



## Characterizing Image Motion

Konstantinos G. Derpanis

Technical Report CS-2006-06

June 12, 2006

Department of Computer Science

4700 Keele Street North York, Ontario M3J 1P3 Canada

# Characterizing Image Motion

Konstantinos G. Derpanis

Supervisors: Richard P. Wildes and John K. Tsotsos

Committee Members: Minas Spetsakis and Hugh R. Wilson

Qualifying Examination Report

Department of Computer Science and Engineering

York University, Toronto

## **Abstract**

This paper contains a review of the characterization and recovery of motion from image sequences. The main dichotomy of considered approaches is in terms of the extent of the region of description, namely, local (infinitesimal) versus regional. The local consideration concentrates on the recovery of optical flow, while the regional considerations provide richer, potentially semantically meaningful quantitative or qualitative descriptions of motion. The paper concludes with an outline of open problems.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	In the beginning ...	1
1.2	Areas of motion analysis	2
1.3	From local to regional descriptors	4
1.4	Outline of paper	5
<b>2</b>	<b>Fundamental principles of motion analysis</b>	<b>7</b>
2.1	Spatiotemporal image representation	7
2.1.1	Image sequence representation: The plenoptic function	7
2.1.2	Image sequence representation: Derived measurements	11
2.2	Relating image brightness and motion	12
2.2.1	The visual motion field	12
2.2.2	Motion analysis: The brightness constancy assumption	15
	Local motion	15
	Motion as orientation in spatiotemporal images	18
	Motion in the frequency domain	19
2.2.3	Motion field versus optical flow	22
2.2.4	Revisiting the brightness constancy assumption	23
2.2.5	Multiple motion representations	31
<b>3</b>	<b>Recovering optical flow</b>	<b>37</b>
3.1	Matching methods	38
3.2	Differential-based estimation approaches	42
3.2.1	Local estimation methods: Least-squares and variants	42
3.2.2	Confidence measures	46
3.2.3	Global estimation methods	46
3.2.4	Robust methods	50
3.2.5	Probabilistic methods	55
3.3	Frequency-based approaches	57
3.4	Coarse-to-fine processing	60
3.5	Equivalences	62

3.6	Discussion . . . . .	67
<b>4</b>	<b>Regional Descriptors of Motion</b>	<b>73</b>
4.1	Parametric motion models . . . . .	73
4.2	Layered motion model representations . . . . .	77
4.3	Temporal textures . . . . .	79
4.4	Optical flow-based reasoning . . . . .	82
4.5	Spatiotemporal structure-based reasoning . . . . .	88
4.5.1	Structure tensor methods . . . . .	88
4.5.2	Spectral analysis-based approaches . . . . .	90
<b>5</b>	<b>Open problems</b>	<b>93</b>
5.1	Modeling theory . . . . .	93
5.2	Estimation . . . . .	94
5.3	Experimental evaluation . . . . .	95
	<b>References</b>	<b>95</b>

# List of Figures

2.1	Plenoptic function . . . . .	8
2.2	Summary of spatiotemporal representations . . . . .	10
2.3	Camera coordinate system . . . . .	13
2.4	One-dimensional optical flow constraint . . . . .	16
2.5	Optical flow constraint equation in velocity space . . . . .	18
2.6	Aperture problem . . . . .	19
2.7	An example image sequence is depicted in the spatial, spatiotemporal and frequency domains . . . . .	20
2.8	Motion as orientation . . . . .	21
2.9	Motion in the Fourier domain . . . . .	22
2.10	Example of the optical flow field not always equal to the motion field	23
2.11	Optical snow in frequency domain . . . . .	34
2.12	Layer decomposition of motion . . . . .	35
3.1	Region-based matching . . . . .	39
3.2	Least-squares versus total least-squares . . . . .	46
3.3	Minimization along countour . . . . .	48
3.4	Robust estimator example . . . . .	50
3.5	Constraint line clustering . . . . .	52
3.6	Common error norms and their influence functions . . . . .	54
3.7	Motion energy . . . . .	59
3.8	Gaussian and Laplacian pyramid examples . . . . .	61
3.9	Hierarchical motion estimation . . . . .	62
3.10	Equivalence between correlation and energy model . . . . .	66
3.11	<i>Yosemite</i> sequence . . . . .	70
3.12	<i>Yosemite</i> sequence performance . . . . .	71
4.1	Kinematic motion . . . . .	75
4.2	Affine motion expressed as a linear combination of basis flows . . . . .	76
4.3	Layered model decomposition . . . . .	78
4.4	Examples of temporal textures . . . . .	80

4.5	Phase portrait classification . . . . .	84
4.6	Principle axes classification of structure tensor . . . . .	90
4.7	Primitive spatiotemporal patterns . . . . .	91
5.1	Otte and Nagel's <i>Marbled-Block</i> sequence . . . . .	96



# List of Tables

2.1	Summary of common image sequence representations for motion analysis	11
2.2	Summary of conservation assumptions and their corresponding differential “optical flow” constraints . . . . .	30
2.3	Summary of spatiotemporal and spectral definitions of motion . . . .	36
3.1	Common region-based match measures . . . . .	41
3.2	Summary of <i>Yosemite</i> image velocity results . . . . .	69
4.1	Eigenvalue-based classification of first-order planar phase portraits . .	86
4.2	Kinematic-based classification of first-order planar phase portraits . .	87
4.3	Regional classification by eigenvalue analysis. . . . .	88

# Chapter 1

## Introduction

THIS paper contains a review of the analysis of the apparent motion from image sequences. The apparent motion is assumed to be the result of the relative motion between objects in the world and a camera. The relative motion may be due to camera motion, the motion of objects in the world or both. Broadly speaking, the area of motion analysis in computer vision encompasses methods that distill information of the scene from the apparent motion of brightness patterns in the image sequence or methods that remain purely in the image domain, such as image mosaicing.

### 1.1 In the beginning ...

The majority of the early work in motion analysis in the context of machine vision was application driven; Nagel ([Nagel, 1978](#); [Nagel, 1981](#)) provides an extensive catalogue of these application-oriented investigations. Here a mere sampling of early (pre-1980) application-oriented manifestations of motion analysis is presented.

The origins of the analysis of biological movements via image sequences can be traced as far back as the 18th century to the work of Muybridge ([Muybridge, 1887](#)). Video compression is an area where motion estimation received/receives significant attention (e.g., ([Mounts, 1969](#); [Haskell, 1974](#); [Limb & Murphy, 1975a](#); [Limb & Murphy, 1975b](#); [Netravali & Robbins, 1979](#))). Closely following the advent of television, researchers (e.g., ([Kretzmer, 1952](#))) discovered that there was a significant amount of temporal redundancy between television picture elements. By computing the motion of elements (regions) significant bandwidth reductions were realized by transmitting only those elements in subsequent frames that were insufficiently characterized by the motion of elements in previous frames. Satellite and aerial video was used to measure the growth of forests ([Goldberg & Kourtz, 1977](#)) and analyze meteorological processes such as cloud movement patterns ([Fujita, 1969](#)). Applications of motion analysis also appeared in traffic monitoring systems ([Onoe et al., 1973](#)). In the area of

medicine, Tsotsos ([Tsotsos et al., 1979](#)) proposed a system for describing normal and abnormal dynamics of the human left ventricle. The system was based on mapping trajectories of markers implanted on the heart to natural-language concepts defined by cardiologists.

By 1986, application-independent contributions addressing general issues of motion analysis had started to dominate the literature, indicating the maturity of the field of motion analysis as a discipline of study concerned with the fundamental problems of analysis and interpretation of image sequences shared by many of the aforementioned applications ([Nagel, 1986](#)). However, it should be pointed out that important application independent analyses had already appeared by this time (e.g., ([Horn & Schunck, 1981](#))). Evidence of this maturity can be seen in the emergence of workshops (e.g., Workshop on Motion first held in 1979 ([Aggarwal & Badler, 1979](#))) and conference sessions dedicated to motion analysis such as the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) and the IEEE International Conference on Computer Vision (ICCV).

## 1.2 Areas of motion analysis

For the purpose of organizing the following discussion, here a broad categorization of motion analysis is provided, as follows: optical flow (local) analysis, regional motion analysis, structure from motion, tracking, motion detection and segmentation and motion understanding. This organization is not meant to imply that each of these areas are independent of each other, in fact many of these areas are interdependent. For example, optical flow estimation represents an important component of structure from motion. Optical flow and regional motion analysis are the main focus of this paper and are not considered further in this section.

*Visual tracking* is concerned with localizing a particular image region from frame-to-frame. Generally, tracking approaches can be categorized as either non-predictive or predictive. Non-predictive trackers make decisions based on motion information extracted from the current frame. Whereas, predictive trackers process the motion history of the tracked region for the purpose of predicting its future position. The two types of predictive-tracking approaches prominent in the literature are Kalman and particle filter trackers. The *Kalman filter* ([Kalman, 1960](#)) is a general framework adapted from the *optimal estimation theory* literature. A Kalman filter in the context of visual tracking is a recursive algorithm that estimates the predicted position and uncertainty of a moving target in the next frame. The predicted position coupled with the uncertainty defines the search region in the next frame. In brief, the Kalman filter is based on two equations, the *state transition model* and the *observation model*. The state model encapsulates the relationship between position and the hidden state parameters of the model (e.g., velocity). While the observation model

relates the system's state to a set of image measurements. Key assumptions of these two equations are that they are linear and corrupted by additive Gaussian noise. These assumptions allow for a close-form solution. Unfortunately, for many realistic models the linear and Gaussian assumptions are too restrictive. The *extended Kalman filter* (EKF) and *iterated extended Kalman filter* (Bar-Shalom & Li, 1993) relax the linearity assumption by linearizing all non-linear models. A drawback of the (I)EKF is that the linearization may lead to filter instability if the models are not sufficiently linear within the time-step interval of consideration (Julier et al., 1995). To address these limitations Julier and Uhlmann (Julier et al., 1995) developed the *unscented Kalman filter* (UKF). The UKF replaces the linearization steps of the (I)EKF with a deterministic sampling approach to capture the mean and covariance estimates with a minimal set of sampling points. *Particle filtering* methods (also known as the CONDENSATION algorithm, survival of the fittest, bootstrap filter, sequential Monte Carlo, etc.) are concerned with relaxing the Kalman filtering assumptions, namely, admitting non-linear models and non-Gaussian noise while foregoing closed-form solutions (e.g., (Isard & Blake, 1998)). Particle filtering solutions rely on stochastic simulations to estimate the position and potentially multi-modal uncertainty.

*Structure from motion* is concerned with the recovery of three-dimensional structure and velocity or displacement from relationships between these variables and image motion. Structure from motion represents one of the oldest problems considered in computer vision (e.g., (Ullman, 1979)). This is due in part to the fact that many early researchers held the view that the ultimate goal of computer vision was the recovery of three-dimensional surface shape, the so-called reconstructionist view. Much of structure from motion research can be traced back even further to the field of photogrammetry which is concerned with measuring and processing lengths and angles in photographs for mapping purposes (McCurdy, 1944). There are three distinct methodologies for the computation of structure from motion: *finite displacement*, *infinitesimal* and *direct* methods. The *finite displacement* approach recovers the three-dimensional structure and displacement (in terms of rotation and translation) based on the displacement between consecutive images of a sparse set of highly discriminative image features, such as corners and lines. The *infinitesimal* approach recovers the three-dimensional structure and velocity (in terms of rotation and translation) based on optical flow estimates. *Direct* methods recover unknown parameters directly from image quantities (e.g., time varying brightness) at each pixel, thus foregoing flow estimates altogether. No matter which methodology is followed the estimates have proven to be notoriously sensitive to noise (e.g., (Daniilidis & Spetsakis, 1997)). Current research is concerned with both modeling the noise process and developing robust algorithms (e.g., (Oliensis, 2002)). For a more detailed overview of the subject of structure from motion the reader is referred to the following review papers (Huang & Tsai, 1981; Aggarwal & Nandhakumar, 1988; Huang & Netravali, 1994).

*Motion detection and segmentation* (also known as change detection, foreground detection and background detection) is concerned with identifying temporal changes in the image. The temporal changes in an image are assumed to be the result of the motion of objects in the scene. The computationally simplest of these algorithms computes the difference between the current image and a previous image and labels those pixels whose difference exceeds a noise threshold (e.g., (Mounts, 1969)). A more sophisticated approach, made possible by the significant increases in computing power and storage capacity, adaptively model the intensity at each pixel as probability distributions and identify motion at points where the observed intensity deviates significantly from the underlying “learned” model (Stauffer & Grimson, 1999; Elgammal et al., 2000). Another class of approaches consider “nulling out” the global scene motion, followed by detecting the independently moving objects (e.g., (Burt et al., 1989)).

*Motion understanding* consists of describing the motion of object(s) over extended observations in terms of human recognizable concepts. These approaches connect spatiotemporal paths of tracked feature points in an image sequence with concepts, such as natural language (motion-related nouns and verbs). The problem of motion understanding can be broken down into the following subproblems (Tsotsos et al., 1980): the computer vision component that extracts features from the image, the representation and organization of the knowledge base, and the recognition system. Examples of motion understanding systems include: interpreting the dynamics of the human left ventricle (Tsotsos et al., 1980) and traffic monitoring (Marburger & Neumann, 1981; Koller et al., 1991; Kollnig et al., 1994).

### 1.3 From local to regional descriptors

In this review paper the focus is on *local* and *regional* descriptions of image motion.

Local descriptions considered in Chapter 3, describe motion within an infinitesimal neighbourhood of space-time, where descriptions are in terms of optical flow. This should not be confused with the fact that in computing a **unique** local description consideration of a region about a point is required. Historically, significant focus has been placed on optical flow analysis as exemplified by the multitude of review papers (Huang & Tsai, 1981; Aggarwal, 1986; Nagel, 1986; Hildreth & Koch, 1987; Aggarwal & Nandhakumar, 1988; Vega-Riveros & Jabbour, 1989; Barron et al., 1994a; Beauchemin & Barron, 1995; DuFaux & Moscheni, 1995; Mitiche & Bouthemy, 1996; Haußecker & Spies, 1999; Stiller & Konrad, 1999; Fleet & Weiss, 2005) and comparative evaluation papers (Burt et al., 1982; Little & Verri, 1989; Willick & Yang, 1991; Barron et al., 1994b; Otte & Nagel, 1994; Liu et al., 1996; Bainbridge-Smith & Lane, 1997; Galvin et al., 1998; McCane et al., 2001) dedicated to the topic.

Regional methods, considered in Chapter 4, relax the infinitesimal consideration

and attempt to describe motion or the image structure within a (non-infinitesimal) region of space-time. Both quantitative and qualitative approaches have appeared in the literature for the purpose of describing the motion and image structure within a region.

## 1.4 Outline of paper

The remainder of this paper is organized into three main chapters. Chapter 2 reviews the fundamental principles underlying the methods of motion analysis considered in the subsequent chapters. Chapter 3 reviews approaches for recovering optical flow (i.e., infinitesimal region of analysis). Chapter 4 reviews literature that relaxes the infinitesimal spatiotemporal consideration with the goals of recovering richer, potentially semantically meaningful descriptions of the motion or spatiotemporal image structure based on quantitative or qualitative means. Finally, chapter 5 concludes this paper with an outline of open problems.



# Chapter 2

## Fundamental principles of motion analysis

IN this chapter we consider the fundamental principles underlying the analysis of motion within local and regional extents of space-time, that will be discussed in Chapters 3 and 4, respectively. The fundamental principles considered are the spatiotemporal representation of an image sequence (Section 2.1) and the relationship between the image brightness and motion (Section 2.2).

### 2.1 Spatiotemporal image representation

#### 2.1.1 Image sequence representation: The plenoptic function

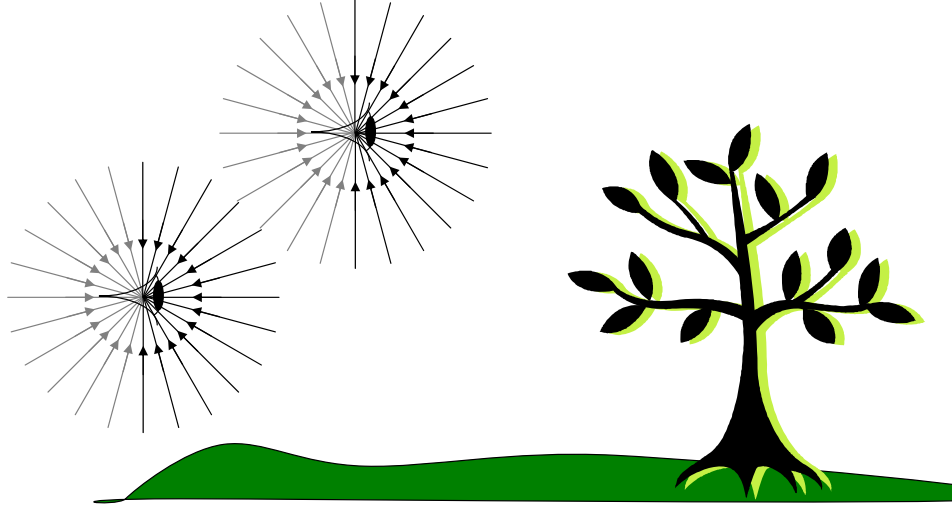
In this section the representations of image sequences considered in this paper are outlined. To bring the various representations into a common framework of understanding, the representations are presented as instances of the *plenoptic function*.

Adelson and Bergen ([Adelson & Bergen, 1991](#)) introduced the formalization of the plenoptic function  $P$  (from plenus, complete or full, and optic) to describe the potential information available to an observer at any point in space and time. This information is captured by the pencil of light rays passing through the imaging device. In its most general form, the plenoptic function is a seven-dimensional function parameterized by the position of the imaging sensor in three-dimensional space ( $V_x, V_y, V_z$ ), the wavelength of light  $\lambda$ ,  $(\theta, \phi)$  representing the azimuth and elevation angles that index the viewable rays and  $t$  representing time (see Fig. 2.1 for an illustrative example of the plenoptic function):

$$P = P(\theta, \phi, t, \lambda, V_x, V_y, V_z). \quad (2.1)$$

The spherical parameterization makes it explicit that the light impinges a given point





**Figure 2.1.** Plenoptic function. Two viewpoint samples are depicted represented by the eyes observing pencils of light rays. The grey lines represent rays coming from behind the optical sensor. Adapted from (Adelson & Bergen, 1991).

from all directions.

The plenoptic function is an idealized model, containing all potential information available. In practice, machine vision implementations consider subsets and samples of the dimensions. The representations for analysis considered in this paper and presented below can be seen as instances of the plenoptic function distinguished by the considered dimensions, the extent of the region of analysis (or in discrete terms the number of samples taken) within each available dimension and the resolution within the available dimensions.

Two specializations common among all the image representations considered in this paper are as follows: a pinhole camera model (Horn, 1986) is assumed throughout the paper, thus the imaging rays passing through the imaging plane (from the front) are available at any given time instant, and a single sample of the viewpoint is available at any given time. In this light, it is more natural to adopt the standard Cartesian parameterization of the rays  $(x, y)$ , where  $x$  and  $y$  represent the spatial coordinates of the points in the image plane:

$$P = P(x, y, t, \lambda). \quad (2.2)$$

Note although not considered here, motion analysis approaches exist that consider simultaneously rays over a hemi-spherical field of view through the use of omnidirectional devices (e.g., (Daniilidis et al., 2002)); for a recent review article on these class of sensors see (Yagi, 1999).

