representation for a gait pattern in image sequences, we construct a model of a walking person in terms of local space-time features. For this purpose we consider the upper sequence in Figure 6.8 and manually select a time interval $(t_0, t_0 + T)$ corresponding to one period T of the gait pattern. Then, given n local space-time features $f_i^m = (x_i^m, y_i^m, t_i^m, \sigma_i^m, \tau_i^m, c_i^m)$, $i = 1, ..., n$ (m stands for model) defined by the positions (x_i^m, y_i^m, t_i^m), scales (σ_i^m, τ_i^m) and classes c_i^m corresponding to the detected events in the time interval $t_i^m \in (t_0, t_0+T)$, we define the walking model by a set of periodically repeating features $M = \{f_i + (0, 0, kT, 0, 0, 0, 0) | i = 1, ..., n, k \in \mathcal{Z}\}$. Furthermore, to account for the variations of the person in the image sequences, we introduce a model state X defined by the vector $X = (x, y, \theta, s, \xi, v_x, v_y, v_s)$. The components of X describe the position of the person in the image (x, y), her size

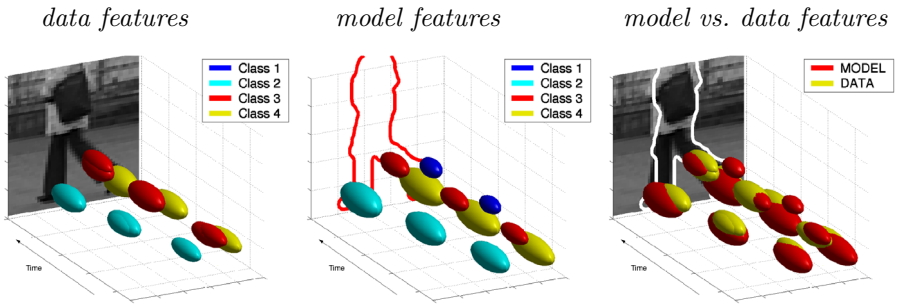


Figure 8.3: Matching of spatio-temporal data features with model features: (a) Features detected from the data sequence over a time interval corresponding to three periods of the gait cycle; (b) Model features minimizing the distance to the features in (a); (c) Model features and data features overlaid. The estimated silhouette overlaid on the current frame confirms the correctness of the method.

s, the frequency of the gait ξ, the phase of the gait cycle θ at the current time moment as well as the temporal variations (v_x, v_y, v_s) of (x, y, s). The parameters v_x and v_y describe the velocity of the person in the image, while v_s describes how fast size changes occur. Given the state X, the parameters of each model feature $f \in M$ transform according to

$$\begin{aligned} \tilde{x}^m &= x + sx^m + \xi v_x(t^m + \theta) + s\xi x^m v_s(t^m + \theta) \\ \tilde{y}^m &= y + sy^m + \xi v_y(t^m + \theta) + s\xi y^m v_s(t^m + \theta) \\ \tilde{t}^m &= \xi(t^m + \theta) \\ \tilde{\sigma}^m &= s\sigma^m + v_s s\sigma^m(t^m + \theta) \\ \tilde{\tau}^m &= \xi\tau^m \\ \tilde{c}^m &= c^m \end{aligned} \tag{8.1}$$

It follows that this type of scheme is able to handle translations and uniform rescalings in the image domain as well as uniform rescalings in the temporal domain. Hence, it allows for matching of patterns with different image velocities as well as with different frequencies over time.

To estimate the boundary of the person in test sequences, we extract silhouettes $S = \{x^s, y^s, \theta^s | \theta^s = 1, ..., T\}$ from each frame of the model sequence in Figure 6.8(top) by thresholding the intensity values. Each obtained silhouette corresponds to a discrete value of the phase parameter θ. The silhouette is used here only for visualization purpose and allows us to approximate the boundary of the person in the current frame using the model state X and a set of points $\{(x^s, y^s, \theta^s) \in S | \theta^s = \theta\}$ transformed according to $\tilde{x}^s = sx^s + x$, $\tilde{y}^s = sy^s + y$.

Figure 8.4: The result of matching a spatio-temporal walking model to sequences of outdoor scenes.

8.2.2 Model matching

Given a model state X, a current time t_0, a length of the time window t_w, and a set of data features $D = \{f^d = (x^d, y^d, t^d, \sigma^d, \tau^d, c^d) | t^d \in (t_0, t_0 - t_w)\}$ detected from the recent time window of the data sequence, the match between the model and the data is defined by a weighted sum of distances h between the model features f_i^m and the data features f_j^d

$$\mathcal{H}(\tilde{M}(X), D, t_0) = \sum_i^n h(\tilde{f}_i^m, f_j^d) e^{-(\tilde{t}_i^m - t_0)^2/\xi}. \tag{8.2}$$

Here, $\tilde{M}(X)$ denotes a set of n model features in the time window $(t_0, t_0 - t_w)$ transformed according to (8.1), i.e. $\tilde{M} = \{\tilde{f}^m | t^m \in (t_0, t_0 - t_w)\}$, $f_j^d \in D$ denotes instances of data feature that minimize the distance h with respect to model features f_i^m while ξ stands for the variance of the exponential weighting function that gives higher preference to recent events compared to other events in the past.