In terms of the wavelength, historically analysis has been predominately limited

to a single sample in the form of the *image brightness* (formally, the *image irradiance* assuming a calibrated system (Horn, 1986)), where the image brightness is recovered by way of averaging over the wavelengths of the visible spectrum. This was due in part to the unavailability, until fairly recently, of inexpensive colour cameras, and computational speed and storage capacity to process more than one wavelength sample (e.g., colour). Various recent attempts have considered multispectral samples in the form of colour and infrared images (Ohta, 1989; Markandey & Flinchbaugh, 1990; Golland & Bruckstein, 1997; Irani & Anandan, 1998; Barron & Klette, 2002; Andrews & Lovell, 2003; van de Weijer & Gevers, 2004). Woodham (Woodham, 1990) considered images of scenes observed under several illumination sources. It is important to note that the motion is implicitly assumed to be consistent across the considered spectrum. In the sequel, the discussion will be limited to image brightness, however, each of the approaches may be generalized to multispectral samples.

The main distinctions among approaches for motion analysis considered in this paper are as follows:

- the set of spatiotemporal dimensions considered
- whether an image sequence is treated as a sequence of snapshots or as temporally continuous.

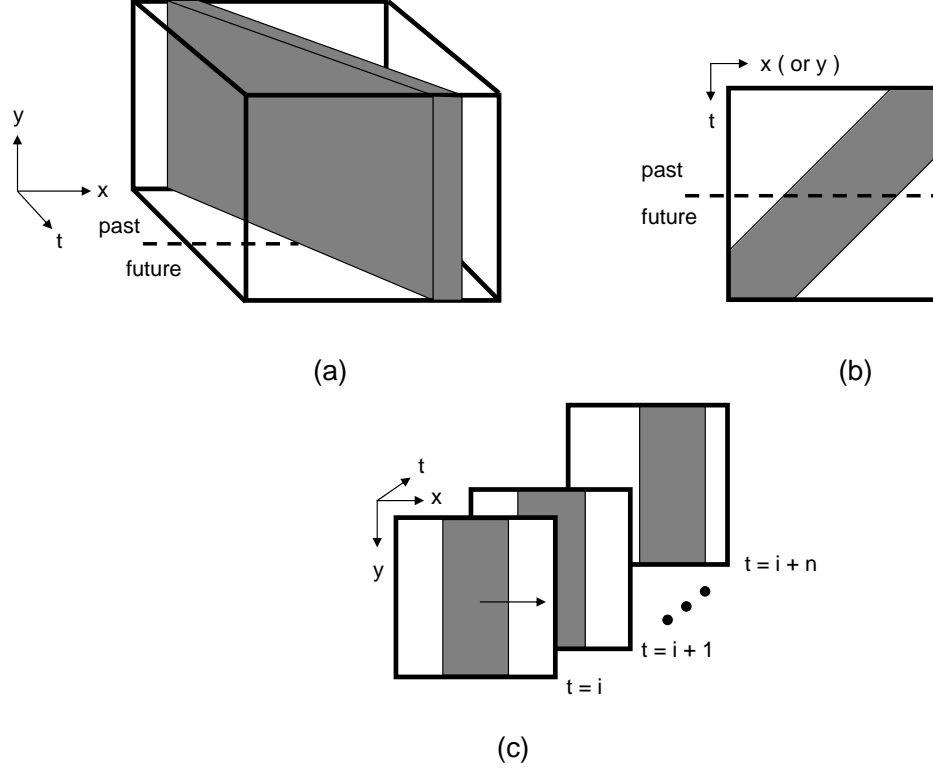
Here it is important to point out that although in implementation we are dealing with discrete image representations, continuity can be assumed in the case where the spatiotemporal dimensions have been sampled rapidly enough, as defined by the *Nyquist frequency* (Oppenheim et al., 1997), such that the representation provides a close approximation of the underlying continuous distribution.

The parameterization of the plenoptic function representing the consideration of both spatial  $(x, y)$  and temporal  $t$  dimension, where each dimension is assumed continuous, is given as follows,

$$P = P(x, y, t). \quad (2.3)$$

One can think of this representation as a spatiotemporal volume formed by “stacking” temporally consecutive images (Fahle & Poggio, 1981; Watson & Ahumada, 1985; Adelson & Bergen, 1985; Jähne, 1990) (see Fig. 2.2 (a)). Although beyond the scope of this paper, the spatiotemporal volume has also appeared in the context of motion detection (e.g., (Liou & Jain, 1989)) and the recovery of scene structure (Bolles & Baker, 1985; Bolles et al., 1987; Faugeras, 1990).

Rather than consideration of both spatial dimensions, motion analysis approaches have been proposed (e.g., (Fennema & Thompson, 1979; Adelson & Bergen, 1985; Adelson & Bergen, 1986; Wildes & Bergen, 2000)) that limit analysis to a single spatial dimension,  $x$  or  $y$ , and time,  $t$  (Fig. 2.2 (b)). Although, only a single spatial dimension is considered at a time, the results of this two-dimensional (space-time)



**Figure 2.2.** Summary of spatiotemporal representations: (a) volumetric representation of a moving gray bar, (b) spatiotemporal slice of (a), and (c) a sequence of snapshots of (a).

analysis are intended to be integrated in subsequent stages to cover both spatial dimensions plus time. The plenoptic function of these two-dimensional spatiotemporal representations are denoted as follows,

$$P = P(x, t) \quad \text{and} \quad P = P(y, t). \quad (2.4)$$

The representations above model space-time as continuous or equivalently in the discrete domain as non-aliased. Alternatively, space-time has been represented as a coarsely time sampled set of images or “snapshots”, where significant temporal aliasing precludes the recovery of measurements from the underlying continuous signal over time; the spatial domain is assumed to be sampled rapidly enough such that strong aliasing effects are not present. The plenoptic function is denoted as follows (see Fig. 2.2 (c) for an illustrative depiction),

$$P = P(x, y, \{t = i, i + \Delta t, \dots, i + n\Delta t\}) \quad (2.5)$$

where the temporal samples may not necessarily adhere to the Nyquist sampling rate (see Fig. 2.2 (c)).

Representation	Plenoptic function parameterization
<i>Spatiotemporal volume</i>	$P(x, y, t)$
<i>Spatiotemporal frequency volume</i>	$\hat{P}(\omega_x, \omega_y, \omega_t)$
<i>Spatiotemporal slice</i>	$P(x, t)$ $P(y, t)$
<i>Spatiotemporal frequency slice</i>	$\hat{P}(\omega_x, \omega_t)$ $\hat{P}(\omega_y, \omega_t)$
<i>Temporal snapshot</i>	$P(x, y, \{t = i, i + \Delta t, \dots, i + n\Delta t\})$
<i>Temporal snapshot spatial frequency</i>	$\hat{P}(\omega_x, \omega_y, \{t = i, i + \Delta t, \dots, i + n\Delta t\})$

**Table 2.1.** Summary of common image sequence representations for motion analysis and their respective plenoptic function parameterizations. Note that  $\hat{P}$  denotes the plenoptic function represented in the frequency domain.

In addition, each of these representations can be considered within the frequency domain by replacing each of their respective spatiotemporal dimensions with their frequency counterpart,  $\hat{P}(\omega_x, \omega_y, \omega_t)$ , or in the case of the “snapshot” representation considering each of the frames separately in the spatial frequency domain,  $\hat{P}(\omega_x, \omega_y, \{t = i, i + 1, \dots, i + n\})$ .

Table 2.1 summarizes the various spatiotemporal image representations considered in this section.

### 2.1.2 Image sequence representation: Derived measurements

Rather than direct consideration of image brightness for motion analysis, local functions of image brightness have been considered.

Wohn et al. (Wohn et al., 1983) considered images preprocessed by the following local functions: the spatial gradient magnitude, curvature and moments. Several authors (e.g., (Buxton & Buxton, 1984; Hildreth, 1984; Waxman et al., 1988; Gong, 1989; Duncan & Chou, 1992)) have analyzed motion in images preprocessed with a spatial edge detector operator. Mitiche et al. (Mitiche et al., 1987) considered images preprocessed by the local functions: average, contrast, entropy, median, power content and variance.

Several authors (e.g., (Haralick & Lee, 1983; Tretiak & Pastor, 1984; Giroi et al., 1989; Verri & Poggio, 1989; Uras et al., 1989; Simoncelli, 1993b; Tistarelli, 1994; Tistarelli, 1995; del Bimbo et al., 1996)) analyzed images prefiltered by spatiotemporal

first derivatives. Simoncelli points out that derivative prefilters may be extended to higher orders (Simoncelli, 1993b). Furthermore, the derivatives may be computed over a number of scales and orientations in space or space-time, such as in (Weber & Malik, 1995). More generally, the derivative filters are instances of linear oriented, bandpass filters. The Gabor filter (Gabor, 1946; Daugman, 1980) represents another instance of a linear oriented, bandpass prefilter that has been used by several authors (e.g., (Adelson & Bergen, 1986; Heeger, 1987; Heeger, 1988; Fleet & Jepson, 1989; Fleet & Jepson, 1993; Spetsakis, 1994)). The output of linear prefilters provides local information about the amplitude, phase and frequency content (Adelson & Bergen, 1986). Fleet and Jepson (Fleet & Jepson, 1989; Fleet & Jepson, 1993), consider the local phase while omitting the amplitude information from further consideration.

Another form of preprocessing are multiresolution representations that describe the input image at multiple spatial scales of resolution. Two prominent examples of multiresolution representations is the Gaussian pyramid (Burt, 1981) and the Laplacian pyramid (Burt & Adelson, 1983). We return to the topic of multiresolution representations in Chapter 3.4.

## 2.2 Relating image brightness and motion

### 2.2.1 The visual motion field

The *visual motion field* (or motion field for short) is defined as the perspective projection on the image plane of the three-dimensional instantaneous scene velocity.

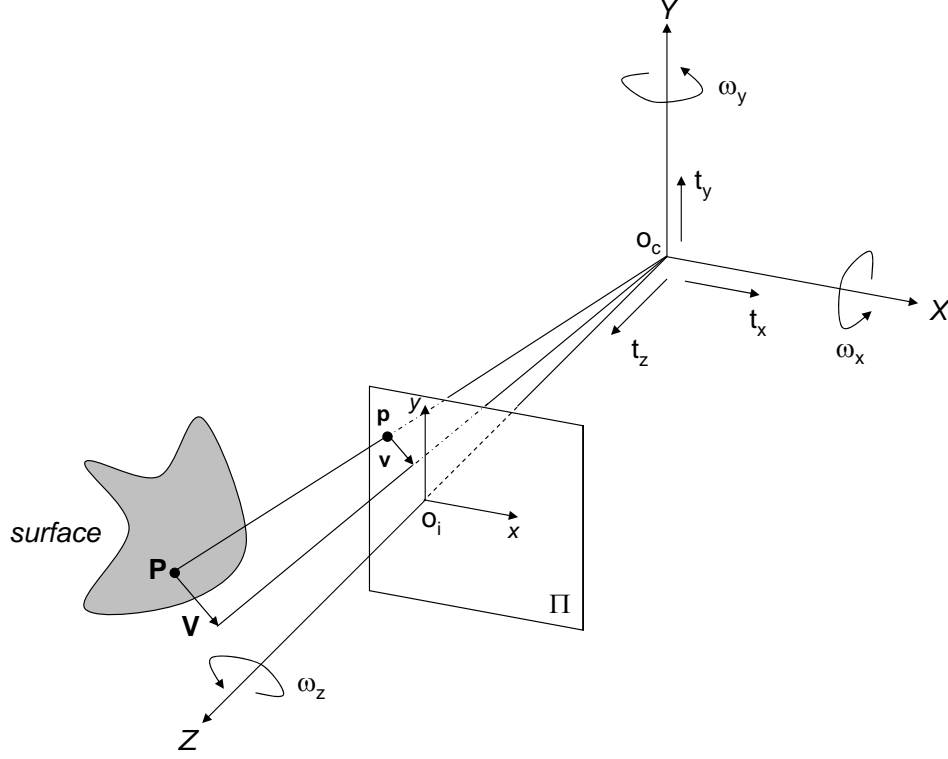
The visual motion field  $(u, v)$  was first formulated by Koenderink and van Doorn (Koenderink & van Doorn, 1975) in spherical coordinates and later by Longuet-Higgins and Prazdny (Longuet-Higgins & Prazdny, 1980) in Cartesian coordinates. Less formal use was known earlier (e.g., (Gibson, 1950)). In this section a brief review of the development of the motion field equation in Cartesian coordinates is presented.

Given a point  $\mathbf{P} = (X, Y, Z)^\top$  in the three-dimensional world space, its corresponding perspective projection onto the image plane with focal length  $f$  (without loss of generality the focal length  $f = 1$ ) is denoted in homogeneous coordinates by  $\mathbf{p} = (x, y, 1)^\top$  and related as follows (see Fig. 2.3),

$$\mathbf{p} = \frac{\mathbf{P}}{Z}. \quad (2.6)$$

The motion field is obtained by differentiating (2.6) with respect to time,

$$\begin{aligned} \dot{x} &= u = \frac{\dot{X}}{Z} - \frac{X\dot{Z}}{Z^2} \\ \dot{y} &= v = \frac{\dot{Y}}{Z} - \frac{Y\dot{Z}}{Z^2}, \end{aligned} \quad (2.7)$$



**Figure 2.3.** Camera coordinate system. The camera coordinate system is depicted with origin  $O_c$ . The image plane denoted by  $\Pi$  with origin  $o_i$  is located at  $Z = f$  where  $f$  denotes the focal length. Perspective projection maps a point  $\mathbf{P} = (X, Y, Z)$  to  $\mathbf{p} = (x, y)$ . The parameters  $t_x$ ,  $t_y$  and  $t_z$  represent the translational velocities in the  $X$ ,  $Y$  and  $Z$  directions respectively,  $\omega_x$ ,  $\omega_y$  and  $\omega_z$  represent the infinitesimal angle of rotation about  $X$ ,  $Y$  and  $Z$  conducted about the camera origin.  $\mathbf{V}$  and  $\mathbf{v}$  denote the world and image velocity of the point  $\mathbf{P}$ , respectively.

where  $(u, v)$  represents the projected image velocity of points in the scene and  $\dot{X}$ ,  $\dot{Y}$  and  $\dot{Z}$  represent the temporal derivatives of the respective spatial parameter.

The three-dimensional instantaneous velocity of the scene point  $\mathbf{P}$  can be modeled by the combination of translation  $\mathbf{T} = (t_x, t_y, t_z)$  and rotational velocity  $\mathbf{\Omega} = (\omega_x, \omega_y, \omega_z)$ , as follows,

$$\dot{\mathbf{P}} = -(\mathbf{\Omega} \times \mathbf{P} + \mathbf{T}). \quad (2.8)$$

Substituting (2.6) and (2.8) into (2.7), yields,

$$\mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{Z}(xt_z - t_x) + xy\omega_x - (x^2 + 1)\omega_y + y\omega_z \\ \frac{1}{Z}(yt_z - t_y) + (y^2 + 1)\omega_x - xy\omega_y - x\omega_z \end{pmatrix} \quad (2.9)$$

where  $(x, y)$  represents the image coordinates and  $Z$  the depth coordinate of the three-dimensional point. Eq. (2.9) is termed the *visual motion field equation*. An

important implication of the motion field equation is that the structure component  $Z$  is non-linearly coupled only with the translational components, thus the velocity field is invariant to the scene structure in the case where no translation is present.

So far the structural component  $Z(X, Y)$  in the motion field equation (2.9) has been left unspecified. Let us next consider the specialization of the structure of the surface as a plane. The importance of considering planar structures lies in the fact that much of the structure that we encounter in the world can be approximated locally as smooth and by extension, locally approximated as a plane,

$$\alpha X + \beta Y + \gamma Z = 1 \quad (2.10)$$

parameterized by  $\alpha, \beta$  and  $\gamma$ . Substituting the equation of the plane (2.10) into the motion field equation (2.9) and some algebraic manipulation, yields (Longuet-Higgins & Prazdny, 1980),

$$\mathbf{u}(x, y, t) = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a_0 + a_1x + a_2y + a_6x^2 + a_7xy \\ a_3 + a_4x + a_5y + a_6xy + a_7y^2 \end{pmatrix} \quad (2.11)$$

where,

$$\begin{aligned} a_0 &= -t_x\gamma - \omega_y \\ a_1 &= -t_x\alpha + t_z\gamma \\ a_2 &= -t_x\beta + \omega_z \\ a_3 &= -t_y\gamma + \omega_x \\ a_4 &= -t_y\alpha - \omega_z \\ a_5 &= -t_y\beta + t_z\gamma \\ a_6 &= -\omega_y + t_z\alpha \\ a_7 &= t_z\beta + \omega_x. \end{aligned} \quad (2.12)$$

Notice that the planar function (2.10) may be considered as a local first-order Taylor series expansion of a surface, similar representations have been derived using various higher-order Taylor series expansions of a surface (Longuet-Higgins & Prazdny, 1980; Waxman & Ullman, 1985; Negahdaripour, 1992). Also, due to the polynomial nature of the motion field equation, the parameters of the flow model (2.11) directly represent the structure of the local flow field:

$$\mathbf{u}(x, y, t) = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} + \mathbf{J}(x, y)^\top + \frac{1}{2}(x, y)\mathbf{H}(x, y)^\top \quad (2.13)$$

where  $(u_0, v_0)^\top$  denotes the instantaneous velocity at  $(x_0, y_0)^\top$ ,  $\mathbf{J}$  the Jacobian matrix of the motion field,

$$\mathbf{J} = \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} \quad (2.14)$$

and  $\mathbf{H}$  the Hessian matrix,

$$\mathbf{H} = \begin{bmatrix} u_{xx} & u_{xy} \\ v_{yx} & v_{yy} \end{bmatrix}. \quad (2.15)$$

When the distance between the surface(s) and the camera is large compared to the distance variation within the surface, it is usually possible to approximate the polynomial flow by a linear flow field by ignoring the second-order terms of (2.11) (Negahdaripour & Lee, 1991; Bergen et al., 1992; Campani & Verri, 1992), leading to the following affine model,

$$\mathbf{u}(x, y, t) = \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a_0 + a_1x + a_2y \\ a_3 + a_4x + a_5y \end{pmatrix} \quad (2.16)$$

$$= \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} + \mathbf{J}(x, y)^\top. \quad (2.17)$$

### 2.2.2 Motion analysis: The brightness constancy assumption

Optical flow, first coined by Gibson (Gibson, 1950), is a two-dimensional vector field representing the apparent velocities of the brightness patterns in an image (Horn, 1986). Optical flow is defined by the assumptions that are imposed on both the spatiotemporal image and image velocity field structures. Thus, there is not one unique optical flow, rather, there potentially are many. The most common assumption imposed on the spatiotemporal image structure, is that of brightness constancy of points as they move from frame-to-frame.

The following discussion outlines the manifestation of motion within the various image sequence representations discussed in Section 2.1.1. Throughout the following discussion the brightness constancy is assumed to hold.

#### Local motion

Let  $I(x, y, t)$  be the volumetric (i.e., continuous) representation of the image sequence brightness, where  $(x, y)$  represents the spatial  $x$  and  $y$  dimensions,  $t$  denotes time, and  $u(x, y, t)$  and  $v(x, y, t)$  are the horizontal and vertical components of the optical flow, respectively. Assume that the intensity of a point remains the same as it moves by  $\delta x, \delta y$  for a time interval  $\delta t$ , given formally as,

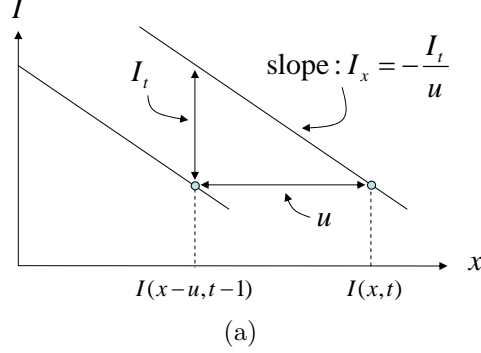
$$I(x - \delta x, y - \delta y, t - \delta t) = I(x, y, t). \quad (2.18)$$

This assumption is commonly termed the *brightness constancy assumption*. Assuming that the brightness varies smoothly (i.e., is differentiable) with  $x, y$  and  $t$ , we can take the Taylor series expansion of the left-hand side of (2.18),

$$I(x, y, t) - I_x\delta x - I_y\delta y - I_t\delta t + \text{h.o.t.} = I(x, y, t) \quad (2.19)$$

$$I_x\delta x + I_y\delta y + I_t\delta t + \text{h.o.t.} = 0 \quad (2.20)$$





**Figure 2.4.** The one-dimensional (i.e., motion restricted to a single dimension) geometric interpretation of the optical flow constraint is depicted, where  $I(x-u, t-1) = I(x, t)$  and the spatiotemporal structure is assumed linear. Adapted from (Fennema & Thompson, 1979).

where h.o.t. represents the higher order terms (i.e., second- and higher-order) in  $\delta x, \delta y$  and  $\delta t$ . From (2.20) there are two standard ways to proceed. The first approach consists of dividing through by  $\delta t$  and taking the limit as  $\delta t \rightarrow 0$ , while, the second approach consists of fixing  $\delta t = 1$  and neglecting h.o.t. by assuming that the spatiotemporal structure around the point is locally linear, both approaches yield,

$$I_x u + I_y v + I_t = 0. \quad (2.21)$$

Interpreting  $u, v$  as velocity without  $\delta t \rightarrow 0$  differs in that strictly speaking  $u, v$  corresponds to displacement that coincides with velocity in the case where the assumption of local spatiotemporal linearity holds (see Fig. 2.4 for a one-dimensional geometric interpretation of this case), while in the case of allowing  $\delta t \rightarrow 0$  no assumption on the series expansion of the spatiotemporal structure is made. Eq. (2.21) relates the image velocity to the spatiotemporal derivatives of the image at a particular location and is commonly referred to as the *optical flow constraint equation* (Horn, 1986), *brightness constancy constraint equation* (Jähne, 2005), *gradient constraint equation* (Fleet & Weiss, 2005) and *motion constraint equation* (Haußecker & Spies, 1999). One can interpret the optical flow constraint (2.21) as algebraically expressing that the (non-normalized) directional derivative in the direction  $(u, v, 1)^\top$  of the spatiotemporal image structure vanishes.

An alternative development of the optical flow constraint assumes that each point in a given image moves along a path in which the intensity is conserved (i.e., brightness constancy), formally,

$$I(x(t), y(t), t) = c. \quad (2.22)$$

Taking the total derivative of both sides with respect to time, yields,

$$\frac{d}{dt} I(x(t), y(t), t) = 0. \quad (2.23)$$

Eq. (2.23) is termed the *conservation of brightness*. Expanding the left-hand side of (2.23) by the chain rule yields the optical flow constraint (2.21).

Although acceleration is not considered in the sequel, it is instructive to point out that differentiating (2.23) once again with respect to  $t$ ,

$$\frac{d^2}{dt^2}I(x(t), y(t), t) = 0, \quad (2.24)$$

leads to a motion constraint equation relating the image structure with optical flow and optical acceleration, given as follows (Arnsperg, 1988),

$$(I_{xx}u + I_{xy}v + I_{xt})u + I_xa + (I_{yx}u + I_{yy}v + I_{yt})v + I_yb + (I_{tx}u + I_{ty}v + I_{tt}) = 0 \quad (2.25)$$

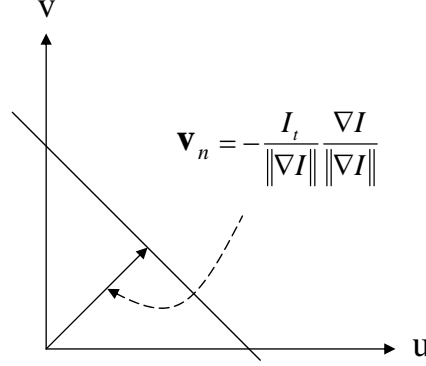
where the optical flow is represented by  $(u, v) = (\frac{dx}{dt}, \frac{dy}{dt})$  and the optical acceleration by  $(a, b) = (\frac{d^2x}{dt^2}, \frac{d^2y}{dt^2})$ .

Given a single point, the optical flow constraint (2.21) is underconstrained since it represents a single linear constraint in two unknowns. The optical flow constraint admits a continuum of possible solutions that lie along a line in velocity space (see Fig. 2.5). A special solution along this line is the component of motion in the direction of the spatial gradient of the intensity function, termed the *normal flow* (Horn, 1986),

$$\mathbf{v}_n = -\frac{I_t}{\|\nabla I\|} \frac{\nabla I}{\|\nabla I\|} \quad (2.26)$$

where  $I_t$  represents the partial derivative of the image brightness function with respect to time  $t$ ,  $\nabla I$  represents the spatial gradient and  $\|\nabla I\|$  represents the  $\ell^2$ -norm of the gradient. The normal flow vector lies perpendicular to the velocity constraint line or equivalently stated is the velocity vector with the smallest magnitude (i.e., speed) that lies along the constraint line. The component of motion perpendicular to the spatial gradient remains to be recovered. The underconstrained nature of the optical flow constraint represents an instance of the *aperture problem*.

The aperture problem (Marr & Ullman, 1979; Ullman, 1979; Adelson & Movshon, 1982) is a common problem that all motion recovery approaches face. It occurs within regions of analysis where the brightness pattern is only a function of one image coordinate and not the other. In such cases, the motion in the direction of the spatial image gradient can only be resolved while the motion in the orthogonal direction remains ambiguous. Figure 2.6 illustrates regions afflicted by the aperture problem and a region where unambiguous flow recovery can be made. Analytically, the aperture problem appears in the situation where the number of unknown variables is greater than the number of given constraints. An open problem related to the aperture problem is in selecting the appropriate aperture size such that the aperture problem disappears while at the same time the given assumptions hold, Jepson and Black (Jepson & Black, 1993a) term this problem the *generalized aperture problem*.



**Figure 2.5.** Optical flow constraint equation in velocity space. The optical flow constraint equation in velocity space  $u, v$  is depicted. The normal velocity  $\mathbf{v}_n$  represents the component of the optical flow in the direction of the gradient of the spatial intensity structure.

### Motion as orientation in spatiotemporal images

For the case of translational motion where an initial image  $I_0$  is translated with constant velocity over time while respecting the brightness constancy assumption over time, formally,

$$I(x, y, t) = I_0(x - ut, y - vt) \quad (2.27)$$

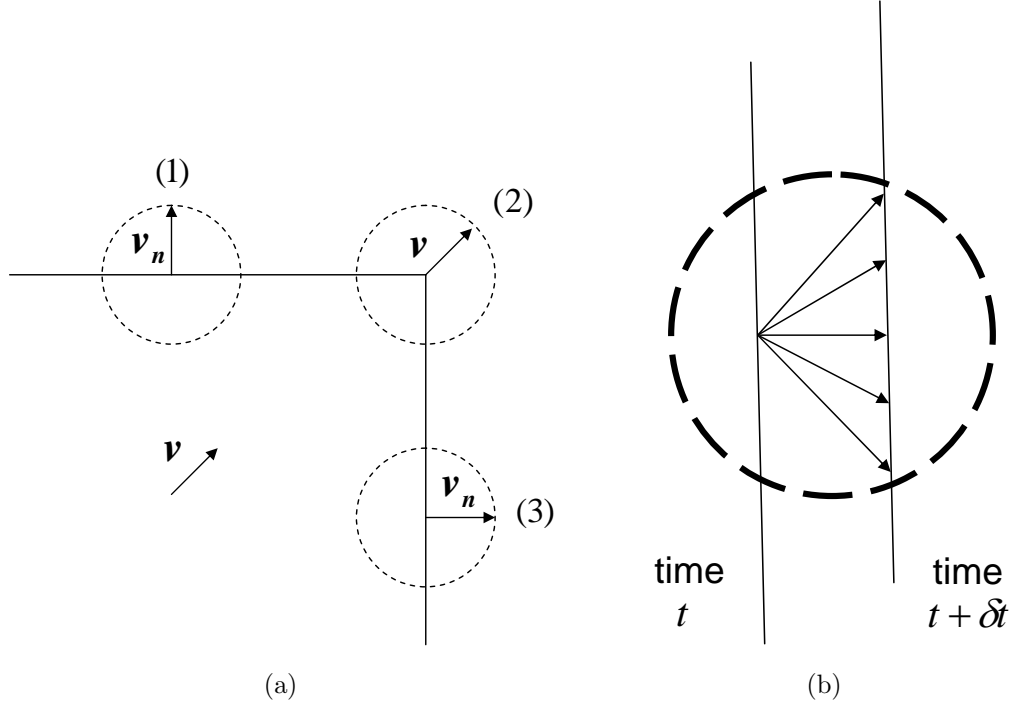
the velocity can be directly linked to the orientation in the spatiotemporal volume or slice (Fahle & Poggio, 1981; Watson & Ahumada, 1985; Adelson & Bergen, 1985; Jähne, 1990). In the case of a two-dimensional space-time image (one spatial dimension and time) the velocity  $u$  is given as (see Fig. 2.7 and Fig. 2.8 (a)),

$$u = -\tan(\theta), \quad (2.28)$$

where  $\theta$  represents the angle between the time axis  $t$  and the direction that the grey values are constant. More generally, in the case of a three-dimensional space-time volume the velocity  $(u, v)^\top$  is given as (see Fig. 2.8 (b)),

$$\begin{pmatrix} u \\ v \end{pmatrix} = - \begin{pmatrix} \tan(\theta_x) \\ \tan(\theta_y) \end{pmatrix}, \quad (2.29)$$

where the angles  $\theta_x$  and  $\theta_y$  represent the angles between the time axis  $t$  and the direction that the grey values are constant in terms of the  $x$  and  $y$  axes, respectively. Thus, the problem of velocity estimation can be recast as a problem of orientation estimation. Although we have limited the motion to constant velocity, one could generalize (2.27) to include acceleration in which case the space-time surfaces would exhibit curvature. Thus, the problem of acceleration estimation may be cast as problem of curvature estimation.



**Figure 2.6.** Aperture problem. (a) regions (1),(3) denote where the aperture problem is present, while at region (2) a full estimate can be made since there is distinct local structure. (b) depicts the uncertainty in a region where the aperture problem is present. An isobrightness line is observed through an aperture at time  $t$ . At time  $t + \delta t$  the line has moved to a new position. From the information contained within the aperture only the component perpendicular to the line can be recovered. The tangential component to the line cannot be recovered.

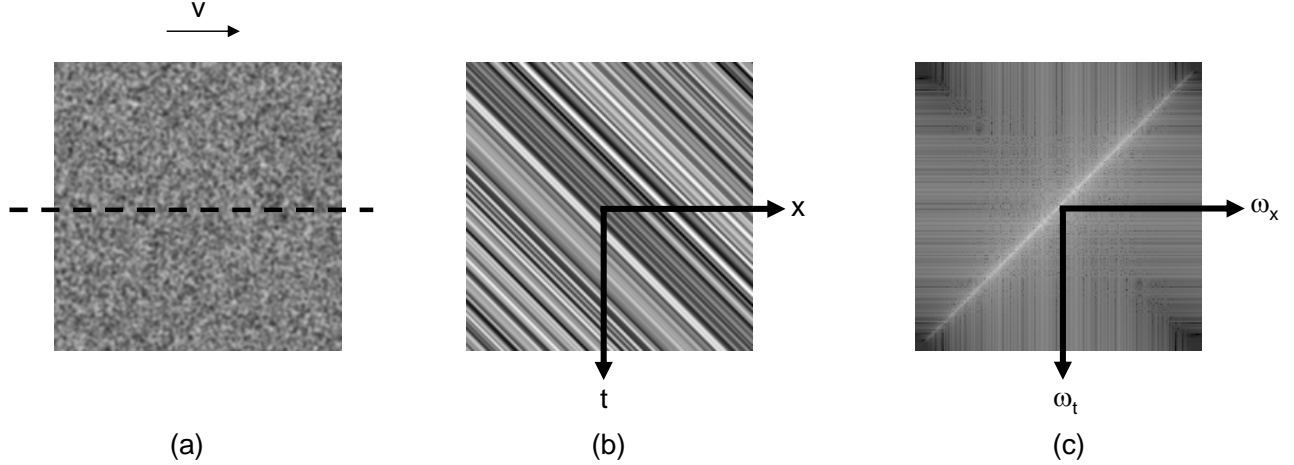
### Motion in the frequency domain

Assume as in the volumetric representation that an initial image  $I_0$  is translated with constant velocity over time while respecting the brightness constancy assumption from frame-to-frame, formally,

$$I(x, y, t) = I_0(x - ut, y - vt). \quad (2.30)$$

The three-dimensional Fourier transform of (2.30) is,

$$\hat{I}(\omega_x, \omega_y, \omega_t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I_0(x - ut, y - vt) e^{-j2\pi(x\omega_x + y\omega_y + t\omega_t)} dx dy dt. \quad (2.31)$$



**Figure 2.7.** A white noise pattern moving to the right at a pixel/frame in the spatial, spatiotemporal and frequency domains is depicted: (a) a single frame of the sequence (the spatial domain) (b) the spatiotemporal X-T slice of (a) delineated by the dashed line and (c) the spatiotemporal slice in the frequency domain.

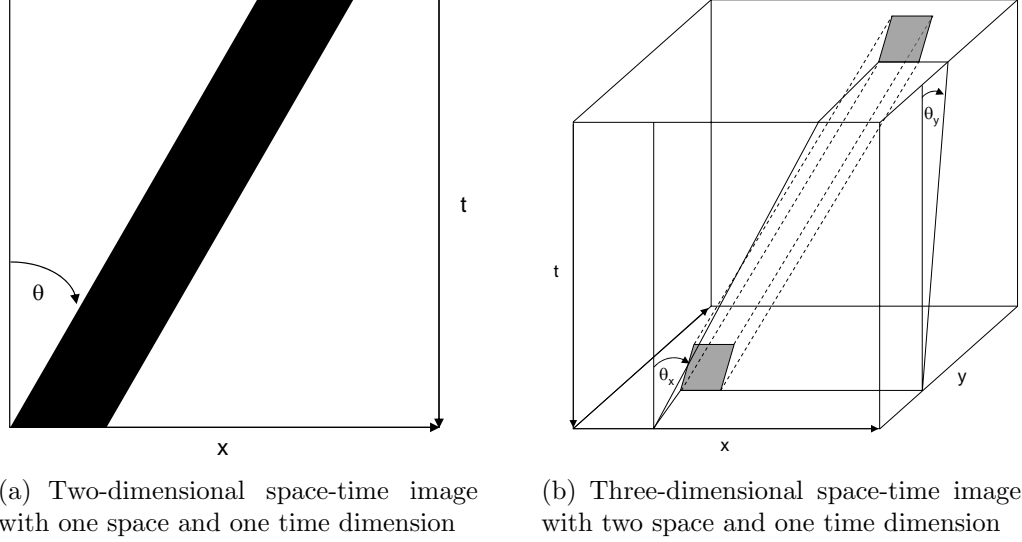
Utilizing properties of the Fourier transform in addition to some algebraic manipulation one can arrive at the following,

$$\hat{I}(\omega_x, \omega_y, \omega_t) = \hat{I}_0(\omega_x, \omega_y) \delta(\omega_x u + \omega_y v + \omega_t) \quad (2.32)$$

where  $\delta(\cdot)$  represents the *Dirac delta* function (Oppenheim et al., 1997). This equation algebraically expresses that the non-zero components of the Fourier transform of a translating scene lie on a plane with normal  $(u, v, 1)^\top$  that passes through the origin (see Fig. 2.9 (a)). In the case of a two-dimensional space-time image, one spatial dimension and time, its Fourier spectrum lies along a line through the origin with slope  $-1/v_x$  (see Fig. 2.7 and Fig. 2.9 (b)). This insight forms the basis for several motion estimation algorithms that estimate velocity by identifying the orientation of this plane/line (Fahle & Poggio, 1981; Watson & Ahumada, 1985; Adelson & Bergen, 1985; Heeger, 1988; Jähne, 1990; Simoncelli, 1993b). If within the region of analysis the spatial image structure is a function of only one of the spatial parameters the planar spectra will collapse to a line passing through the origin. In such cases the full velocity will not be recoverable, instead only the normal component of the velocity will be recoverable, since a continuum of planes and by extension a continuum of velocities will be consistent with this line spectra. This situation represents an instance of the aperture problem in the frequency domain.

Lastly, let us consider two neighbouring frames at times  $t$  and  $t - 1$  displaced spatially with respect to each other by  $(u, v)$ , formally,

$$I(x, y, t) = I(x - u, y - v, t - 1). \quad (2.33)$$



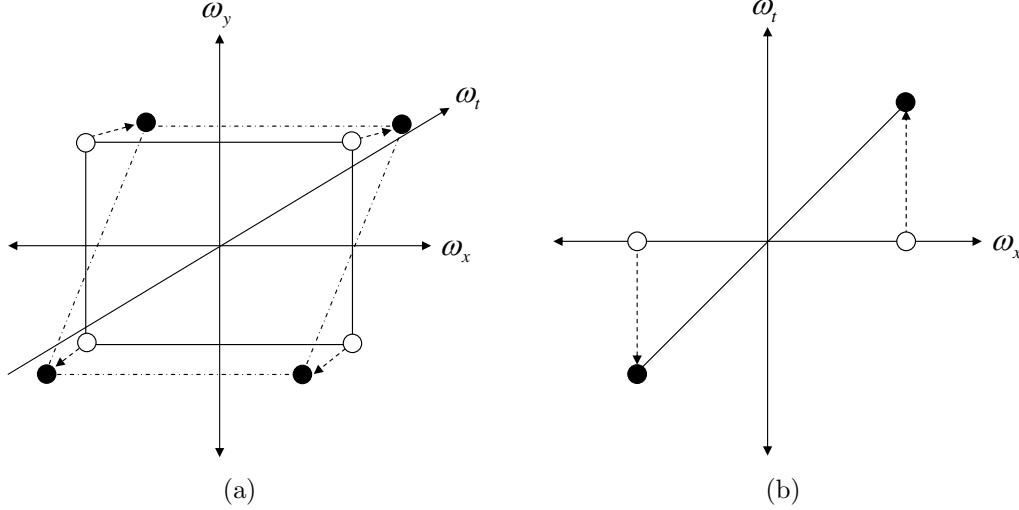
**Figure 2.8.** Motion as orientation. Adapted from (Jähne, 2005).

Utilizing the *shift property* of the Fourier transform (Lim, 1990), the spatial Fourier transform of these two frames are related by a linear phase shift, as follows,

$$\hat{I}(\omega_x, \omega_y, t) = \hat{I}(\omega_x, \omega_y, t-1)e^{-j2\pi(\omega_x u + \omega_y v)} \quad (2.34)$$

where  $\hat{I}(\omega_x, \omega_y, t)$  and  $\hat{I}(\omega_x, \omega_y, t-1)$  denote the spatial Fourier transforms of  $I(x, y, t)$  and  $I(x, y, t-1)$ , respectively.