The distance h between two features of the same class is defined as a Euclidean distance between two points in space-time, where the spatial and the temporal dimensions are weighted with respect to a parameter ν as well as by the extents of the features in space-time

$$h^2(f^m, f^d) = (1 - \nu) \frac{(x^m - x^d)^2 + (y^m - y^d)^2}{(\sigma^m)^2} + \nu \frac{(t^m - t^d)^2}{(\tau^m)^2}. \tag{8.3}$$

Here, the distance between features of different classes is regarded as infinite. Alternatively, one could measure the feature distance by taking into account their descriptors and distances from several of the nearest cluster means.

To find the best match between the model and the data, we search for the model state $\tilde{X}$ that minimizes $\mathcal{H}$ in (8.2)

$$\tilde{X} = \text{argmin}_X \, \mathcal{H}(\tilde{M}(X), D, t_0) \tag{8.4}$$

using a standard Gauss-Newton optimization method. The result of such an optimization for a sequence with data features in Figure 8.3(a) is illustrated in Figure 8.3(b). Here, the match between the model and the data features was searched over a time window corresponding to three periods of the gait pattern or approximately 2 seconds of video. As can be seen from Figure 8.3(c), the overlaps between the model features and the data features confirm the match between the model and the data. Moreover, the model silhouette transformed according to $\tilde{X}$ matches with the contours of the person in the current frame and confirms a reasonable estimate of the model parameters.

8.2.3 Results and discussion

Figure 8.4 illustrates results of the proposed approach obtained for two outdoor sequences with walking persons. The first sequence contains variations in the projected size of a person due to camera zoom and illustrates the invariance of the method with respect variations in the spatial scale. The second sequence shows the successful detection and pose estimation of a person despite the presence of the complex non-stationary background and occlusions. Note that these results have been obtained by re-initializing model parameters before optimization at each frame. Hence, the approach appears to be stable and could be improved further by tracking the model parameters $\tilde{X}$ over time.

The presented method is an example of how the positions and the local image descriptors of space-time features could be used for representing and localizing motion patterns in image sequences. The success of this approach and the sufficiency of the involved matching method indicate the low ambiguity as well as the high stability of space-time features with respect to their positions and assigned labels. Whereas the model construction in this method was done semi-manually, it would be interesting to investigate other alternatives for fully automatic learning of motion models as for example in (Song et al., 2003).

Chapter 9

Summary and discussion

In this work, we have investigated the issue of motion representation for the purpose of recognizing motion patterns in video sequences. The initial motivation for the approach we have taken was to overcome the need of spatial segmentation, spatial recognition and temporal tracking prior to motion interpretation, since such methods might be unstable in complex scenes. Inspired by related findings in psychology, which have shown that people tend to segment the motion into action units or events (see Chapter 4), we introduced the notion of *local motion events* and proposed to use such events as primitives when representing and recognizing complex patterns of motion in video sequences. We explored local properties of events in space-time and showed how such properties can be used to overcome several problems originating from the variations of the data in real image sequences. Such variations include scale transformations, Galilean transformations, variations in lighting conditions, variations in cluttered non-static backgrounds and others. In particular, we proposed and analyzed methods for detecting, adapting and describing local motion events in image sequences.

To detect motion events, we extended a previous method of local feature detection in static images (Förstner and Gülch, 1987; Harris and Stephens, 1988) and derived an approach for detecting local space-time features in image sequences (see Chapter 4). To detect such features, we proposed to maximize a measure of local variation of the image function over space and time and to define the positions of features in space-time by the obtained maxima. Using the numerous examples throughout this thesis, we have shown that such a method detects meaningful and stable events in image sequences.

Image structures in space-time depend on scale and velocity transformations and so do the results of feature detection. To detect motion events independently of transformations in the data, we proposed an iterative method in Chapter 5 for adapting the positions and the neighborhoods of space-time features with respect to estimated values of scaling and velocity transformations. To obtain such estimates, we considered the framework of spatio-temporal scale-space and analyzed the be-

havior of local space-time image structures under changes in scale and velocity. We derived two methods for scale selection and velocity adaptation and combined them into an approach for invariant detection of spatio-temporal features. The invariance of this approach has been proved in theory for prototype image structures and has been demonstrated in practice for real image sequences in Chapter 7. Moreover, the obtained methods for scale selection and velocity adaptation are not restricted to the adaptation of local features and can be applied to other image structures in space-time. Local velocity adaptation at all points of image sequences has been applied in Section 5.1 to obtain a dense and velocity-invariant representation of motion patterns. The advantages of such a representation have been demonstrated in practice in Section 7.4.

In Chapter 6, we demonstrated how local neighborhoods of motion events can be used for describing and matching corresponding events in image sequences. For this purpose, we defined several types of local image descriptors and matched similar space-time patterns in image sequences with human actions. When formulating image descriptors, we considered either spatio-temporal derivatives or optic flow as local measurements and combined these measurements in terms of local jets, local histograms, local position-dependent histograms and principal component analysis. Moreover, using adapted neighborhoods of motion events, we obtained image descriptors that are independent of scale and velocity transformations in the data. The stability and the discriminative power of the proposed descriptors has been compared in an experimental evaluation in Chapter 7. In particular, we found that position-dependent histograms in combination with scale-adapted features performed best among the tested methods when evaluated on the problem of human action recognition within a given video database. This result is consistent with the a related finding in the spatial domain, where the SIFT descriptor (Lowe, 1999), which can be seen as a local position-dependent histogram of gradient directions, was shown to give the best performance among the other evaluated descriptors (Mikolajczyk and Schmid, 2003) for the task of matching local features in static images. A similar descriptor in terms of position-dependent histograms of wavelet coefficients has also been successfully applied to the problem of detecting cars and faces in static images by Schneiderman (2000).