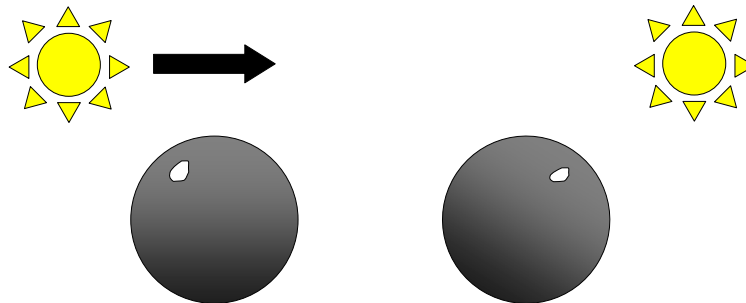
An important point to keep in mind about the frequency analyses presented above is that it is global in nature (i.e., infinite spatiotemporal extents). In practice, we are interested in analyzing local spatiotemporal regions. This is achieved by windowing the local region by a preferably smooth function, such as a Gaussian. This spatiotemporal windowing process is equivalent to convolving the spectrum of the image with the spectrum of the window. Hence, in the local analysis of motion, idealized structures in the frequency domain are blurred. To reduce the blurring one could increase the window size at the risk of violating assumptions imposed on the region. Furthermore, increasing the window size reduces the ability to resolve the location of the motion type. This is a consequence of the *Heisenberg-Gabor uncertainty principle* that states that one can not simultaneously localize a signal in time (or in our case in space-time) and frequency well (Gabor, 1946).



**Figure 2.9.** Motion in the Fourier domain. (a) depicts the effect of motion on a 3D space-time image (two spatial and one temporal component) in the Fourier domain. The plane outlined by the solid line lying within the  $\omega_x - \omega_y$  plane represents the spectrum of a static scene; the open circles denote points in the plane. The plane outlined by the dotted line represents the spectrum as the scene translates with velocity  $\mathbf{v}$ . Motion results in a shearing displacement (indicated by the dotted arrows) of the static plane. As a result of the shearing displacement, the open circles are mapped to the black circles. Furthermore, the plane intersects the origin. (b) depicts the effect of motion on a 2D space-time image (one spatial and one temporal component) in the Fourier domain. The open circles represent a single frequency subtending a line spectrum representing a static scene. The solid black circles represent the locations of the open circles as the image translates with speed  $s$ . Each open circle undergoes a shearing displacement (indicated by the dotted arrows) with the line spectrum crossing the origin.

### 2.2.3 Motion field versus optical flow

In the ideal case the optical flow would correspond to the *motion field* (see Chapter 2.2.1) which represents the projection of the three-dimensional velocity of objects in a scene onto the image plane. In such a case, the quantitative measurements of the three-dimensional scene structure may be recovered. In general, the optical flow does not equal the motion field. The reason for the lack of strict equivalence between the two flow fields lies in the fact that the motion field is a purely geometric construct, whereas optical flow depends on assumptions placed on the structure of the brightness and velocity of the image sequence. Examples for which the definition of optical flow based on the brightness constancy assumption does not coincide with the motion field are well known (Horn, 1986). Consider a moving blank wall, the motion field is non-zero, while the optical flow is zero everywhere since the image brightness does not change (Horn, 1986). Figure 2.10 illustrates an additional example where equivalence between the motion field and optical flow does not hold. A stationary textureless



**Figure 2.10.** Example of the optical flow field not always equal to the motion field. A stationary textureless sphere illuminated by a moving light source is depicted. Although, the resulting shading in the image changes leading to a non-zero optical flow, the motion field is zero. Adapted from ([Horn, 1986](#)).

sphere illuminated by a moving light source is depicted, where the image brightness changes as the illumination source moves. In this case, the shading changes resulting in a non-zero optical flow, while the motion field is zero everywhere ([Horn, 1986](#)).

Verri and Poggio ([Verri & Poggio, 1987a](#); [Verri & Poggio, 1989](#)) demonstrate that the optical flow under a perspective projection model and the assumption of brightness constancy, is generally different from the motion field unless the following special conditions are satisfied:

- the illumination is given by a distant point light source such that the direction of the light rays are parallel
- the imaged surface is Lambertian (i.e., the surface is perfectly matte such that the luminance is the same regardless of the viewing angle)
- the imaged surface has sufficient texture
- the imaged surface translates or rotates about an axis parallel to the direction of illumination

These conditions ensure that the underlying surface radiance remains fixed. In reality, these conditions rarely hold but empirically have proven to provide a good local approximation ([Barron et al., 1994b](#)).

### 2.2.4 Revisiting the brightness constancy assumption

In Section 2.2.2 the assumption of brightness constancy (conservation) was imposed on the spatiotemporal structure of the imaged scene in order to derive a constraint on the the optical flow (2.21). As noted, the optical flow constraint has two main deficiencies:



1. the validity of the constraint only holds under fairly restrictive conditions (see Section 2.2.3)
2. we are faced with an underconstrained problem in the form of the aperture problem (see Section 2.2.2).

This section reviews further “optical flow” constraints to address these issues arrived at by way of alternative or complimentary assumptions to brightness constancy imposed on the spatiotemporal structure and assumptions of the structure of the velocity field. The success of flow recovery approaches based on these “optical flow” constraints hinges on the validity of the assumptions used to derive the respective constraints.

Cornelius and Kanade ([Cornelius & Kanade, 1983](#)) relax the assumption of conservation of brightness (2.23) by allowing for a linearly additive change (with respect to time) in the intensity of a surface patch as it moves relative to the camera, formally,

$$\frac{d}{dt}I(x(t), y(t), t) = c, \quad (2.35)$$

where  $c$  is a constant. In the case where  $c = 0$ , (2.35) reduces to the conservation of brightness assumption (2.23). Evaluating the left-hand side of (2.35), yields,

$$I_x u + I_y v + I_t = c. \quad (2.36)$$

Assuming *Phong’s shading model* (diffuse and specular lighting effects) ([Phong, 1975](#)) adequately represents the radiance of the scene, Mukawa ([Mukawa, 1989](#)) derived an optical flow constraint equation identical to (2.36). Nagel ([Nagel, 1989](#)) considered the geometry of a locally rigid scene and the radiometric effects of an optical system, while assuming isotropic constant lighting and Lambertian surfaces and arrived at the following constraint,

$$I_x u + I_y v + I_t = c \quad (2.37)$$

where

$$c = 4I \left( \frac{\hat{\mathbf{z}} \dot{\mathbf{P}}^\top}{\hat{\mathbf{z}} \mathbf{P}^\top} - \frac{\mathbf{P} \dot{\mathbf{P}}^\top}{\mathbf{P} \mathbf{P}^\top} \right), \quad (2.38)$$

$\mathbf{P}$  is a three-dimensional scene point,  $\dot{\mathbf{P}}$  the corresponding velocity and  $\hat{\mathbf{z}}$  is a unit vector in the direction of the optical axis.

More generally, Gennert and Negahdaripour ([Gennert & Negahdaripour, 1987](#)) and Negahdaripour and Yu ([Negahdaripour & Yu, 1993](#)) propose the following *generalized brightness change model*,

$$I(x - \delta x, y - \delta y, t - \delta t) = M(x, y, t)I(x, y, t) + C(x, y, t), \quad (2.39)$$

where  $M(x, y, t)$  (contrast change) and  $C(x, y, t)$  (mean intensity shift) incorporate non-motion related brightness changes of a point through a linear transformation. Assuming that both  $M$  and  $C$  are linear functions with respect to time  $t$ , Negahdaripour and Yu (Negahdaripour & Yu, 1993) derive a motion constraint equation in an analogous fashion to the derivation of the optical flow constraint (2.21) by taking a Taylor series expansion of the left-hand side of (2.39) followed by taking the limit as  $\delta t \rightarrow 0$ , yielding,

$$I_x u + I_y v + I_t = m_t I(x, y, t) + c_t, \quad (2.40)$$

where  $m_t$  and  $c_t$  are the temporal changes of  $M$  and  $C$ , respectively. Lai and Fang (Lai & Fang, 1999) generalize the two illumination factors  $M$  and  $C$  by assuming that they are slowly varying functions of the spatial parameters  $(x, y)$  modeled by a low-order polynomial expansion.

Haußecker and Fleet (Haußecker & Fleet, 2000; Haußecker & Fleet, 2001) propose a generalization of the brightness constancy assumption by defining a path  $\mathbf{x}(t)$  where the brightness changes according to a parametric function  $h$  that models a time-dependent physical process, formally,

$$\frac{dI}{dt} I(x(t), y(t), t) = \frac{d}{dt} h(I(x(0), y(0), 0), t, \mathbf{a}), \quad (2.41)$$

where  $\mathbf{a}$  contains the parameters of the model. Expanding the total derivative on the left-hand side of (2.41), yields,

$$I_x u + I_y v + I_t = \frac{d}{dt} h(I(x(0), y(0), 0), t, \mathbf{a}). \quad (2.42)$$

Haußecker and Fleet term (2.42) the *generalized brightness change constraint equation*. Note that when  $h$  is constant with respect to time, (2.42) reduces to the optical flow constraint (2.21). This model represents a more general formulation of constraints (2.36) and (2.39) in that it allows for the modeling of higher-order brightness changes over a region of time (i.e., sequence of images) as opposed to the linear brightness change model assumed locally in (2.39). Physical models encapsulated by  $h$  considered by Haußecker and Fleet include: changing surface orientation with respect to a directional illuminant, a moving illuminant, and physical models of heat transport (diffusion and decay) in infrared images.

Next we turn our attention to a set of motion constraints based on the assumption of the conservation of analytical quantities of the image brightness for elementary deformations of the image. These elementary deformations consist of translation, curl (rate of rotation), divergence (rate of expansion/constriction) and shear (area preserving deformation). Excluding translation, the remaining deformations are encapsulated in the first-order spatial derivatives of the flow field  $(u, v)$ , given in matrix form by the *Jacobian* matrix,  $\mathbf{J}$ ,

$$\mathbf{J} = \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix}. \quad (2.43)$$

To bring the following set of constraints into a common framework of understanding, the constraints are driven by assumptions and specializations imposed on the identity:

$$\nabla \frac{dI}{dt} = \frac{d}{dt} \nabla I + \mathbf{J}^\top \nabla I. \quad (2.44)$$

where  $\nabla$  represents the gradient operator  $\nabla = (\partial/\partial x, \partial/\partial y)$ . Note that no assumptions are present in (2.44) (other than differentiability), such as brightness constancy. This common framework is adapted from (Verri et al., 1990). The first assumption imposed on (2.44) and assumed throughout the following constraints is the conservation of brightness (i.e.,  $dI/dt = 0$ ), (2.44) becomes,

$$\frac{d}{dt} \nabla I + \mathbf{J}^\top \nabla I = 0. \quad (2.45)$$

Though not considered here (and in any paper for that matter), one could relax the brightness constancy assumption imposed in (2.44) and instead assume that brightness changes according to some parameterized function, such as in (2.41). The next set of constraints are based on restricting the motion of the flow through specializations of the Jacobian. Assuming that the Jacobian,  $\mathbf{J}$ , in (2.45) represents only diverging flow and with some algebraic manipulation, (2.45) can be written as (Verri et al., 1990),

$$\frac{d}{dt} \left( \frac{I_y}{I_x} \right) = 0, \text{ if } I_x \neq 0, \quad (2.46)$$

which algebraically expresses that the direction of the spatial gradient is conserved. This constraint holds true for textured planar patterns that are parallel to the image plane and move parallel to the optical axis (Verri et al., 1990). Assuming that the Jacobian  $\mathbf{J}$  in (2.45) represents only a rotating flow, (2.45) can be written as (Verri et al., 1990),

$$\frac{d}{dt} (I_x^2 + I_y^2) = 0, \quad (2.47)$$

which algebraically expresses that the magnitude of the spatial gradient is conserved. Assuming that the Jacobian,  $\mathbf{J}$ , in (2.45) represents only a shearing flow, (2.45) can be written as (Verri et al., 1990),

$$\frac{d}{dt} (I_x I_y) = 0, \quad (2.48)$$

or

$$\frac{d}{dt} (I_x^2 - I_y^2) = 0. \quad (2.49)$$

Assuming that “enough” image structure is present, each of the constraints (2.46)-(2.49) can be combined with the optical flow constraint (2.21) to yield a fully constrained solution of the flow  $(u, v)$ , thus resolving the aperture problem. So far each

of the assumptions resulted in  $\mathbf{J} \neq \mathbf{0}$ . Assuming that  $\mathbf{J} = \mathbf{0}$  (i.e., only translational motion is assumed), which is the case for a planar patch that is parallel to the image plane and translating parallel to the image plane, (2.45) reduces to,

$$\frac{d}{dt} \nabla I(x, y, t) = \mathbf{0}, \quad (2.50)$$

which states that the gradient is conserved. Further expansion of the total derivative on the left-hand side of (2.50) yields the following constraints,

$$I_{xx}u + I_{yx}v + I_{tx} = 0 \quad (2.51)$$

$$I_{xy}u + I_{yy}v + I_{ty} = 0 \quad (2.52)$$

or more compactly,

$$\mathbf{H}\mathbf{u} = -\nabla I_t \quad (2.53)$$

where  $\mathbf{H}$  represents the spatial Hessian matrix,

$$\mathbf{H} = \begin{bmatrix} I_{xx} & I_{yx} \\ I_{xy} & I_{yy} \end{bmatrix}. \quad (2.54)$$

This system can be solved whenever the Hessian matrix (2.54) is invertible (i.e.,  $\det \mathbf{H} \neq 0$ ). Note that the determinant of the Hessian,  $\mathbf{H}$ , represents the Gaussian curvature of the intensity surface of the image, thus good candidate points are those that exhibit large intensity curvature in two principle directions. In addition, one can estimate the optical flow by coupling (2.51) and (2.52) with the optical flow constraint (2.21) which results in an over-determined system of equations in the flow components  $(u, v)$ . Constraints (2.51) and (2.52) have repeatedly appeared in the literature through various diverse constructions in an attempt to resolve the aperture problem through local means (see (Haralick & Lee, 1983; Tretiak & Pastor, 1984; Nagel, 1987; Giroi et al., 1989; Verri & Poggio, 1989; Uras et al., 1989; Sobey & Srinivasan, 1991; Simoncelli, 1993b; Tistarelli, 1994; Tistarelli, 1995; del Bimbo et al., 1996)). These constraints can be seen as the application of “brightness” constancy on derived measurements of the image in the form of first derivative output images (as discussed in Section 2.1.2). A drawback of these higher-order approaches is that reliable estimates of second- and higher-order derivatives from images is problematic due to the high-pass nature of the operators.

Several authors have derived “optical flow” constraints based on the analogy of optical flow as a fluid flow, as studied in fluid mechanics, where the image brightness is considered as density. The optical flow constraint (2.21) is equivalent to the *continuity equation for an incompressible fluid* (i.e., conservation of density). A drawback of the optical flow constraint (2.21) is that it is valid for a restricted class of scene motions. For instance, the optical flow constraint is invalid for scene motion consisting of

rotation out of the plane of the imaging surface. To address this issue, Schunck (Schunck, 1984; Schunck, 1985) proposed that the image adhered to the *conservation of mass*<sup>1</sup> (i.e., total brightness), resulting in a more general constraint equation,

$$I_x u + I_y v + I_t + I \operatorname{div}(u, v)^\top = 0 \quad (2.56)$$

where  $\operatorname{div}(u, v)^\top$  represents the divergence of the flow field. Eq. (2.56) is termed the *extended optical flow constraint* (cf., the *continuity of mass equation*). From a structural point of view, the extended optical flow constraint (2.56) differs from the optical flow constraint (2.21) by the addition of the divergence term. Note that both constraints are brought into agreement if the motion is parallel to the image plane in which case the term containing the divergence vanishes. The validity of Schunck's analogy is questionable given the lack of an analytical or experimental justification. In the context of measuring fluid flow from image sequences, Wildes et al. (Wildes et al., 1997; Wildes et al., 2000) demonstrate that the two-dimensional image of a three-dimensional fluid flow that conforms to the conservation of mass in three-dimensions results in a two-dimensional flow that conforms to the conservation of mass in two-dimensions and in turn to (2.56), provided that there is no material loss due to normal flow across the boundaries of the fluid. Keeping with the analogy of optical flow as a fluid, a logical next step would be the introduction of the *conservation of momentum* assumption (Aris, 1989). To this author's knowledge no such work has been reported.

In the problem of image tracking, the recovery of optical flow or displacement represents an intermediate step for frame-to-frame registration. In this context, several constancy assumptions have been introduced. Black and Jepson (Black & Jepson, 1998), propose the *subspace constancy assumption* which formalizes the notion of appearance constancy with respect to a “learned” eigenspace representation (Oja, 1983), formally,

$$I(\mathbf{x} + u(\mathbf{x}, \mathbf{a}), y + v(\mathbf{x}, \mathbf{a})) = (\mathbf{U}\mathbf{c})(\mathbf{x}), \quad \forall \mathbf{x} \text{ within the tracked region} \quad (2.57)$$

where  $(u(\mathbf{x}, \mathbf{a}), v(\mathbf{x}, \mathbf{a}))$  represents an image transformation (or motion) parameterized by  $\mathbf{a}$ ,  $\mathbf{U}$  is a matrix that encapsulates the eigenspace (subspace) basis,  $\mathbf{c}$  is a vector of basis coefficients and  $(\mathbf{U}\mathbf{c})(\mathbf{x})$  is defined as the value of  $(\mathbf{U}\mathbf{c})$  at pixel position

---

<sup>1</sup> The *conservation of mass* requires that the rate of accumulation of mass within the volume plus the net rate of outflow of mass from the delimiting surface be zero (Aris, 1989), formally,

$$\underbrace{\int_V \frac{\partial I}{\partial t} dV}_{\text{rate of accumulation of mass}} + \underbrace{\int_V \operatorname{div}(I\mathbf{u}) dV}_{\text{net rate of outflow of mass}} = 0 \quad (2.55)$$

$\mathbf{x} = (x, y)$ . Expanding, the left-hand side of (2.57) using a first-order Taylor series expansion and reorganizing terms leads to the following differential constraint,

$$I_x u(\mathbf{x}, \mathbf{a}) + I_y v(\mathbf{x}, \mathbf{a}) + (I(\mathbf{x}) - (\mathbf{U}\mathbf{c})(\mathbf{x})) = 0. \quad (2.58)$$

Notice the similarity of (2.58) with the optical flow constraint (2.21),  $(I(\mathbf{x}) - (\mathbf{U}\mathbf{c})(\mathbf{x}))$  takes the place of the temporal derivative  $I_t$ . Hager and Belhumeur (Hager & Belhumeur, 1996; Hager & Belhumeur, 1998) relax the brightness constancy assumption (2.18) by allowing an additive deviation that models explicitly illumination changes such as shadows and changing lighting. Illumination changes are modeled as a linear combination of basis vectors  $\mathbf{b}_i, i = 1, \dots, N$ . The basis vectors  $\mathbf{b}_i$  are “learnt” offline using principle components analysis (Oja, 1983) on a set of training exemplars of the target region taken under varying illumination conditions. Hager and Belhumeur’s conservation assumption is given as follows,

$$I(x+u(\mathbf{x}, \mathbf{a}), y+v(\mathbf{x}, \mathbf{a}), t+1) = I(x, y, t=0) + (\mathbf{B}\mathbf{c})(\mathbf{x}), \quad \forall \mathbf{x} \text{ within the tracked region} \quad (2.59)$$

where  $(u(\mathbf{x}, \mathbf{a}), v(\mathbf{x}, \mathbf{a}))$  represents an image transformation (or motion) parameterized by  $\mathbf{a}$ ,  $I(x, y, t=0)$  denotes the reference template,  $\mathbf{B} = [\mathbf{b}_1 | \mathbf{b}_2 | \dots | \mathbf{b}_N]$  is a matrix of basis vectors and  $\mathbf{c} = (c_1, c_2, \dots, c_N)^\top$  a vector of basis coefficients. Expanding the left-hand side of (2.2.4) via a first-order Taylor series expansion and reorganizing terms leads to the following differential constraint,

$$I_x u(\mathbf{x}, \mathbf{a}) + I_y v(\mathbf{x}, \mathbf{a}) + (I(\mathbf{x}, t+1) - I(0, t_0)) = (\mathbf{B}\mathbf{c})(\mathbf{x}). \quad (2.60)$$

Notice the similarity of (2.60) with Cornelius and Kanade’s linear brightness change differential constraint (2.36),  $(I(\mathbf{x}, t+1) - I(0, t_0))$  has replaced the temporal derivative and the model of brightness change  $(\mathbf{B}\mathbf{c})(\mathbf{x})$  has replaced the linear brightness term  $c$ . A drawback of these tracking approaches is the requirement of training prior to tracking in order to learn the appearance basis.

Table 2.2 provides a summary of the conservation assumptions outlined in this section and their corresponding differential constraints. To reiterate, the success of flow recovery approaches based on any of these constraints depends on the validity of their respective underlying assumptions highlighted in this section.

Conservation Assumption	Differential “Optical flow” constraint
Brightness (cf., density): $\frac{d}{dt}I = 0$	$I_x u + I_y v + I_t = 0$
Generalized brightness change: $I(\mathbf{x} - \delta\mathbf{x}, t - \delta t) = M(\mathbf{x}, t)I(\mathbf{x}, t) + C(\mathbf{x}, t)$	$I_x u + I_y v + I_t = m_t I(x, y, t) + c_t$
Total brightness (cf., mass): $\int_V \frac{\partial I}{\partial t} dV + \int_V \text{div}(I\mathbf{u}) dV = 0$	$I_x u + I_y v + I_t + I \text{div}(u, v)^\top = 0$
Gradient direction: $\frac{d}{dt}(I_x/I_y) = 0$	$(I_y I_{xx} - I_x I_{xy})u + (I_y I_{xy} - I_x I_{yy})v$ $+ I_y I_{xt} - I_x I_{yt} = 0$
Gradient magnitude: $\frac{d}{dt}(I_x^2 + I_y^2) = 0$	$(I_x I_{xx} + I_y I_{xy})u + (I_x I_{xy} + I_y I_{yy})v$ $+ I_x I_{xt} + I_y I_{yt} = 0$
Product of gradient components: $\frac{d}{dt}(I_x I_y) = 0$	$(I_x I_{xy} - I_y I_{xx})u + (I_x I_{yy} - I_y I_{xy})v$ $+ I_x I_{yt} - I_y I_{xt} = 0$
Difference of squared gradient components: $\frac{d}{dt}(I_x^2 - I_y^2) = 0$	$(I_x I_{xx} - I_y I_{xy})u + (I_x I_{xy} - I_y I_{yy})v$ $+ I_x I_{xt} - I_y I_{yt} = 0$
Gradient: $\frac{d}{dt}\nabla I = \mathbf{0}$	$I_{xx}u + I_{xy}v + I_{xt} = 0$ $I_{xy}u + I_{yy}v + I_{yt} = 0$
Subspace: $I(\mathbf{x} + \mathbf{u}(\mathbf{x}, \mathbf{a}), y + v(\mathbf{x}, \mathbf{a})) = (\mathbf{Uc})(\mathbf{x})$	$I_x u(\mathbf{x}, \mathbf{a}) + I_y v(\mathbf{x}, \mathbf{a}) + (I(\mathbf{x}) - (\mathbf{Uc})(\mathbf{x})) = 0$
Subspace brightness change: $I(\mathbf{x} + \mathbf{u}(\mathbf{x}, \mathbf{a}), t + 1) = I(\mathbf{x}, t = 0) + (\mathbf{Bc})(\mathbf{x})$	$I_x u(\mathbf{x}, \mathbf{a}) + I_y v(\mathbf{x}, \mathbf{a}) + (I(\mathbf{x}, t + 1) - I(0, t_0))$ $= (\mathbf{Bc})(\mathbf{x})$

**Table 2.2.** Summary of conservation assumptions and their corresponding differential “optical flow” constraints.

### 2.2.5 Multiple motion representations

This section focuses on spatiotemporal and frequency representations of multiple motions at a point or within a region. The multiple motion cases considered below are as follows: *transparency* (i.e., additive transparency), *translucency* (i.e., multiplicative transparency), *occlusion* and *optical snow*. Note we limit the discussion to constant velocity motion, accelerated motion is beyond the scope of this discussion but may be found in (Balanza & Cortelazzo, 1989). Also, we consider two spatial dimensions and time, similar results may be developed for the case of one spatial dimension and time.

In Section 2.1.1 the description of the *constant velocity model* that observes brightness constancy was given in the spatiotemporal domain as,

$$I(\mathbf{x}, t) = I(\mathbf{x} - \mathbf{v}t), \quad (2.61)$$

and in the frequency domain, as,

$$\hat{I}(\omega, \omega_t) = \hat{I}(\omega)\delta(\omega \cdot \mathbf{v} + \omega_t), \quad (2.62)$$

where  $\mathbf{x} = (x, y)^\top$  represents the spatial variables,  $\omega = (\omega_x, \omega_y)^\top$  the spatial frequency variables and  $\mathbf{v} = (u, v)$  the velocity of the moving surface. Eq. (2.62) may be geometrically interpreted as an impulse plane passing through the origin with normal  $(u, v, 1)$ . These two representations of the constant velocity model play fundamental roles in the development of the multiple motion cases to follow. More specifically, the spatiotemporal representation of the multiple motion cases are based on compositions of multiple layers, each of which is considered as a constant velocity model. The key differentiating aspects among the model descriptions involve the combination rule and the spatial description of the layers (i.e., compact versus global in extent). A common assumption among the multiple motion cases detailed below is that the constituent layers observe brightness constancy,  $dI_i(x, y, t)/dt = 0$  for  $i = 1, 2, \dots, n$ .

In the case where the imaged surfaces are transparent, more than one motion may be present at a single image point. For two image sequences that have been additively combined<sup>2</sup>, the spatiotemporal structure may be modeled, as follows,

$$I(\mathbf{x}, t) = \alpha I(\mathbf{x} - \mathbf{v}_1 t) + (1 - \alpha)I(\mathbf{x} - \mathbf{v}_2 t), \quad (2.63)$$

where  $\alpha \in (0, 1)$ . Taking the Fourier transform of (2.63), yields,

$$\hat{I}(\omega, \omega_t) = \alpha \hat{I}(\omega, \omega_t)\delta(\omega \cdot \mathbf{v}_1 + \omega_t) + (1 - \alpha)\hat{I}(\omega)\delta(\omega \cdot \mathbf{v}_2 + \omega_t) \quad (2.64)$$

---

<sup>2</sup>In the case where the transparency is multiplicative, one can simply take the logarithm of the image sequence which results in an additive transparency of the logarithm of the individual image sequences. However, the application of the logarithm has numerical implications due to the compression of the dynamic range.



Geometrically, the frequency spectrum corresponds to the superposition of the oriented planes from the respective layers. Shizawa and Mase (Shizawa & Mase, 1990; Shizawa & Mase, 1991a; Shizawa & Mase, 1991b) introduce the chaining of the linear operators used in the optical flow constraint<sup>3</sup> (2.21), to form the following constraint,

$$\left( \prod_{i=1}^2 (\mathbf{v}_i, 1)^\top \cdot \nabla \right) I(x, y, t) = 0 \quad (2.67)$$

where  $I(x, y, t) = \sum_{i=1}^2 I_i(x, y, t)$  represents the superposition of the two images  $I_i(x, y, t)$ ,  $\mathbf{v}_i = (u_i, v_i)$  denotes the image velocity in the  $i$ th layer and  $\nabla = \left( \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial t} \right)^\top$ . This constraint may be generalized to  $N$  layers by simply extending the limit of the product of operators to  $N$ . Langley et al. (Langley et al., 1992) apply this same model to the phase-based representation of the image sequence. Interestingly, Mulligan (Mulligan, 1992) has reported that human observers can easily perceive two coherently moving patterns (layers) of white noise that have been combined additively. However, beyond two layers, humans are no longer able to perceive all layers simultaneously.

*Translucency motion* considers the multiplicative combination of moving layers, formally,

$$I(x, y, t) = I_1(\mathbf{x} - \mathbf{v}_1 t) I_2(\mathbf{x} - \mathbf{v}_2 t). \quad (2.68)$$

In the frequency domain translucency manifests as,

$$\hat{I}(\omega, \omega_t) = \hat{I}_1(\omega) \delta(\omega \cdot \mathbf{v}_1 + \omega_t) * \hat{I}_2(\omega_x) \delta(\omega \cdot \mathbf{v}_2 + \omega_t), \quad (2.69)$$

where  $*$  denotes the convolution operator. The geometric interpretation of the spectrum is dependent on the spectral content of the constituent layers. As a simple example, assume that a translating impulse (*layer 1*) is modulated by a moving sinusoid (*layer 2*), the resulting spectrum consists of the spectrum of the impulse (oriented line) but translated such that it does not contain the origin. As mentioned above, translucency motion may be converted to transparency motion if the image is pre-processed by taking the logarithm of the image sequence.

---

<sup>3</sup>The optical flow constraint (2.21) using operator notation is defined as follows,

$$D(\mathbf{v})I(x, y, t) = 0 \quad (2.65)$$

where  $D(\mathbf{v})$  is a linear operator defined as follows,

$$D(\mathbf{v}) = \begin{pmatrix} v_x \\ v_y \\ 1 \end{pmatrix} \cdot \nabla. \quad (2.66)$$

The spectral analysis of *occlusion* was first analyzed by Fleet and Langley (Fleet & Langley, 1994) and later by Beauchemin and Barron (Beauchemin & Barron, 1994; Beauchemin & Barron, 2000a; Beauchemin & Barron, 2000b) and Yu et al. (Yu et al., 1999; Yu et al., 2002; Yu et al., 2003). The spatial domain description consists of three layers, the occluder  $I_1(\mathbf{x})$  moving with velocity  $\mathbf{v}_1 = (u_1, v_1)$ , a function delineating the spatial support (window) of the occluder  $w(\mathbf{x})$  (e.g., two-dimensional Heaviside unit step function as considered in (Yu et al., 1999; Yu et al., 2002; Yu et al., 2003)) moving with velocity  $\mathbf{v}_1$  and the occluded layer  $I_2(\mathbf{x})$  moving with velocity  $\mathbf{v}_2 = (u_2, v_2)$ , combined as follows,

$$I(x, y, t) = w(\mathbf{x} - \mathbf{v}_1 t) I_1(\mathbf{x} - \mathbf{v}_1 t) + (1 - w(\mathbf{x} - \mathbf{v}_1 t)) I_1(\mathbf{x} - \mathbf{v}_2 t). \quad (2.70)$$

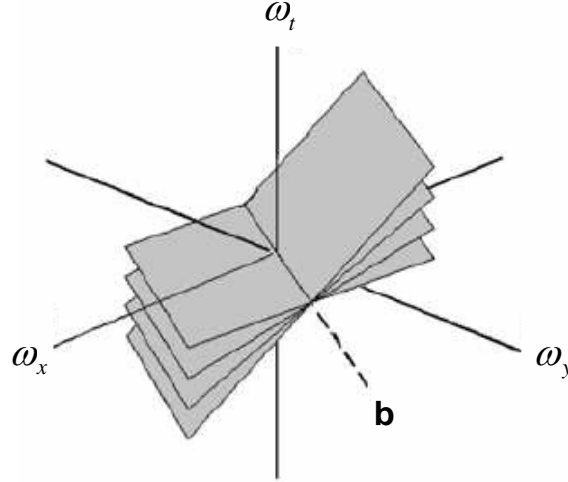
Notice that replacing the window function  $w(\mathbf{x})$  with a real constant  $\alpha \in (0, 1)$  yields the transparency motion case detailed above. The spectral domain description of occlusion is given as follows,

$$\begin{aligned} \hat{I}(\omega, \omega_t) = & \hat{w}(\omega) \delta(\omega \cdot \mathbf{v}_1 + \omega_t) * \hat{I}_1(\omega) \delta(\omega \cdot \mathbf{v}_1 + \omega_t) \\ & + \hat{I}_2(\omega) \delta(\omega \cdot \mathbf{v}_2 + \omega_t) \\ & - \hat{w}(\omega) \delta(\omega \cdot \mathbf{v}_1 + \omega_t) * \hat{I}_2(\omega) \delta(\omega \cdot \mathbf{v}_2 + \omega_t). \end{aligned} \quad (2.71)$$

The first two terms of (2.71) are the superposition of the spectral planes of the occluder and occluded layers. The occluder's spectral plane is additionally subjected to a distortion within its plane. The third term accounts for a distortion plane emanating from the occluded's spectral plane and parallel to the occluder's spectra. Furthermore, the distortion decreases rapidly in a hyperbolic fashion and is in most cases negligible or only comparable to the noise present (Yu et al., 1999; Yu et al., 2002; Yu et al., 2003). This insight led Yu et. al (Yu et al., 1999; Yu et al., 2002; Yu et al., 2003) to conclude that the spectrum of the occlusion is dominated by the two spectral planes of the constituent layers and that making the use of the distortion to assign the dominant planes to the occluder and occluded layers unreliable. Furthermore, an analysis limited to the frequency domain would not allow one to distinguish between the cases of transparency and occlusion motion. Furthermore, if the background layer has very little power or the power is concentrated at the origin, then (2.71) becomes indistinguishable from a translating pattern.

Recently, Langer and Mann (Langer & Mann, 2001; Langer & Mann, 2003; Mann & Langer, 2005) reported a new class of multiple motions, termed *optical snow*. An example where optical snow arises is when an observer moves laterally in a static three-dimensional scene containing many surfaces at different depths. They extend the *constant velocity model* by assuming that within a local image region of analysis there is a one parameter family of velocities of the form,

$$\mathbf{v} = (u_x, v_x) = (u_x + \alpha \tau_x, u_y + \alpha \tau_y), \quad (2.72)$$



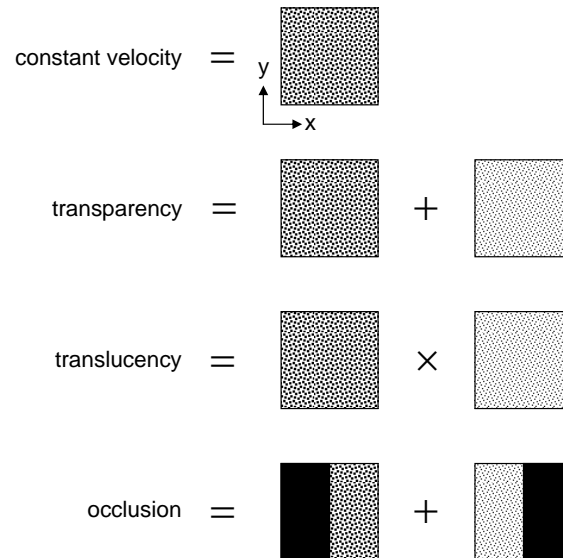
**Figure 2.11.** Optical snow. The “bowtie” spectrum of optical snow is depicted. The family of planes intersect a common line denoted by **b** through the origin.

where  $u_x, v_y, \tau_x$  and  $\tau_y$  are constants and  $\alpha$  is a free parameter. Substituting (2.72) into the constant velocity model (2.62) produces a one-parameter family of planes in the frequency domain,

$$(u_x + \alpha\tau_x)\omega_x + (u_y + \alpha\tau_y)\omega_y + \omega_t = 0 \quad (2.73)$$

where each plane intersects a common line through the origin. They liken this family of planes to a “bowtie” signature (see Fig. 2.11). It should be noted that the bowtie signature represents an abstraction from the true spectrum of optical snow since the distortions introduced by the occlusions (see discussion above) of the surfaces are not explicitly modeled.

Table 2.3 provides a summary of the spatiotemporal and frequency domain definitions of each of the multiple motion cases detailed above and Fig. 2.12 provides an illustrative depiction of the composition of each case.



**Figure 2.12.** Layer decomposition of motion. The layer decomposition is depicted for: constant velocity, transparency, translucency and occlusion. In the case of occlusion, the occluder layer is modulated by a local windowing function (e.g., two-dimensional Heaviside unit step function) moving with the same velocity of the occluder and the occluded layer is modulated by the inverse windowing function of the occluder moving with the same velocity of the occluder. The case of optical snow (not depicted) is formed by a series of layers combined sequentially by depth order with occlusion (furthest to closest from the camera).

---

Motion type	Spatial domain	Frequency domain
<i>Constant velocity</i>	$I(\mathbf{x}, t) = I(\mathbf{x} - \mathbf{v}t)$	$\hat{I}(\omega, \omega_t) = \hat{I}(\omega)\delta(\omega \cdot \mathbf{v} + \omega_t)$
<i>Transparency</i>	$I(\mathbf{x}, t) = \alpha I(\mathbf{x} - \mathbf{v}_1 t) + (1 - \alpha)I(\mathbf{x} - \mathbf{v}_2 t)$	$\hat{I}(\omega, \omega_t) = \alpha \hat{I}(\omega, \omega_t)\delta(\omega \cdot \mathbf{v}_1 + \omega_t) + (1 - \alpha)\hat{I}(\omega)\delta(\omega \cdot \mathbf{v}_2 + \omega_t)$ <p>where  <math>\alpha \in (0, 1)</math></p>
<i>Translucency</i>	$I(x, y, t) = I_1(\mathbf{x} - \mathbf{v}_1 t)I_2(\mathbf{x} - \mathbf{v}_2 t)$	$\hat{I}(\omega, \omega_t) = \hat{I}_1(\omega)\delta(\omega \cdot \mathbf{v}_1 + \omega_t) * \hat{I}_2(\omega_x)\delta(\omega \cdot \mathbf{v}_2 + \omega_t)$
<i>Occlusion</i>	$I(x, y, t) = w(\mathbf{x} - \mathbf{v}_1 t)I_1(\mathbf{x} - \mathbf{v}_1 t) + (1 - w(\mathbf{x} - \mathbf{v}_1 t))I_1(\mathbf{x} - \mathbf{v}_2 t)$	$\hat{I}(\omega, \omega_t) = \hat{w}(\omega)\delta(\omega \cdot \mathbf{v}_1 + \omega_t) * \hat{I}_1(\omega)\delta(\omega \cdot \mathbf{v}_1 + \omega_t) + \hat{I}_2(\omega)\delta(\omega \cdot \mathbf{v}_2 + \omega_t) - \hat{w}(\omega)\delta(\omega \cdot \mathbf{v}_1 + \omega_t) * \hat{I}_2(\omega)\delta(\omega \cdot \mathbf{v}_2 + \omega_t)$
<i>Optical flow</i>	N/F	$\hat{I}(\omega_x, \omega_y, \omega_t) = \sum_{i=0}^N \hat{I}(\omega)\delta(\omega \cdot \mathbf{v}_i + \omega_t)$ <p>where  <math>v_i = u + \alpha_i \tau</math> and  <math>u</math> and <math>\tau</math> are constants</p>

**Table 2.3.** Summary of spatiotemporal (X-Y-T) and spectral ( $\omega_x$ - $\omega_y$ - $\omega_t$ ) domain definitions of motion. N/F  $\equiv$  not formulated.