Finally, in Chapter 8 we demonstrated that local motion events can be used for robust motion interpretation in complex scenes. To challenge the proposed methods, we considered image sequences with human actions recorded in city environments. By combining local motion events with a simple nearest neighbor classifier, we demonstrated that reliable recognition of human actions can be achieved in the presence of (a) motions in the background, (b) variations in spatial scale, (c) variations in camera motion, (c) three-dimensional view variations, (d) individual variations of cloth and motion of subjects and (e) occlusions. The resulting recognition performance is reasonably high and consistent with the previous results obtained for simple scenes in Chapter 7. When taking the positions of motion events into account, we also demonstrated how motion events can be used for matching pairs of image sequences and for detecting walking people in scenes with scale variations

and complex non-static backgrounds.

Summarizing the main contributions of this thesis, we have addressed the issue of motion representation and demonstrated how motion events can be used to overcome the need for spatial segmentation, spatial recognition and temporal tracking prior to motion interpretation. Moreover, we have shown that motion events provide generic primitives that can be used for expressing and recognizing different classes of motion patterns. We presented a theory and evaluated methods for local adaptation of events with respect to scale and velocity transformations in image sequences. Finally, we demonstrated that motion events in combination with the proposed local image descriptors and adaptation mechanisms provide a robust and discriminative representation for action recognition in complex scenes.

9.1 Future work

Automatic interpretation of video increases in importance due to the large amount of available video data and the ease by which new data can be acquired. Applications of automatic video interpretation can be expected to have a high impact in the areas of surveillance, robotics, human-computer interaction and content-based search and browsing of video databases. From this perspective and from the results obtained in this thesis, there is a strong motivation for further development of event-based schemes for motion interpretation.

With regard to extensions, there is a number of open issues which could be investigated to improve the current method. First of all, local motion events in this work have mostly been treated independently of each other. Whereas this approach has been taken in order to analyze the performance of pure local descriptors, there is an obvious direction for improvement by taking the dependencies of local motion events into account. For this purpose, the relative positions of events in space-time could be used to represent and recognize motion patterns in image sequences. When using relative positions of events for recognition, however, special care should be taken to preserve the invariance of such representations under global variations of motion patterns in space-time. To address this non-trivial task, one could consider a learning approach (Song et al., 2003), combine events into groups (Brown and Lowe, 2002) or use global qualitative descriptors such as order types (Carlsson, 1999).

The detection of motion events in this work has been accomplished by maximizing the variation of image values over local neighborhoods in image sequences. Whereas this approach allows for detection of well-localized and stable points in space-time, the criteria for selecting such points is basically heuristic and does not provide any guarantees that the resulting set of events is optimal for motion interpretation. Moreover, as can be seen from Figure 6.1, the detected events are often rather sparse and ignore some parts of motion patterns that intuitively seem important for the interpretation. Hence, it would be interesting to investigate alternative methods for event detection, for example, by extending other existing methods of

feature detection in the spatial domain (Kadir and Brady, 2001; Köthe, 2003) or by using Galilean invariants in space-time (Fageström, 2004; Lindeberg et al., 2004). To overcome the sparseness of event-based representations, it could also be interesting to consider representations in terms of combinations of local spatial and local spatio-temporal features and to investigate the use of such representations for integrated spatio-temporal recognition.

Concerning other applications of motion events, it could be interesting to combine events with more traditional methods for tracking in order to improve the performance of tracking at motion discontinuities. For this purpose, the occurrence of events could be predicted from the state of a model, while the events in image sequences could be used to verify the model state and to emphasize hypotheses with correct predictions.

Finally, to use motion events in real-time applications, the current schemes for event detection and adaptation will have to be modified in order to take the causality of the temporal domain into account. Whereas event detection in this work has been formulated in terms of Gaussian derivatives, a similar scheme could be implemented using recursive spatio-temporal filters as described in Section 3.3. Moreover, the iterative scheme for event adaptation could be substituted by non-iterative selection of scales and velocities in a densely sampled spatio-temporal scale-space. To make this approach computationally feasible, the scale-space representation of image sequences could be implemented in terms of image pyramids in the spatial domain (Lindeberg and Bretzner, 2003) and time-recursive scale-space representations in the temporal domain (Lindeberg and Fagerström, 1996).

Bibliography

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

Aggarwal, J. and Cai, Q. (1999). Human motion analysis: A review, *Computer Vision and Image Understanding* **73**(3): 428–440.

Allmen, M. and Dyer, C. (1990). Cyclic motion detection using spatiotemporal surfaces and curves, *Proc. International Conference on Pattern Recognition*, pp. I:365–370.