## Chapter 3

# Recovering optical flow

GIVEN a definition of optical flow, how does one estimate the flow from image data? The estimation relies on pooling information within a region (i.e., aperture) or at a point (by way of introducing further constraints) in order to avoid the aperture problem. Further considerations include: robustness to noise and outliers (deviations from assumptions), and computational complexity. Optical flow estimation methods can be classified into three main groups:

1. **Matching methods:** compute image displacements by matching image features or image regions across two or more images. These methods are also referred to as *correspondence*, *block-based*, *area-based* and *correlation-based* methods in the literature.
2. **Differential methods:** compute optical flow using spatiotemporal derivatives. These methods are also termed gradient-based methods.
3. **Frequency-based methods:** compute optical flow using local energy or phase information.

Though motivated differently, it will be demonstrated in Section 3.5 that these three groups of approaches are broadly analytically equivalent (Adelson & Bergen, 1985; Adelson & Bergen, 1986). In practice, the differences in their respective implementations can result in dramatic differences in performance (Barron et al., 1994a).

The organization of the rest of this chapter is as follows. In Sections 3.1-3.4 approaches for estimating the optical flow are reviewed. Sections 3.1-3.3 review matching, differential and frequency-based optical flow estimation methods, respectively. Section 3.4 reviews coarse-to-fine processing of motion within image pyramid structures. In Section 3.5 the analytical equivalence between certain versions of the matching-based, differential and frequency-based approaches is outlined. Finally, Section 3.6 provides a discussion.

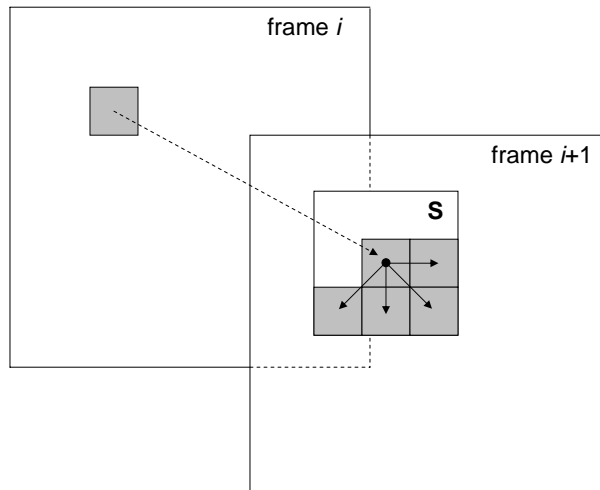
### 3.1 Matching methods

Matching methods are conceptually the simplest out of the three motion estimation methods. These methods also appear in the context of stereo vision (Brown et al., 2003). Generally, matching methods consider the problem of matching image measurements in one image across a sequence of image “snapshots” (see Chapter 2.1.1) or put another way matching methods attempt to resolve the so-called *correspondence problem* (Ullman, 1979). Matching methods are broadly classified as either feature-based or region-based approaches. An advantage of these methods over the differential- and frequency-based approaches discussed later is that they do not rely on measurements of the underlying temporal continuous signal which may be potentially poor due to significant temporal aliasing or lack of sufficient temporal support.

Feature-based approaches consist of extracting highly discriminable points in one image and seeking their match in another. The selection of features is motivated by the desire for features to exhibit a high degree of invariance under local image deformations, in particular, photometric distortions (brightness variations) and geometric distortions brought on by change in viewpoint. Examples of simple features that have appeared in the literature include lines and corners (e.g., (Moravec, 1977; Beaudet, 1978; Barnard & Thompson, 1980; Kitchen & Rosenfeld, 1982; Förstner & Güllich, 1987; Harris & Stephens, 1988)). A recent trend in the literature considers defining features based on local image descriptors constructed at points identified by an interest operator (Schmid & Mohr, 1997; Lowe, 1999; Schmid et al., 2000; Lowe, 2004; Carneiro & Jepson, 2003; Carneiro & Jepson, 2005; Mikolajczyk & Schmid, 2005). The interest operator is selected on the basis of its invariance to photometric and geometric distortions, while image descriptors are selected on the basis of discriminability as well as invariance to photometric and geometric distortions. A drawback of these approaches is that they yield highly sparse flow fields, requiring an interpolation post-processing step for the recovery of dense flows. Furthermore, defining what constitutes a “good” feature is non-trivial (Schmid et al., 2000).

Region-based methods can generally be described as follows. Given a local patch  $I_1(x, y)$  in one image frame and the displaced patch in the following frame  $I_2(x, y)$ , search for the displacement  $(u, v)$  (discrete values) that optimizes some similarity measure  $M(I_1, I_2)$  over a search region (see Fig. 3.1). The local patch must be defined such that it has adequate spatial structure to avoid the aperture problem. Note that matching can be applied to various derived local image measurements; for a discussion on potential derived (local) measurements the reader is referred to Chapter 2.1.2. The search region is usually limited to the range bounded by the expected maximum displacement. Due to the quantized nature of the search, matching methods do not yield subpixel precision. To recover subpixel precision several authors (e.g., (Anandan, 1989)) have proposed interpolation strategies on the matching surface.

The *sum of squared differences* (SSD) is a computationally simple matching cri-



**Figure 3.1.** Region-based matching. The delineated region in frame  $i$  is compared to several regions in frame  $i + 1$  within the search space  $S$ .

teria that measures the total squared difference among pixels in two regions. The implicit assumption is that the brightness at each point within the patch is conserved. There is also a normalized version that makes SSD invariant to additive and multiplicative illumination changes. *Sum of absolute differences* (SAD) measures the total absolute difference among pixels in two regions. *Cross-correlation* measures the similarity between two image regions by multiplying pixel-wise and summing the results; by viewing the regions as vectors this procedure can be seen as a sliding dot product. Zabih and Woodfill (Zabih & Woodfill, 1994) introduce methods based on non-parametric local transforms of the images, namely the *rank* and *census* transform, prior to matching. Non-parametric local transforms are local transformations that rely on the relative order of intensities within a region as opposed to the intensity values themselves. The *rank transform* of a pixel is defined as the number of pixels within the region where the intensity is less than the intensity at the centre pixel. After applying the rank transform, region-based matching is executed using the sum of absolute differences. This transform is invariant to additive and multiplicative brightness changes. In addition, Zabih and Woodfill propose a variation of the rank transform, termed the *census transform*, that preserves the spatial distribution of ranks by encoding them in a bit string for each pixel. Matching is performed by using the Hamming distance (the number of bits that differ between two bit strings) between bit strings encoded for each pixel. Table 3.1 provides a listing of common match measures in the motion and stereo literature; for a comparison of these measures and their variants see (Burt et al., 1982; Aschwanden & Guggenbuhl, 1993).

In the presence of large velocities, the matching methods described above become susceptible to false matches due to increases in the search space. Furthermore, the



time complexity grows quadratically with the maximum possible displacement of a pixel (Camus, 1995). Barnea and Silverman (Barnea & Silverman, 1972) introduced the idea of terminating (“early jump-out”) the computation of the match measure when the partial similarity result exceeds a predefined threshold to reduce computations. To avoid the quadratic increase in search area, Camus (Camus, 1995) proposed a “real-time”<sup>1</sup> algorithm that fixes the spatial search region to a small value and searches over time. Optical flow is determined by the best-matching spatial shift divided by the corresponding frame delay. For improving both computational efficiency and match quality, approaches have been proposed that consider matching within pyramid structures (e.g., *Gaussian* and *Laplacian* pyramid (Burt & Adelson, 1983)) which reduces the search range and enhances salient image structure (Anandan, 1989) (see Section 3.4 for more details).

To improve match quality, especially around occlusion boundaries, Okutomi and Kanade (Okutomi & Kanade, 1992) proposed a method to adaptively change the window size and shape. More recent work on adaptive window matching can be found in (Fusiello et al., 1997; Hirschmüller et al., 2002; Okutomi et al., 2002). Singh (Singh, 1990; Singh, 1991) proposed a Kalman filter approach to integrate velocity estimates over time for the purpose of reducing the search uncertainty (i.e., search range).

---

<sup>1</sup>Liu et al. (Liu et al., 1998) reported that Camus’ optical flow algorithm (Camus, 1995) is capable of computing the flow on  $64 \times 64$  images, with the temporal search region of 10 frames, at up to 9 frames per second on a 80MHz HyperSparc computer. Limiting the temporal search range to only 3 frames the algorithm is capable of running at below frame rate.

<i>Match Measure</i>	<i>Definition</i>
Sum of Squared Differences (SSD)	$\sum_{j=-n}^n \sum_{i=-n}^n (I_1(x+i, y+j) - I_2(x+i+u, y+v+j))^2$
Normalized SSD	$\sum_{j=-n}^n \sum_{i=-n}^n \left( \frac{I_1(x+i, y+j) - \bar{I}_1}{\sqrt{\sum_{j=-n}^n \sum_{i=-n}^n (I_1(x+i, y+j) - \bar{I}_1)^2}} - \frac{I_2(x+i+u, y+v+j) - \bar{I}_2}{\sqrt{\sum_{j=-n}^n \sum_{i=-n}^n (I_2(x+i+u, y+v+j) - \bar{I}_2)^2}} \right)^2$
Sum of Absolute Differences (SAD)	$\sum_{j=-n}^n \sum_{i=-n}^n  I_1(x+i, y+j) - I_2(x+i+u, y+v+j) $
Cross-Correlation	$\sum_{j=-n}^n \sum_{i=-n}^n I_1(x+i, y+j) \cdot I_2(x+i+u, y+v+j)$
Normalized Cross-Correlation	$\frac{\sum_{j=-n}^n \sum_{i=-n}^n (I_1(x+i, y+j) - \bar{I}_1) \cdot (I_2(x+i+u, y+v+j) - \bar{I}_2)}{\sum_{j=-n}^n \sum_{i=-n}^n (I_1(x+i, y+j) - \bar{I}_1)^2 \cdot (I_2(x+i+u, y+v+j) - \bar{I}_2)^2}$
Rank	$\sum_{j=-n}^n \sum_{i=-n}^n I'_1(x+i, y+j) - I'_2(x+i+u, y+v+j)$ <p>where <math>I'_k(x, y) = \sum_{j=-n}^n \sum_{i=-n}^n I_k(x+i, y+j) &lt; I_k(x, y)</math></p>
Census	$\sum_{j=-n}^n \sum_{i=-n}^n \text{HAMMING}(I'_1(x+i, y+j), I'_2(x+i+u, y+v+j))$ <p>where <math>I'_k(x, y) = (I_k(x+1, y+1) &lt; I_k(x, y), \dots, I_k(x+n, y+n) &lt; I_k(x, y))</math></p>

**Table 3.1.** Common region-based match measures.  $\bar{I}_1$  and  $\bar{I}_2$  denote the local average intensity.  $\text{HAMMING}(\cdot, \cdot)$  returns the number of bits that differ in the two bit strings.

## 3.2 Differential-based estimation approaches

In the following sections, differential-based approaches for the estimation of optical flow are assumed. Unlike the matching methods discussed above, the differential-based approaches assume that the image representation is locally continuous in both the spatial and temporal dimensions. In practice, computations are done on a discrete representation of the image sequence. In this case, the image sequence is assumed to be sampled rapidly enough such that measurements in the form of derivatives of the underlying continuous representation may be recovered (i.e., minimal spatial or temporal aliasing is introduced). Throughout the following discussion, the presentation will be limited to recovering flow with the optical flow constraint (2.21). However, the general formulation of these estimators are not specific to the optical flow constraint (2.21), similar estimators can be formulated with the “relaxed” differential constraints detailed in Chapter 2.2.4.

### 3.2.1 Local estimation methods: Least-squares and variants

As pointed out in Chapter 2.2.2 the full recovery of optical flow at a point using the optical flow constraint (2.21) alone is an underconstrained problem. To arrive at a full estimate, at least one more independent constraint is required. Additional constraints may be realized by assuming that the velocity is locally constant (Glazer, 1981; Lucas & Kanade, 1981a; Lucas & Kanade, 1981b), thus allowing enough constraints to be pooled to define a unique solution. More generally a parametric model of velocity may be assumed to describe the local motion (see Chapter 4.1). Note that the locally constant velocity model represents a zeroth-order parametric model. Given the optical flow constraints within a region centred about the point whose velocity estimate is desired, the estimate of the velocity can be obtained by forming a (*weighted*) *least squares* estimate that seeks the velocity that minimizes the sum of the squared deviation from brightness constancy for each point within the region of analysis, formally,

$$\arg \min_{(u,v)} \sum_{(x,y,t) \in \Omega} w(x,y,t) (I_x u + I_y v + I_t)^2, \quad (3.1)$$

where  $\Omega$  defines the region of analysis and  $w(\cdot, \cdot, \cdot)$  represents a weighting function; for example a Gaussian that is centered at the middle of the region of analysis. Differentiating (3.1) with respect to the velocity components and setting the derivatives to zero yields the closed-form solution,

$$\mathbf{v} = \underbrace{\begin{bmatrix} \sum w(x,y,t) I_x I_x & \sum w(x,y,t) I_x I_y \\ \sum w(x,y,t) I_x I_y & \sum w(x,y,t) I_y I_y \end{bmatrix}^{-1}}_{(\mathbf{A}^\top \mathbf{W} \mathbf{A})^{-1}} \underbrace{\begin{pmatrix} -\sum w(x,y,t) I_x I_t \\ -\sum w(x,y,t) I_y I_t \end{pmatrix}}_{\mathbf{A}^\top \mathbf{W} \mathbf{b}}, \quad (3.2)$$

where

$$\mathbf{A} = (\nabla I(x_1, y_1), \dots, \nabla I(x_n, y_n))^\top \quad (3.3)$$

$$\mathbf{b} = -(I_t(x_1, y_1), \dots, I_t(x_n, y_n))^\top \quad (3.4)$$

and  $\mathbf{W}$  is a diagonal matrix with  $w(x_1, y_1, t_1), \dots, w(x_n, y_n, t_n)$  along its diagonal. Even with two or more constraints it may be the case that the minimum number of independent constraints has not been met (i.e., the aperture problem) and thus no estimate of the full velocity can be obtained.

It is also instructive to view the least-squares minimization (3.1) in the Fourier domain. Applying *Parseval's theorem*<sup>2</sup> to (3.1),

$$\epsilon(u, v) = \sum_{\omega_x, \omega_y, \omega_t} |\mathcal{F}\{I_x(x, y, t)\}u + \mathcal{F}\{I_y(x, y, t)\}v + \mathcal{F}\{I_t(x, y, t)\}|^2, \quad (3.6)$$

where  $\mathcal{F}\{f(\cdot)\}$  denotes the Fourier transform of  $f(\cdot)$ . Applying the *derivative theorem*<sup>3</sup> of the Fourier transform to (3.6), yields,

$$\epsilon(u, v) = \sum_{\omega_x, \omega_y, \omega_t} (\omega_x u + \omega_y v + \omega_t)^2 |\hat{I}(\omega_x, \omega_y, \omega_t)|^2, \quad (3.7)$$

where  $\hat{I}(\omega_x, \omega_y, \omega_t)$  represents the Fourier transform of  $I(x, y, t)$ . Recalling from Chapter 2.1.1 that the spectrum of a translating pattern in the frequency domain lies on a plane passing through the origin (2.32), Eq. (3.7) algebraically expresses that the minimization problem consists of finding the parameters  $(u, v)$  of the plane  $\omega_x u + \omega_y v + \omega_t = 0$  that accounts for (“nulls out”) the energy in the Fourier spectrum.

A source of significant imprecision in the motion estimates is temporal aliasing which corrupt the temporal derivative estimates. In this case, precision can be improved by an iterative alignment procedure (Netravali & Robbins, 1979; Lucas & Kanade, 1981a; Lucas & Kanade, 1981b). This procedure can be likened to Newton-Raphson's root finding method (Korn & Korn, 1968). Given the initial velocity estimate the first image is shifted towards the second image. The motion estimation procedure is repeated between the newly shifted image and the second image. This procedure is iterated until convergence.

---

<sup>2</sup>*Parseval's theorem* states that the power of a signal computed in the spatiotemporal  $f(x, y, t)$  and the frequency  $F(\omega_x, \omega_y, \omega_t)$  domains are equal (Lim, 1990), formally,

$$\sum_{x, y, t} f(x, y, t)^2 = \sum_{\omega_x, \omega_y, \omega_t} |F(\omega_x, \omega_y, \omega_t)|^2. \quad (3.5)$$

<sup>3</sup>The *derivative theorem* states that the Fourier transform of the derivative of a function  $f(x, y, t)$  equals the Fourier transform of  $f$  multiplied by an complex ramp function  $j\omega$  (Lim, 1990). For example, the Fourier transform of  $f_x(x, y, t)$  is  $j\omega_x F(\omega_x, \omega_y, \omega_t)$  where  $F$  is the Fourier transform of  $f$ .

From a statistical point of view, the least squares solution corresponds to the *maximum likelihood* solution in the case where the temporal derivative measurements of the image intensity are contaminated by additive independently identically distributed (i.i.d.) Gaussian noise (Press et al., 1992) and the spatial derivatives are noise-free. As an aside, though the least-squares method is commonly referred to as the Lucas and Kanade algorithm (Lucas & Kanade, 1981a; Lucas & Kanade, 1981b), this method had been previously known (Netravali & Robbins, 1979; Horn & Schunck, 1993).

The least-squares solution implicitly assumes that the spatial derivatives of the image intensity are noise-free. In the estimation of motion this assumption does not hold in practice. A consequence is that the estimate from the least-squares approach can be shown to be statistically inconsistent and biased toward zero (Van Huffel & Vandewalle, 1991).

An alternative approach is to assume that all measurements are contaminated by noise, where the noise is assumed to be i.i.d.. This generalization of the least-squares solution is called *total least-squares* (Chu & Delp, 1989; Wang et al., 1992; Weber & Malik, 1995) or the *structure tensor* approach (Jähne, 1990; Jähne et al., 1998; Middendorf & Nagel, 2001) (in the statistical literature this approach also appears under the names *error-in-variables* and *orthogonal regression* (Van Huffel & Vandewalle, 1991)). In the case where the noise is i.i.d. having the same covariance, the total least-squares solution represents a *maximum likelihood estimator* (Van Huffel & Vandewalle, 1991). Note that the assumption of Gaussian sensor noise is rarely verified in practice, one appeals to the central limit theorem to justify the approximation.

The key conceptual difference between the least-squares and total least-squares approaches can be seen by looking at the one-dimensional case of fitting a line  $y = mx + b$  to a set of data points  $\{(x_i, y_i)\}$ , depicted in Fig. 3.2. The least-squares solution seeks the parameterization of the line that minimizes the vertical distance between the data points and the line, whereas in the case of total least-squares, the solution seeks to minimize the perpendicular distance.

The formulation of the total least-squares solution in the context of optical flow estimation begins with the following generalization of the optical flow constraint,

$$I_x u + I_y v + I_t w = 0 \quad (3.8)$$

where  $w$  represents an additional degree of freedom; in the case of the optical flow constraint (2.21)  $w = 1$ .

To obtain further constraints and combat the effects of noise,  $\mathbf{u} = (u, v, w)^\top$  is estimated by minimizing (3.8) locally in a least-squares sense, as follows,

$$E(\mathbf{u}) = \min_{\|\mathbf{u}\|=1} \sum_{(x,y,t) \in \Omega} g(x, y, t) (\nabla I(\mathbf{x}, t)^\top \mathbf{u})^2 \quad (3.9)$$

where  $\Omega$  defines the region of analysis, the constraint  $\|\mathbf{u}\| = 1$  is enforced in order to avoid the trivial solution  $\mathbf{u} = (0, 0, 0)^\top$  and  $g(x, y, t)$  is a local spatiotemporal weighting term, such as a Gaussian.