Almansa, A. and Lindeberg, T. (2000). Fingerprint enhancement by shape adaptation of scale-space operators with automatic scale-selection, *IEEE Transactions on Image Processing* **9**(12): 2027–2042.

Alvarez, L., Lions, P. and Morel, J. (1992). Image selective smoothing and edge detection by nonlinear diffusion, *SIAM J. Numer. Anal.* **29**(3): 845–866.

Alvarez, L. and Morel, J. (1994). Morphological approach to multiscale analysis, *Geometry-Driven Diffusion in Computer Vision*, Kluwer Academic Publishers, p. Chapter 8.

Amit, Y. and Geman, D. (1999). A computational model for visual selection, *Neural Computation* **11**(7): 1691–1715.

Ballester, C. and Gonzalez, M. (1998). Affine invariant texture segmentation and shape from texture by variational methods, *Journal of Mathematical Imaging and Vision* **9**(2): 141–171.

Barron, J., Fleet, D. and Beauchemin, S. (1994). Performance of optical flow techniques, *International Journal of Computer Vision* **12**(1): 43–77.

Barth, E., Stuk, I. and Mota, C. (2002). Analysis of motion and curvature in image sequences, *Southwest Symposium on Image Analysis and Interpretation*, pp. 206–210.

Bascle, B. and Blake, A. (1998). Separability of pose and expression in facial tracing and animation, *Proc. International Conference on Computer Vision*, pp. 323–328.

Baumberg, A. (2000). Reliable feature matching across widely separated views, *Proc. Computer Vision and Pattern Recognition*, pp. I: 774–781.

Baumberg, A. M. and Hogg, D. (1996). Generating spatiotemporal models from examples, *Image and Vision Computing* **14**(8): 525–532.

Beardsley, P., Zisserman, A. and Murray, D. (1997). Sequential updating of projective and affine structure from motion, *International Journal of Computer Vision* **23**(3): 235–259.

Bigün, J. and Granlund, G. H. (1987). Optimal orientation detection of linear symmetry, *Proc. International Conference on Computer Vision*, London, Great Britain, pp. 433–438.

Black, M. and Jepson, A. (1998a). Eigentracking: Robust matching and tracking of articulated objects using view-based representation, *International Journal of Computer Vision* **26**(1): 63–84.

Black, M. and Jepson, A. (1998b). A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions, *Proc. Fifth European Conference on Computer Vision*, Freiburg, Germany, pp. 909–924.

Black, M., Sapiro, G., Marimont, D. and Heeger, D. (1998). Robust anisotropic diffusion, *IEEE Transactions on Image Processing* **7**(3): 421–432.

Black, M., Yacoob, Y., Jepson, A. and Fleet, D. (1997). Learning parameterized models of image motion, *Proc. Computer Vision and Pattern Recognition*, pp. 561–567.

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

Bobick, A. and Davis, J. (2001). The recognition of human movement using temporal templates, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(3): 257–267.

Bobick, A. and Wilson, A. (1995). Using configuration states for the representation and recognition of gesture, *Proc. International Conference on Computer Vision*, pp. 382–388.

Brand, M. (1997). The "inverse hollywood problem": From video to scripts and storyboards via causal analysis, *National Conference on Artificial Intelligence (AAAI)*, pp. 132–137.

Bregler, C. (1997). Learning and recognizing human dynamics in video sequences, *Proc. Computer Vision and Pattern Recognition*, pp. 568–574.

Bregler, C. and Malik, J. (1998). Tracking people with twists and exponential maps, *Proc. Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. 8–15.

Bretzner, L. and Lindeberg, T. (1998). Feature tracking with automatic selection of spatial scales, *Computer Vision and Image Understanding* **71**(3): 385–392.

Brown, M. and Lowe, D. (2002). Invariant features from interest point groups, *British Machine Vision Conference*, pp. 253–262.

Carlsson, S. (1999). Order structure, correspondence, and shape based categories, *in* D. A. Forsyth, J. L. Mundy, V. D. Gesù and R. Cipolla (eds), *Shape, Contour and Grouping in Computer Vision*, Vol. 1681 of *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg, pp. 58–71.

Caspi, Y. and Irani, M. (2002). Spatio-temporal alignment of sequences, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(11): 1409–1424.

Cedras, C. and Shah, M. (1995). Motion-based recognition: A survey, *Image and Vision Computing* **13**(2): 129–155.

Chomat, O. and Crowley, J. (1999). Probabilistic recognition of activity using local appearance, *Proc. Computer Vision and Pattern Recognition*, pp. II:104–109.

Chomat, O., de Verdiere, V., Hall, D. and Crowley, J. (2000). Local scale selection for Gaussian based description techniques, *Proc. Sixth European Conference on Computer Vision*, Vol. 1842 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. I:117–133.

Chomat, O., Martin, J. and Crowley, J. (2000). A probabilistic sensor for the perception and recognition of activities, *Proc. Sixth European Conference on Computer Vision*, Vol. 1842 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. I:487–503.

Cutting, J., Proffitt, D. and Kozlowski, L. (1978). A biomechanical invariant for gait perception, *Journal of Experimental Psychology: Human Perception and Performance* **4**: 357–372.

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

Deutscher, J., Blake, A. and Reid, I. (2000). Articulated body motion capture by annealed particle filtering, *Proc. Computer Vision and Pattern Recognition*, Hilton Head, SC, pp. II:126–133.

Doretto, G., Chiuso, A., Wu, Y. and Soatto, S. (2003). Dynamic textures, *International Journal of Computer Vision* **51**(2): 91–109.

Doretto, G., Cremers, D., Favaro, P. and Soatto, S. (2003). Dynamic texture segmentation, *Proc. Ninth International Conference on Computer Vision*, Nice, France, pp. 1236–1242.

Duda, R., Hart, P. and Stork, D. (2001). *Pattern Classification*, Wiley.

Efros, A., Berg, A., Mori, G. and Malik, J. (2003). Recognizing action at a distance, *Proc. Ninth International Conference on Computer Vision*, Nice, France, pp. 726–733.

Elder, J. and Zucker, S. (1998). Local scale control for edge detection and blur estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(7): 699–716.

Elgammal, A., Shet, V., Yacoob, Y. and Davis, L. (2003). Learning dynamics for exemplar-based gesture recognition, *Proc. Computer Vision and Pattern Recognition*, pp. I:571–578.

Engel, S. and Rubin, J. (1986). Detecting visual motion boundaries, *Workshop on Motion: Representation and Analysis*, Charleston, S.C., pp. 107–111.

Fablet, R., Bouthemy, P. and Perez, P. (2002). Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval, *IEEE Transactions on Image Processing* **11**(4): 393–407.

Fageström, D. (2004). Galilean differential geometry of moving images, *Proc. of European Conference on Computer Vision, to appear.*

Faugeras, O. (1993). *Three-Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press.

Faugeras, O., Luong, Q. and Papadopoulo, T. (2001). *The Geometry of Multiple Images*, MIT Press.

Fei-Fei, L., Fergus, R. and Perona, P. (2003). A bayesian approach to unsupervised one-shot learning of object categories, *Proc. Ninth International Conference on Computer Vision*, Nice, France, pp. 1134–1141.

Fergus, R., Perona, P. and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning, *Proc. Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. II:264–271.

Fleet, D., Black, M., Yacoob, Y. and Jepson, A. (2000). Design and use of linear models for image motion analysis, *International Journal of Computer Vision* **36**(3): 171–193.

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

Florack, L., Salden, A., ter Haar Romeny, B., Koenderink, J. and Viergever, M. (1995). Nonlinear scale-space, *Image and Vision Computing* **13**(4): 279–294.

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

Gavrila, D. (1999). The visual analysis of human movement: A survey, *Computer Vision and Image Understanding* **73**(1): 82–98.

Gavrila, D. and Davis, L. (1995). Towards 3-D model-based tracking and recognition of human movement, *in* M. Bichsel (ed.), *Int. Workshop on Face and Gesture Recognition*, IEEE Computer Society, pp. 272–277.

Gavrila, D. and Davis, L. (1996). 3-D model-based tracking of humans in action: a multi view approach, *Proc. Computer Vision and Pattern Recognition*, pp. 73–80.

Gibson, J. (1950). The perception of the visual world, *Houghton Mifflin*.

Gould, K. and Shah, M. (1989). The trajectory primal sketch: A multi-scale scheme for representing motion characteristics, *Proc. Computer Vision and Pattern Recognition*, pp. 79–85.

Granlund, G. and Knutsson, H. (1995). *Signal Processing for Computer Vision*, Kluwer Academic Publishers.

Griffin, L. (1996). Critical point events in affine scale space, *in* J. Sporring, M. Nielsen, L. Florack and P. Johansen (eds), *Gaussian Scale-Space Theory*, Kluwer Academic Publishers, Copenhagen, Denmark, pp. 165–180.

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

Hadjidemetriou, E., Grossberg, M. and Nayar, S. (2002). Resolution selection using generalized entropies of multiresolution histograms, *Proc. Seventh European Conference on Computer Vision*, Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, pp. I:220–235.

Hanna, K. (1991). Direct multi-resolution estimation of ego-motion and structure from motion, *Proc. Workshop Visulal Motion*, pp. 156–162.

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

Hartley, R. (1994). Projective reconstruction and invariants from multiple images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**(10): 1036–1041.

Hartley, R. and Zisserman, A. (2000). Multiple view geometry in computer vision, *Cambridge*.

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

Heeger, D. and Pentland, A. (1986). Seeing structure through chaos, *Workshop on Motion: Representation and Analysis*, pp. 131–136.

Hoey, J. and Little, J. (2000). Representation and recognition of complex human motion, *Proc. Computer Vision and Pattern Recognition*, Hilton Head, SC, pp. I:752–759.

Hoey, J. and Little, J. (2003). Bayesian clustering of optical flow fields, *Proc. Ninth International Conference on Computer Vision*, Nice, France, pp. 1086–1093.

Horn, B. (1987). Motion fields are hardly ever ambiguous, *International Journal of Computer Vision* **1**(3): 239–258.

Horn, B. and Schunck, B. (1981). Determining optical flow, *Artificial Intelligence* **17**(1–3): 185–203.

Irani, M., Anandan, P. and Cohen, M. (2002). Direct recovery of planar-parallax from multiple frames, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(11): 1528–1534.