Expanding (3.9) and applying several matrix manipulation steps, yields,

$$E(\mathbf{u}) = \min_{\|\mathbf{u}\|=1} \mathbf{u}^\top \mathbf{S} \mathbf{u} \quad (3.10)$$

where  $\mathbf{S}$ , termed the *structure tensor* (referred to as the *gradient tensor* in physics), is given by,

$$\mathbf{S} = \sum_{\Omega} g(x, y, t) \begin{bmatrix} I_x I_x & I_x I_y & I_x I_t \\ I_y I_x & I_y I_y & I_y I_t \\ I_t I_x & I_t I_y & I_t I_t \end{bmatrix} \quad (3.11)$$

$$= \sum_{\Omega} g(x, y, t) \begin{pmatrix} I_x \\ I_y \\ I_t \end{pmatrix} (I_x \ I_y \ I_t) \quad (3.12)$$

$$= \sum_{\Omega} g(x, y, t) \nabla I (\nabla I)^\top \quad (3.13)$$

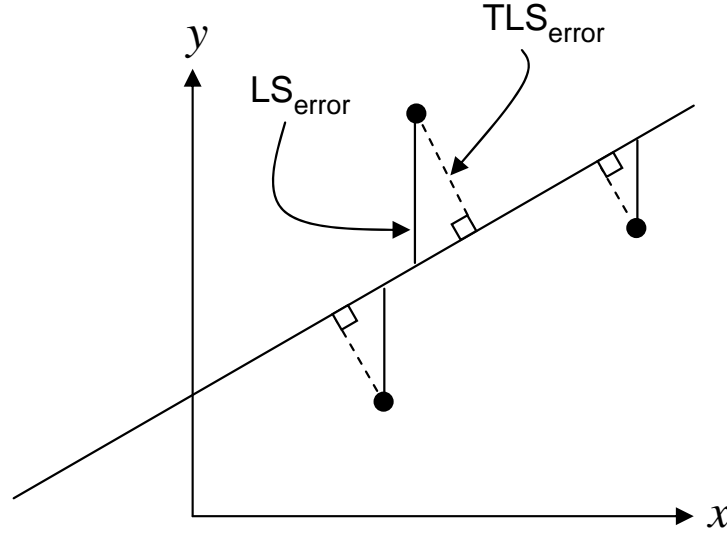
and  $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial t)^\top$ . Notice that the first two rows of the structure tensor matrix are equivalent to the data matrix components used in the least-squares approach (3.2) and the third row accommodates for the variability allowed in the temporal direction by the introduction of  $w$  in (3.8).

One possible route to minimizing (3.10) is to form a Lagrange minimization problem. It can be shown that this minimization is equivalent to the following eigenvalue problem,

$$\mathbf{S} \mathbf{u} = \lambda \mathbf{u} \quad (3.14)$$

where the sought after solution corresponds to the eigenvector with the smallest eigenvalue. If the translational model is ideally met then the rank of  $\mathbf{S}$  is two.

To summarize, the key formulation difference between the least-squares and total least-squares approaches is in their respective optimization procedures. The least-squares estimator varies the two spatial components  $u, v$  of the optical flow constraint while keeping the temporal component fixed to one. The total least-squares approach varies both the spatial and temporal components of the optical flow constraint under the constraint that the magnitude of the direction vector  $(u, v, w)^\top$  equals one. A drawback of total least-squares is that if the assumption that the noise is i.i.d. having the same variance does not hold, the estimate can actually be worse than the least-squares estimate (Weber & Malik, 1995). Theoretically, the i.i.d. assumption is invalid due to the fact that the standard methods for estimating the partial derivatives introduce correlated noise, thus violating the independence assumption (Nagel, 1995).



**Figure 3.2.** Least-squares (LS) versus total least-squares (TLS).

### 3.2.2 Confidence measures

Confidence measures associated with the velocities is also an important but often neglected topic. In many situations, where the computation of optical flow represents an input to further stages of processing, associated confidence measures can be used to weight or discard (i.e., weight of zero) velocity measurements. In the case of the least-squares solution (3.2), several authors have proposed using measures of the normal matrix  $(\mathbf{A}^\top \mathbf{W} \mathbf{A})^{-1}$ . For example, Simoncelli et al. (Simoncelli et al., 1991) suggested the minimum eigenvalue which is an indicator of how close the matrix is to being singular. Similarly, Weber and Malik (Weber & Malik, 1995) for the case of total least-squares used the condition number (Press et al., 1992) of the normal matrix. Waxman et al. (Waxman et al., 1988) proposed using the Gaussian curvature associated with the spatial intensity function. Related confidence measures can also be found in the stereo matching literature, see (Egnal et al., 2004) for an empirical comparison of measures.

### 3.2.3 Global estimation methods

The problem of computing the optical flow based on the brightness constancy constraint alone is an ill-posed problem in the sense of Hadamard (Hadamard, 1902), who defined an ill-posed problem as a problem whose solution does not exist *or* it is not unique *or* it is not stable under perturbations of the data. The least-squares approaches reviewed in Section 3.2.1 are based on the assumption that the motion

within the local region of analysis is restricted to some compact parametric form. In this section methods that explicitly include global constraints to restrict the solution space, making the problem “well-posed”, are reviewed. Horn and Schunck (Horn & Schunck, 1981) introduced a *smoothness constraint* based on the assumption that the flow at neighbouring points are “similar”. The selection of this constraint was guided by the physical consideration that the real world consists of solid objects with smooth surface whose projected velocity field is usually smooth. The smoothness term has the effect of restricting the class of admissible solutions. The measure of smoothness they proposed was the square of the optical flow gradient magnitude, given by,

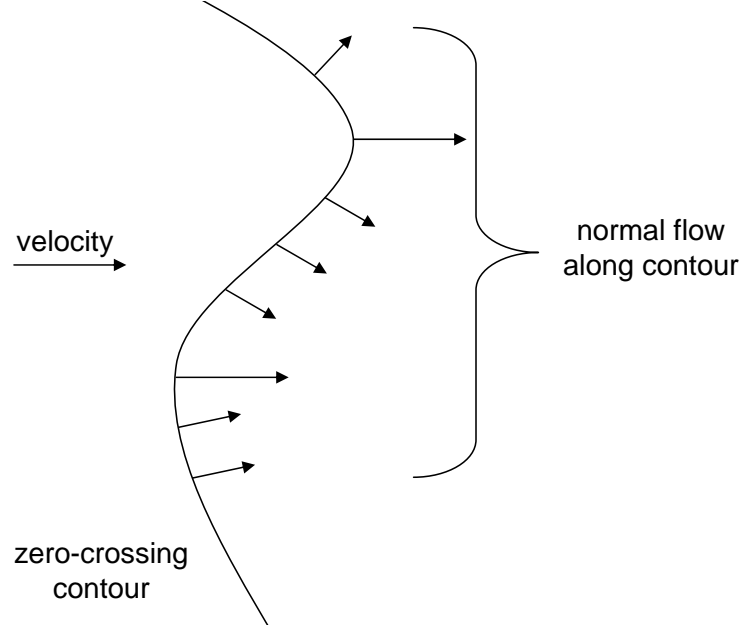
$$\|\nabla u\|_2^2 \text{ and } \|\nabla v\|_2^2. \quad (3.15)$$

Another way to look at this constraint is in terms of regularization theory (Poggio et al., 1984; Poggio et al., 1985; Bertero et al., 1988), where constraints of this form are taken to stabilize the solution (Tikhonov, 1995). Horn and Schunck (Horn & Schunck, 1981) combine the optical flow constraint (data term) (2.21) and smoothness (smoothness term) (3.15) to form the following minimization problem,

$$\min \int \int \underbrace{(\nabla I \cdot \mathbf{v} + I_t)^2}_{\text{data term}} + \alpha^2 \underbrace{(\|\nabla u\|_2^2 + \|\nabla v\|_2^2)}_{\text{smoothness term}} dx dy \quad (3.16)$$

where  $\alpha$  is a weighting parameter that controls the influence of the smoothness term. The solution to this minimization problem is a problem of the calculus of variations (Weinstock, 1974) that yields a system of two partial differential equations to be solved. These partial differential equations admit a family of solutions related by arbitrary constants. To arrive at a unique solution further (boundary) conditions must be introduced. A potential boundary condition that guarantees a unique solution consists of fixing the value of the velocity function along a simple closed curve bounding the region of interest, for example, a zero normal derivative of the velocity along the bounding contour (Horn & Schunck, 1981). The computational realization of Horn and Schunck’s approach consists of discretizing (3.16), yielding a large set of linear equations that may be solved using an iterative method (e.g., *Gauss-Seidel* (Press et al., 1992)). Unlike local methods (e.g., least-squares) that cannot yield velocity estimates in regions of constant brightness, global methods (e.g., Horn and Schunck’s approach) yield full estimates in such regions due to their ability to propagate (fill-in) information from neighbouring regions. It is important to stress that the assumption of smoothness is based on heuristic grounds as opposed to a sound physical justification. For instance, in regions where one object occludes another the assumption of smoothness in the flow does not generally hold. Furthermore, since smoothness is a vague concept, the form of the smoothness term is **not** unique. An important consideration in selecting the form of the smoothness term is that it admits a unique solution.





**Figure 3.3.** Hildreth's method (Hildreth, 1984) estimates the true velocity along the contour by minimizing the squared deviation of the estimated flow along the contour with the projected normal flow of the potential solution (velocity) subject to a smoothness assumption of the flow along the contour.

Hildreth (Hildreth, 1984), proposed a similar scheme to that of Horn and Schunck, however the analysis was limited to contours extracted from zero-crossing of the Laplacian of Gaussian (LOG) filtered image sequence. Estimates of the true velocity along the contour are found by minimizing the squared deviation of the estimated flow along the contour with the projected normal flow of the potential solution (velocity) subject to a first-order smoothness assumption of the flow along the contour (see Fig. 3.3), formally,

$$\oint_{\text{zero-crossing contour}} \alpha (\hat{\mathbf{n}}^\top \mathbf{v} - v_n)^2 + (u_s^2 + v_s^2) ds \quad (3.17)$$

where  $\hat{\mathbf{n}}$  denotes the unit normal vector perpendicular to the contour at arclength  $s$ ,  $\mathbf{v}$  the velocity at  $s$ ,  $v_n$  represents the estimated magnitude of the normal flow (2.26),  $u_s$  and  $v_s$  represent the change of the velocity components along the contour and  $\alpha$  is a weighting factor. The reason cited for limiting processing along contours is based on Marr's theory (Marr, 1982) that initial motion measurements in the human visual system are limited to those locations exhibiting significant intensity change. Gong (Gong, 1989) formulated a similar constraint to that of Hildreth with the addition of

a squared difference term that accounts for the tangential component of the velocity,

$$\oint_{\text{zero-crossing contour}} \alpha(\hat{\mathbf{n}}^\top \mathbf{v} - v_n)^2 + \beta(\hat{\mathbf{t}}^\top \mathbf{v} - v_t)^2 + (u_s^2 + v_s^2) ds \quad (3.18)$$

where  $\hat{\mathbf{n}}$  and  $\hat{\mathbf{t}}$  denote the perpendicular and tangential unit normal vectors to the contour at arclength  $s$ , respectively,  $v_n$  and  $v_t$  represent the perpendicular and tangential components of velocity, respectively, and  $\alpha$  and  $\beta$  are weighting factors. The tangential component, as shown by Gong (Gong, 1989), can be recovered in regions where the determinant of the Hessian of the spatial structure is non-zero. A drawback of these contour-based approaches is that by construction the methods yield sparse velocity estimates. Furthermore, extracted contours may cross motion boundaries resulting in incorrect motion estimates.

A prominent issue with the Horn and Schunck approach is that regions consisting of occluding boundaries in motion are blurred and dislocated. To ameliorate the smoothness across motion discontinuities, Nagel (Nagel, 1983a) and Nagel and Enkelmann (Nagel & Enkelmann, 1986) replace Horn and Schunck’s smoothness term in (3.16) with an *oriented-smoothness* term that essentially prevents smoothing across edges which are equated to points in the image that exhibit steep intensity gradients, formally,

$$\int \int (\nabla I \cdot \mathbf{v} + I_t)^2 + \alpha^2 \text{trace}(\mathbf{J}^\top \mathbf{W} \mathbf{J}) dx dy \quad (3.19)$$

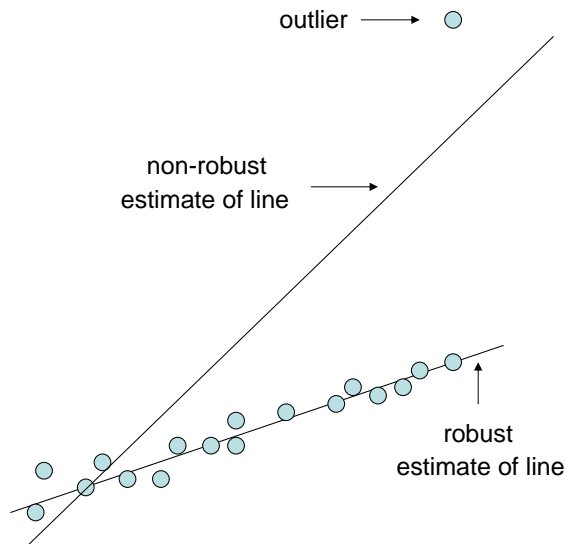
where

$$\mathbf{J} = \begin{bmatrix} u_x & v_x \\ u_y & v_y \end{bmatrix} \quad (3.20)$$

is the Jacobian matrix of velocity  $\mathbf{v}$  and  $\mathbf{W}$  is a weight matrix that encapsulates the local spatial structure through spatial derivatives. Note, if  $\mathbf{W}$  is replaced with a  $2 \times 2$  identity matrix (3.19) becomes (3.16). Interestingly, Nagel (Nagel, 1986) points out that Hildreth’s method (Hildreth, 1984) can be seen as a special case of the oriented-smoothness method. A drawback of the approach is that the assumption that edges coincide with flow discontinuities does not generally hold. For example, edges also appear in regions of strong texture within an object.

Extensions to the Horn and Schunck (Horn & Schunck, 1981) approach include reformulating Horn and Schunck’s data term (Weickert et al., 2003), smoothness term (Schnörr, 1994; Guichard & Rudin, 1996; Weickert & Schnörr, 2001a) or both terms (Black & Anandan, 1996; Roth & Black, 2005). A further extension consists of extending the smoothness constraint in (3.16) to the temporal domain (Yachida, 1981; Nagel, 1990; Weickert & Schnörr, 2001b).

A main drawback common to global methods is their computational efficiency which far exceed those of local methods. Also, the rendered solution is in general



**Figure 3.4.** Robust estimator example. Two estimates of a line are depicted, one using a non-robust formulation (e.g., least-squares) and the other robust, where one grossly outlying data point is present.

different from that of the original problem, even in the case of ideal noise-free data. Finally, there is no principled way of setting the weighting parameters.

### 3.2.4 Robust methods

In the formulation of least-squares methods (Section 3.2.1), several assumptions are made, namely, brightness constancy, and the type of motion model (e.g., translation). For global methods, the least-squares method’s assumption of a single velocity present is relaxed by allowing smooth variations in the flow. In cases where gross deviations (termed outliers) in the assumptions appear, the estimates from these methods may deviate significantly from the true flow. For example, a significant source of outliers for the problem of optical flow estimation are at motion occluding boundaries where for least-squares methods the optical flow constraint and single motion assumption do not hold, and for global methods the optical flow constraint and smoothness assumption do not hold. Figure 3.4 depicts the non-robust nature of the least-squares estimate applied to fitting a line to a set of “noisy” data points. Notice how one outlier moves the least-squares estimate far from the true value.

Estimators that are invariant to a “small number” of outliers present in the data are known collectively as *robust estimators*. Robust estimators developed within the computer vision community include, the *Hough transform* (e.g., (Hough, 1962; Cafforio & Rocca, 1976; Fennema & Thompson, 1979; Adiv, 1983; Bober & Kittler, 1994; Nesi et al., 1995)) and *constraint line clustering* (Schunck, 1984; Schunck, 1988;

Schunck, 1989). Methods adapted from the (robust) statistics literature (Huber, 1981; Hampel et al., 1986; Rousseeuw & Leroy, 2003), include, *least median squares* (e.g., (Mintz & Meer, 1991; Bab-Hadiashar & Suter, 1996)) and *M-estimators* (e.g., (Black & Anandan, 1993; Odobez & Bouthemy, 1995; Black & Anandan, 1996)).

The *Hough transform* (Hough, 1962) is a well established method for detecting parametric curves in an image. In the context of motion estimation, the Hough transform can be described by the following two steps:

1. Transform the motion estimation problem into an optical flow constraint (2.21) intersection problem. Each constraint within the region of analysis votes for the quantized velocities that satisfy it.
2. Determine the velocity with the greatest number of votes (intersections).

Major drawbacks of the Hough transform include, its time and space complexity which grows exponentially in processing time and memory space as a function of the number of parameters of the model, and the quantized nature of the the recovered parameters.

Schunck (Schunck, 1984; Schunck, 1988; Schunck, 1989) proposed the *constraint line clustering* algorithm for estimating optical flow with data contaminated by outliers, for example, regions spanning occlusion boundaries may include data consistent with multiple motions. Constraint line clustering uses a form of cluster analysis to extract an estimate. About each point in the image a set of measurements in the form of the optical flow constraint (2.21) (constraint line) is taken from a spatial neighborhood about the point. The set of intersections of each of the neighboring constraint lines is made with the centre constraint. Assuming constant velocity, all constraints that are consistent with the centre constraint tend to form tight intersection clusters around the true velocity. Any outlying constraints will not intersect the centre constraint line at a consistent point. The algorithm identifies the tightest cluster that contains approximately half of the intersections (see Fig. 3.5) to make the velocity estimate. A weakness of the algorithm is that it is susceptible to inaccuracies of the centre constraint within the analysis region.

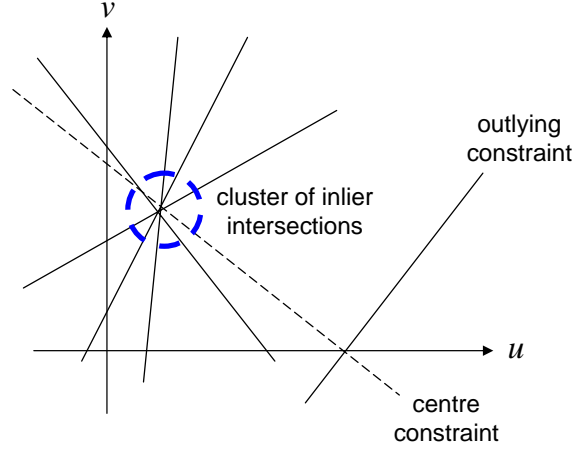
The *least-median-of-squares* (LMedS) method estimates the motion parameters by solving the following non-linear minimization problem (Bab-Hadiashar & Suter, 1996; Rousseeuw & Leroy, 2003):

$$\arg \min_{(u,v)} \text{med}_i \{ (I_x u + I_y v + I_t)_i^2 \}. \quad (3.21)$$

A drawback of LMedS is that its *efficiency*<sup>4</sup> is poor in the case where the data is contaminated by Gaussian noise (Rousseeuw & Leroy, 2003). An approach to improving

---

<sup>4</sup>The *efficiency* of a method is defined as the ratio between the lowest achievable variance for the estimated parameters and the actual variance provided by the given method (Stewart, 1999).