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

Isard, M. and Blake, A. (1998). A mixed-state condensation tracker with automatic model-switching, *Proc. Sixth International Conference on Computer Vision*, Bombay, India, pp. 107–112.

Jägersand, M. (1995). Saliency maps and attention selection in scale and spatial coordinates: An information theoretic approach, *Proc. Fifth International Conference on Computer Vision*, Cambridge, MA, pp. 195–202.

Jähne, B., Haußecker, H. and Geißler, P. (1999). 2. signal processing and pattern recognition, *Handbook of Computer Vision and Applications*, Academic Press, p. Chapter 13.

Johansson, G. (1976). Visual motion perception, *Scientific America* **232**: 75–88.

Kadir, T. and Brady, M. (2001). Saliency, scale and image description, *International Journal of Computer Vision* **45**(2): 83–105.

Knutsson, H. (1989). Representing local structure using tensors, *The 6th Scandinavian Conference on Image Analysis*, Oulu, Finland, pp. 244–251.

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

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

Koenderink, J. J. and van Doorn, A. J. (1992). Generic neighborhood operators, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**(6): 597–605.

Koenderink, J. and van Doorn, A. (1987). Representation of local geometry in the visual system, *Biological Cybernetics* **55**: 367–375.

Koenderink, J. and van Doorn, A. (1999). The structure of locally orderless images, *International Journal of Computer Vision* **31**(2/3): 159–168.

Köthe, U. (2003). Integrated edge and junction detection with the boundary tensor, *Proc. Ninth International Conference on Computer Vision*, Nice, France, pp. 424–431.

Kumar, R., Anandan, P. and Hanna, K. (1994). Direct recovery of shape from multiple views: A parallax based approach, *Proc. International Conference on Pattern Recognition*, pp. 685–688.

Kuniyoshi, Y. and Inoue, H. (1993). Qualitative recognition of ongoing human action sequences, *International Joint Conference on Artificial Intelligence*, pp. 1600–1609.

Laptev, I. and Lindeberg, T. (2003a). A distance measure and a feature likelihood map concept for scale-invariant model matching, *International Journal of Computer Vision* **52**(2/3): 97–120.

Laptev, I. and Lindeberg, T. (2003b). Interest point detection and scale selection in space-time, *in* L. Griffin and M. Lillholm (eds), *Scale-Space'03*, Vol. 2695 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, pp. 372–387.

Laptev, I. and Lindeberg, T. (2003c). Space-time interest points, *Proc. Ninth International Conference on Computer Vision*, Nice, France, pp. 432–439.

Laptev, I. and Lindeberg, T. (2004a). Local descriptors for spatio-temporal recognition, *Proc. of ECCV Workshop on Spatial Coherence for Visual Motion Analysis, to appear.*

Laptev, I. and Lindeberg, T. (2004b). Velocity adaptation of space-time interest points, *Proc. of ICPR, to appear.*

Laptev, I. and Lindeberg, T. (2004c). Velocity-adaptation of spatio-temporal receptive fields for direct recognition of activities: An experimental study, *Image and Vision Computing* **22**(2): 105–116.

Lassiter, G., Geers, A., Apple, K. and Beers, M. (2000). Observational goals and behavior unitization: A reexamination, *Journal of Experimental Social Psychology* **36**: 649–659.

Leibe, B. and Schiele, B. (2003). Interleaved object categorization and segmentation, *British Machine Vision Conference*, Norwich, GB.

Linde, O. and Lindeberg, T. (2004). Object recognition using composed receptive field histograms of higher dimensionality, *Proc. of ICPR, to appear*.

Lindeberg, T. (1993). Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention, *International Journal of Computer Vision* **11**(3): 283–318.

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

Lindeberg, T. (1997a). Linear spatio-temporal scale-space, *in* B. M. ter Haar Romeny, L. M. J. Florack, J. J. Koenderink and M. A. Viergever (eds), *Scale-Space Theory in Computer Vision: Proc. First Int. Conf. Scale-Space'97*, Vol. 1252 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Utrecht, The Netherlands, pp. 113–127.

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

Lindeberg, T. (1998a). Edge detection and ridge detection with automatic scale selection, *International Journal of Computer Vision* **30**(2): 117–154.

Lindeberg, T. (1998b). Feature detection with automatic scale selection, *International Journal of Computer Vision* **30**(2): 77–116.

Lindeberg, T. (2002). Time-recursive velocity-adapted spatio-temporal scale-space filters, *Proc. Seventh European Conference on Computer Vision*, Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, pp. I:52–67.

Lindeberg, T., Akbarzadeh, A. and Laptev, I. (2004). Galilean-corrected spatio-temporal interest operators, *Proc. of ICPR, to appear*.

Lindeberg, T. and Bretzner, L. (2003). Real-time scale selection in hybrid multi-scale representations, *in* L. Griffin and M. Lillholm (eds), *Scale-Space'03*, Vol. 2695 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, pp. 148–163.

Lindeberg, T. and Fagerström, D. (1996). Scale-space with causal time direction, *Proc. Fourth European Conference on Computer Vision*, Vol. 1064 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Cambridge, UK, pp. I:229–240.

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

Lindeberg, T. and Gårding, J. (1997). Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure, *Image and Vision Computing* **15**(6): 415–434.

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

Lowe, D. (2001). Local feature view clustering for 3d object recognition, *Proc. Computer Vision and Pattern Recognition*, Kauai Marriott, Hawaii, pp. I:682–688.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* p. to appear.

Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision, *DARPA Image Understanding Workshop*, pp. 121–130.

Malik, J., Belongie, S., Shi, J. and Leung, T. (1999). Textons, contours and regions: Cue integration in image segmentation, *Proc. Seventh International Conference on Computer Vision*, Corfu, Greece, pp. 918–925.

Mann, R., Jepson, A. and Siskind, J. (1997). The computational perception of scene dynamics, *Computer Vision and Image Understanding* **65**(2): 113–128.

Maturana, H., Lettwin, J., McCulloch, W. and Pitts, W. (1960). Anatomy and physiology of vision in the frog (rana pipins), *Journal of General Physiology* **43**(2): 129–171.

Mel, B. and Fiser, J. (2000). Minimizing binding errors using learned conjunctive features, *Neural Computation* **12**(4): 731–762.

Mikolajczyk, K. and Schmid, C. (2001). Indexing based on scale invariant interest points, *Proc. Eighth International Conference on Computer Vision*, Vancouver, Canada, pp. I:525–531.

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

Mikolajczyk, K. and Schmid, C. (2003). A performance evaluation of local descriptors, *Proc. Computer Vision and Pattern Recognition*, pp. II: 257–263.

Moeslund, T. and Granum, E. (2001). A survey of computer vision-based human motion capture, *Computer Vision and Image Understanding* **81**(3): 231–268.

Nagel, H. and Gehrke, A. (1998). Spatiotemporal adaptive filtering for estimation and segmentation of optical flow fields, *in* H. Burkhardt and B. Neumann (eds), *Proc. Fifth European Conference on Computer Vision*, Vol. 1407 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Freiburg, Germany, pp. II:86–102.

Nan, L., Dettmer, S. and Shah, M. (1997). Visually recognizing speech using eigen sequences, *in* M. Shah and R. Jain (eds), *Motion-Based Recognition*, Kluwer Academic Publishers, Dordrecht Boston, pp. 345–371.

Newtson, D., Engquist, G. and Bois, J. (1977). The objective basis of behavior units, *Journal of Personality and Social Psychology* **35**(12): 847–862.

Nielsen, M. and Lillholm, M. (2001). What do features tell about images?, *in* M. Kerckhove (ed.), *Scale-Space'01*, Vol. 2106 of *LNCS*, Springer, pp. 39–50.

Nistér, D. (2001). *Automatic Dense Reconstruction from Uncalibrated Video Sequences*, Phd thesis, Department of Numerical Analysis and Computer Science (NADA), KTH, KTH, S-100 44 Stockholm, Sweden.

Niyogi, S. A. (1995). Detecting kinetic occlusion, *Proc. Fifth International Conference on Computer Vision*, Cambridge, MA, pp. 1044–1049.

Niyogi, S. and Adelson, H. (1994). Analyzing and recognizing walking figures in XYT, *Proc. Computer Vision and Pattern Recognition*, pp. 469–474.

Perona, P. and Malik, J. (1990). Scale space and edge detection using anisotropic diffusion, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**(7): 629–639.

Piater, J. (2001). *Visual Feature Learning*, Phd thesis, Department of Computer Science, University of Massachusetts, Amherst, MA 01003-9264.

Polana, R. and Nelson, R. (1992). Recognition of motion from temporal texture, *Proc. Computer Vision and Pattern Recognition*, pp. 129–134.

Polana, R. and Nelson, R. (1997). Temporal texture and activity recognition, *in* M. Shah and R. Jain (eds), *Motion-Based Recognition*, Kluwer Academic Publishers, Dordrecht, Boston, London, pp. 87–124.

Rangarajan, K., Allen, W. and Shah, M. (1992). Recognition using motion and shape, *Proc. International Conference on Pattern Recognition*, pp. I:255–258.

Rao, C., Yilmaz, A. and Shah, M. (2002). View-invariant representation and recognition of actions, *International Journal of Computer Vision* **50**(2): 203–226.

Reingold, E. and Tarjan, R. (1981). On a greedy heuristic for complete matching, *SIAM J. Computing* **10**(4): 676–681.

Rohr, K. (1997). Human movement analysis based on explicit motion models, *in* M. Shah and R. Jain (eds), *Motion-Based Recognition*, Kluwer Academic Publishers, Dordrecht Boston, pp. 171–198.

Rothganger, F., Lazebnik, S., Schmid, C. and Ponce, J. (2003). 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints, *Proc. Computer Vision and Pattern Recognition*, pp. II: 272–277.

Rubin, J. and Richards, W. (1985). Boundaries of visual motion, *MIT AI Memo*.

Rui, Y. and Anandan, P. (2000). Segmenting visual actions based on spatio-temporal motion patterns, *Proc. Computer Vision and Pattern Recognition*, Vol. I, Hilton Head, SC, pp. 111–118.

Runeson, S. (1974). Constant velocity – not perceived as such, *Psychological Research* **37**: 3–23.

Schiele, B. and Crowley, J. (1996). Object recognition using multidimensional receptive field histograms, *Proc. Fourth European Conference on Computer Vision*, Vol. I, Cambridge, UK, pp. 610–619.

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

Schmid, C. and Mohr, R. (1997). Local grayvalue invariants for image retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(5): 530–535.

Schmid, C., Mohr, R. and Bauckhage, C. (2000). Evaluation of interest point detectors, *International Journal of Computer Vision* **37**(2): 151–172.

Schneiderman, H. (2000). *A Statistical Approach to 3D Object Detection Applied to Faces and Cars*, Phd thesis, Robotics Institute, Carnegie Mellon University, Pittsburg, PA 15213.

Schüldt, C., Laptev, I. and Caputo, B. (2004). Recognizing human actions: a local SVM approach, *Proc. of ICPR, to appear*.

Shah, M. and Jain, R. (eds) (1997). *Motion-Based Recognition*, Kluwer Academic Publishers, Dordrecht, Boston, London.

Sidenbladh, H. and Black, M. (2001). Learning image statistics for bayesian tracking, *Proc. Eighth International Conference on Computer Vision*, Vancouver, Canada, pp. II:709–716.

Sidenbladh, H., Black, M. and Fleet, D. (2000). Stochastic tracking of 3D human figures using 2D image motion, *Proc. Sixth European Conference on Computer Vision*, Vol. 1843 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. II:702–718.

Siskind, J. and Morris, Q. (1996). A maximum-likelihood approach to visual event classification, *Proc. Fourth European Conference on Computer Vision*, Cambridge, UK, pp. II:347–360.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos, *Proc. Ninth International Conference on Computer Vision*, Nice, France, pp. 1470–1477.

Sminchisescu, C. and Triggs, B. (2003). Kinematic jump processes for monocular 3d human tracking, *Proc. Computer Vision and Pattern Recognition*, pp. I:69–76.

Smith, S. and Brady, J. (1995). ASSET-2: Real-time motion segmentation and shape tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**(8): 814–820.

Soatto, S. and Perona, P. (1998). Reducing structure-from-motion: A general framework for dynamic vision part 1: Modeling, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(9): 933–942.

Song, Y., Goncalves, L. and Perona, P. (2003). Unsupervised learning of human motion, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(7): 814–827.

Sullivan, J. and Carlsson, S. (2002). Recognizing and tracking human action, *Proc. Seventh European Conference on Computer Vision*, Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, p. I:629 ff.

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

Tell, D. and Carlsson, S. (2002). Combining topology and appearance for wide baseline matching, *Proc. Seventh European Conference on Computer Vision*, Vol. 2350 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Copenhagen, Denmark, pp. I:68–83.

ter Haar Romeny, B. (2003). *Front-End Vision and Multi-Scale Image Analysis: Multi-scale Computer Vision Theory and Applications, Written in Mathematica*, Kluwer Academic Publishers.

ter Haar Romeny, B. (ed.) (1994). *Geometry-Driven Diffusion in Computer Vision*, Kluwer Academic Publishers.

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

Tuytelaars, T. and Van Gool, L. (2000). Wide baseline stereo matching based on local, affinely invariant regions, *British Machine Vision Conference*, pp. 412–425.

Tversky, B., Morrison, J. and Zacks, J. (2002). On bodies and events, *in* A. Meltzoff and W. Prinz (eds), *The Imitative Mind*, Cambridge University Press, Cambridge.

Viola, P., Jones, M. and Snow, D. (2003). Detecting pedestrians using patterns of motion and appearance, *Proc. Ninth International Conference on Computer Vision*, Nice, France, pp. 734–741.

Wallach, H. and O'Connell, D. (1953). The kinetic depth effect, *Journal of Experimental Psychology* **45**: 205–217.

Weber, M., Welling, M. and Perona, P. (2000). Unsupervised learning of models for visual object class recognition, *Proc. Sixth European Conference on Computer Vision*, Vol. 1842 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, Dublin, Ireland, pp. I:18–32.

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

Wertheimer, M. (1958). Principles of perceptual organisation, *Readings in Perception* pp. 115–135.

Witkin, A. P. (1983). Scale-space filtering, *Proc. 8th Int. Joint Conf. Art. Intell.*, Karlsruhe, Germany, pp. 1019–1022.

Yacoob, Y. and Black, M. (1998). Parameterized modeling and recognition of activities, *Proc. International Conference on Computer Vision*.

Yacoob, Y. and Black, M. (1999). Parameterized modeling and recognition of activities, *Computer Vision and Image Understanding* **73**(2): 232–247.

Young, R. A. (1987). The Gaussian derivative model for spatial vision: I. Retinal mechanisms, *Springer-Verlag* **2**: 273–293.

Yu, C. and Ballard, D. (2002). Learning to recognize human action sequences, *International Conference on Development and Learning*, Cambridge, Massachusetts, pp. 28–33.

Yuille, A. L. and Poggio, T. A. (1986). Scaling theorems for zero-crossings, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**: 15–25.

Zacks, J., Tversky, B. and Iyer, G. (2000). Perceiving, remembering, and communicating structure in events, *Journal of Experimental Psychology: General* **130**(1): 29–58.

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

Zetzsche, C. and Barth, E. (1991). Direct detection of flow discontinuities by 3d curvature operators, *Pattern Recognition Letters* **12**: 771–779.