**Figure 3.5.** Constraint line clustering. The *constraint line clustering algorithm* identifies the tightest cluster (highlighted by dashed circle) that contains roughly half of the intersections about the centre constraint line (dashed line).

the efficiency consists of refining the motion parameters recovered by LMedS using weighted-least squares motion estimation (3.1) (Rousseeuw & Leroy, 2003).

*M-estimators* represent a generalization of the least-squares estimator (3.1) in that the quadratic error norm  $(\cdot)^2$  is replaced with a robust error norm  $\rho(\cdot)$ , formally,

$$\arg \min_{(u,v)} \sum_{s \in S} \rho(I_x u + I_y v + I_t, \sigma) \quad (3.22)$$

where  $\sigma$  is a scale parameter; (3.22) is termed an M-estimator since it corresponds to a generalization of Maximum-Likelihood estimators (Huber, 1981). The selection of  $\rho(\cdot)$  results in different levels of robustness exhibited by the estimator. One way of understanding the robustness of a particular error norm is by analyzing its *influence function* which characterizes the change in an estimate caused by the inclusion of outlying data as a function of the distance from the uncorrupted estimate. The influence function is proportional to the derivative of  $\rho(\cdot)$ , denoted  $\psi(\cdot)$ . For robustness, the influence function should tend to zero as the distance increases. In the least-squares case, the influence of data points increases linearly and is unbounded (see Fig. 3.6 (a) and (b)). In Fig. 3.6 (c)-(f) several  $\rho(\cdot)$ -functions and their corresponding  $\psi(\cdot)$ -functions are depicted. In practice, the various robust error norms introduced in the statistical literature do not provide accurate models for real computer vision applications (Meer, 2004). Thus, the choice of  $\rho(\cdot)$  in computer vision applications is based on empirical rather than theoretical grounds.

M-estimators do not generally admit a closed-form solution. To arrive at an estimate one may apply an iterative minimization technique, such as *simultaneous over-relaxation* minimization (SOR) (Press et al., 1992). Alternatively, the minimization

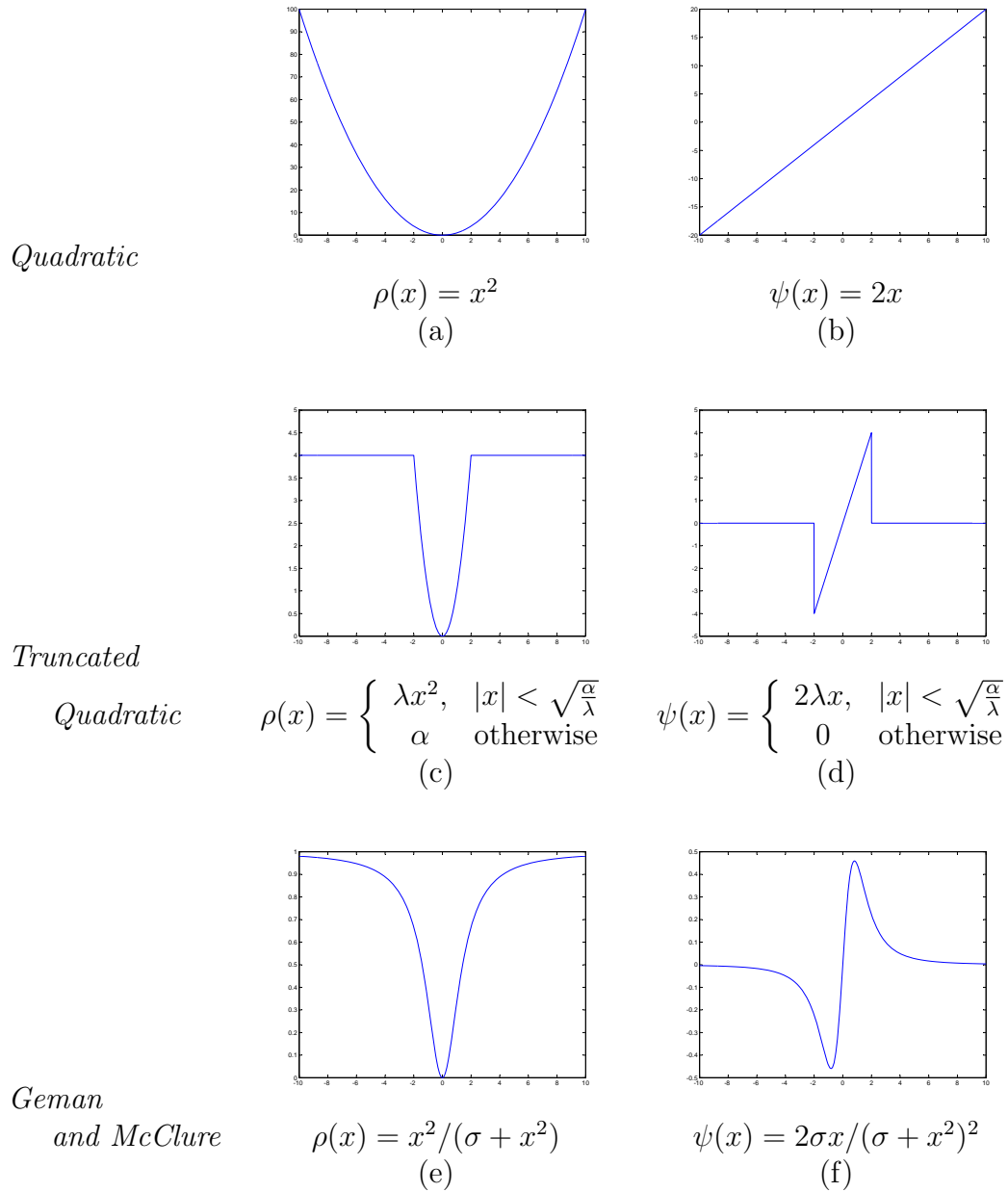
process can be formulated as an *iteratively reweighted least-squares* (IRLS) process (Stewart, 1999; Meer, 2004), which consists of alternating between:

1. calculating weights based on the current parameter estimates
2. estimating the parameters using using weighted least-squares based on the weights calculated in Step 1.

Due to the non-linear nature of the minimization problem, a good starting point is required to avoid getting trapped in local minima.

A measure of the robustness of robust estimators is their *breakdown point*, which is defined as the smallest fraction of outliers in a data set that can cause the estimator to produce arbitrarily bad results (Meer et al., 2000). The least median of squares (Bab-Hadiashar & Suter, 1996) has a breakdown point of 50%. In contrast, the least-squares estimator has a breakdown point of zero, since one arbitrarily bad point can move the estimate arbitrarily far from the true value. The breakdown point of (standard) M-estimators is zero rather than the common misconception in the vision literature (Meer et al., 2000) of  $1/(1 + p)$ , where  $p$  is the number of parameters to be estimated. The satisfactory reported performance of M-estimator-based vision algorithms highlights the fact that the breakdown point is a theoretical worst-case concept and may not be a good indicator of robustness in practice (Meer et al., 2000).

A complete treatment on the subject of robust estimators can be found in the statistics books (Huber, 1981; Hampel et al., 1986; Rousseeuw & Leroy, 2003) and the computer vision specific tutorial papers (Stewart, 1999; Meer, 2004).



**Figure 3.6.** Common error norms  $\rho(\cdot)$  - (a),(c) and (e), and their respective first derivatives  $\psi(\cdot)$  - (b),(d) and (f) (proportional to the influence function). The horizontal axis represents  $x$  and the vertical represents  $\rho(x)$ .

### 3.2.5 Probabilistic methods

In this section, probabilistic methods of motion estimation are considered. In particular, the main focus of this section are Bayesian formulations. Interestingly, from a human perception point of view a motion estimator based on a Bayesian model has been reported to predict a wide range of psychophysical results (Weiss & Adelson, 1998).

The central idea behind Bayesian approaches, as the name implies, is Bayes' theorem, formally stated as,

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \quad (3.23)$$

where,  $p(A|B)$  denotes the posterior density function,  $p(B|A)$  the likelihood,  $p(A)$  the prior, and  $p(B)$  a normalization factor (hereafter omitted). Bayes' theorem can be thought of as a means of reversing the likelihood. The goal of Bayesian approaches for the problem at hand is to calculate the posterior probability of velocity  $(u, v)$  given image data  $I$  (Weiss & Fleet, 2002),

$$p(u, v|I) \propto p(I|u, v)p(u, v). \quad (3.24)$$

A velocity estimate is often attributed to the maximum of the posterior distribution, commonly referred to as the *maximum a posteriori* or MAP for short. Advantages of casting the motion estimation problem in a Bayesian framework include, the quantification of the uncertainty of the best estimate, where uncertainty is related to the inevitable occurrence of the aperture problem and noise in the observation data, and facilitating the principled combination of information from different sources. An often cited drawback, argued by *frequentists*<sup>5</sup>, is that the priors are set subjectively.

The Bayesian approaches assume a parametric flow within the region of interest. For ease of exposition the following description of Bayesian approaches will be in terms of the constant velocity model within the region of interest. The starting point of these approaches is the specification of the likelihood. For the case of translational motion, the image data in the form of the gradient of the image sequence  $\nabla I(x, y, t)$  is related to image velocity by the optical flow constraint 2.21,

$$\nabla I(x, y, t) \cdot (u(x, y, t), v(x, y, t), 1)^\top = 0. \quad (3.25)$$

This constraint defines a plane in the gradient space  $(I_x, I_y \text{ and } I_t)$ . Inevitably, in the real-world noise enters our estimates of the data. A simple model (approximation) is that our data, specifically the temporal derivative  $I_t$ , is contaminated by additive

---

<sup>5</sup>Frequentists define probability as the long-run expected frequency of occurrence  $P(X) = n/N$ , where  $n$  is the number of times event  $X$  is observed in a total of  $N$  observations. Whereas, *Bayesians* define probability as related to degree of belief. It is a measure of the plausibility of an event given incomplete knowledge.



zero-mean Gaussian noise  $N(0, \sigma_n^2)$  while the spatial derivatives are assumed to be exact,

$$\nabla I \cdot (u, v, 1)^\top = N(0, \sigma_n^2). \quad (3.26)$$

Correspondingly, the likelihood function of observing data  $\nabla I$  given velocity  $\mathbf{v} = (u, v)$  is written as follows,

$$p(\nabla I | u, v) \propto e^{-\frac{(I_x u + I_y v + I_t)^2}{2\sigma_n^2}}. \quad (3.27)$$

This can be thought of as Gaussian deviations from the ideal plane (or mean plane) in the gradient space isolated to the  $I_t$  dimension.

To combine multiple measurements taken over a small spatial image region it is assumed that each measurement  $\nabla I^i$ , where  $i = 1, \dots, N$ , is independent and the local velocity is constant (Simoncelli, 2003), formally,

$$p(\nabla I^1, \dots, \nabla I^N | u, v) \propto e^{-\sum_{i=1}^N \frac{(I_x^i u + I_y^i v + I_t^i)^2}{2\sigma_n^2}}. \quad (3.28)$$

To complete the definition of the posterior, a prior must be selected. When the prior is assumed uniform (i.e., all velocities are equally likely), the maximum of the posterior corresponds to the least-squares solution (Lucas & Kanade, 1981b) which in turn corresponds to the maximum likelihood estimate. Simoncelli (Simoncelli, 2003) proposes a zero-mean Gaussian prior  $N(0, \sigma_v^2)$  that favours slower velocities over larger ones,

$$p(u, v) \propto e^{-\frac{(v_x^2 + v_y^2)}{2\sigma_v^2}} \quad (3.29)$$

based on the intuition that in the absence of any specific image information, for example images taken in a dark room, one should assume that things are not moving. Combining the likelihood (3.28) and prior (3.29) yields the following posterior,

$$p(u, v | \nabla I^1, \dots, \nabla I^N) \propto e^{-\sum_{i=1}^N \frac{(I_x^i u + I_y^i v + I_t^i)^2}{2\sigma_n^2} - \frac{(v_x^2 + v_y^2)}{2\sigma_v^2}}. \quad (3.30)$$

Since the posterior (3.30) is a Gaussian, the MAP solution in this case is equivalent to the mean of the posterior.

In practice, the assumption of attributing additive noise solely to the temporal derivative is unrealistic. Nestares et al. (Nestares et al., 2000) counter with the assumption that identically additive zero-mean Gaussian noise pervades all three derivative measurements. This assumption yields the following likelihood function (Nestares et al., 2000),

$$p(\nabla I^1, \dots, \nabla I^N | u, v) \propto e^{-\sum_{i=1}^N \frac{(I_x^i u + I_y^i v + I_t^i)^2}{2\sigma_n^2(1+v_x^2+v_y^2)}}. \quad (3.31)$$

The maximum likelihood estimate (i.e., prior set to uniform) in this case corresponds to the total least-squares velocity estimate (Weber & Malik, 1995).

In contrast to the above approaches that attribute the noise to the measurement process, Cremers and Yuille (Cremers & Yuille, 2003) introduce noise on the velocity itself and assume that the noise in the measurements is relatively negligible.

Probabilistic methods that attempt to capture the global velocity field, model the velocity at each image point as a random variable. The ensemble of random variables has been modeled as a Markov Random Field (MRF). In the MRF model, the value (i.e., velocity) at one location (discrete) in the image is dependent only on the values at neighbouring locations. Approaches in the literature based on MRFs include (Konrad & Dubois, 1992; Francois & Bouthemy, 1993).

### 3.3 Frequency-based approaches

In this section, frequency-based motion estimation techniques are reviewed. The reason for collectively calling these approaches frequency-based, even though the processing in the majority of the approaches discussed below is carried out in the spatial domain, is due to the fact that these approaches rely on the design of velocity-tuned filters in the Fourier domain.

The Fourier Method (Haskell, 1974; Kuglin & Hines, 1975; Arking et al., 1978; Huang & Tsai, 1981) represents the earliest instance of these algorithms. Assuming brightness constancy between two consecutive frames,  $I(x, y, t)$  and  $I(x, y, t - 1)$ , displaced with respect to each other by  $(u, v)$ , the Fourier Method leverages the linear phase shift relationship between the frames in the Fourier domain (developed in Chapter 2.2.2),

$$\hat{I}(\omega_x, \omega_y, t) = \hat{I}(\omega_x, \omega_y, t - 1)e^{-j2\pi(\omega_x u + \omega_y v)} \quad (3.32)$$

where  $\hat{I}(\omega_x, \omega_y, t)$  and  $\hat{I}(\omega_x, \omega_y, t - 1)$  denote the Fourier transforms of their respective images. To compute the displacement, the Fourier Method utilizes the shift property of the Fourier transform (Lim, 1990),

$$\hat{I}(\omega_x, \omega_y, t) = \hat{I}(\omega_x, \omega_y, t - 1)e^{-j2\pi(\omega_x u + \omega_y v)} \quad (3.33)$$

where  $\hat{I}(\omega_x, \omega_y, t)$  and  $\hat{I}(\omega_x, \omega_y, t - 1)$  denote the Fourier transforms of their respective images. To solve for  $(u, v)$ , the difference between the phase angles (denoted by  $\angle$ ) of  $\hat{I}_t$  and  $\hat{I}_{t-1}$ , given by,

$$\angle \hat{I}(\omega_x, \omega_y, t) - \angle \hat{I}(\omega_x, \omega_y, t - 1) = -2\pi(\omega_x u + \omega_y v) \quad (3.34)$$

is taken at two frequency pairs. This yields two linear equation in the unknowns  $(u, v)$  which are then solved. To lessen the spectral effects introduced by the “hard”

image borders, the images should be windowed by a smoothly tapering function, such as a Gaussian window. This approach can be applied to local regions of the image by limiting the region analysis via local windowing. The Fourier method has been generalized to handle in-plane rotational motion (de Castro & Morandi, 1987), polynomial motion (e.g., acceleration) (Chen et al., 1996) and periodic motion (e.g., circular) (Chen et al., 1996).

Instead of limiting the analysis to only two frames, as above, the energy and phase-based approaches treat the image sequence as a spatiotemporal (continuous) volume, where as discussed in Chapter 2.2.2 a translating scene obeying the brightness constancy assumption manifests itself in the Fourier domain as a planar spectrum (or line when considering a single spatial dimension) through the origin with its orientation as a function of velocity, formally,

$$\hat{I}(\omega_x, \omega_y, \omega_t) = \hat{I}_0(\omega_x, \omega_y) \delta(\omega_x u + \omega_y v + \omega_t) \quad (3.35)$$

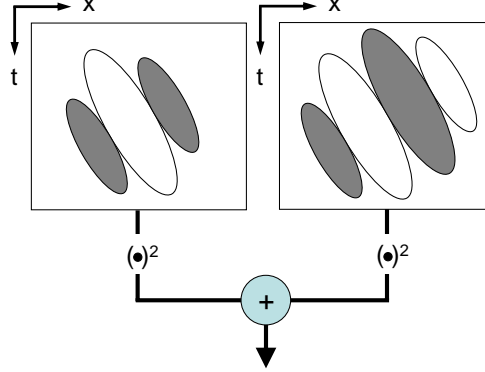
where  $\delta(\cdot)$  represents the Dirac delta function and  $\hat{I}_0(\omega_x, \omega_y)$  represents the spatial Fourier transform of the first image in the sequence  $I(x, y, 0)$ . Given constraint (3.35), both the energy and phase-based approaches are based on the use of velocity-tuned filters.

Adelson and Bergen (Adelson & Bergen, 1985; Adelson & Bergen, 1986) propose a filtering approach to identify the local spatiotemporal orientation, or equivalently the orientation of the linear energy spectrum (see Chapter 2.2.2 for details); their presentation is limited to motion in one spatial dimension, Heeger (Heeger & Pentland, 1986; Heeger, 1987; Heeger, 1988) and later Simoncelli (Simoncelli, 1993b) extended the analysis to motion in two spatial dimensions. The output from a single oriented linear filter (e.g., Gabor filter, derivative of Gaussian) is not only a function of velocity, but also of the local phase and contrast (Adelson & Bergen, 1986). To remove the phase information, the authors take the sum of the squared outputs of a pair of quadrature filters (filters out of phase by  $90^\circ$ ); this process is illustrated in Fig. 3.7. This yields a measure of the local energy (cf. (Granlund, 1978; Knutsson & Granlund, 1983; Malik & Perona, 1990; Freeman & Adelson, 1991)). To arrive at the final velocity estimate that is invariant to contrast the authors propose a ratio that takes the output of filters selective for rightward and leftward motion in an opponent fashion (subtraction) over the output of a filter selective for zero velocity (static), formally,

$$v = \frac{R - L}{S}, \quad (3.36)$$

where  $R$  represents rightward energy,  $L$  leftward energy and  $S$  static energy. The authors liken the approach to that of colour vision, where overlapping cone responses are combined to give a measure of colour that is invariant to brightness.

Heeger (Heeger & Pentland, 1986; Heeger, 1987; Heeger, 1988) proposes an algorithm based on the local energy recovered from a set of oriented filters. Heeger



**Figure 3.7.** Motion energy. The energy image is formed by: (1) filtering the input image by a pair of oriented,  $90^\circ$  out of phase filters (quadrature pairs) (2) individually squaring the outputs and (3) summing.

derives the expected response for a translating white noise image using a set of 12 Gabor energy filters tuned to various spatiotemporal orientations that lie along a cylinder in the frequency domain. To estimate the motion  $(u, v)$ , Heeger formulates a non-linear least-squares estimate that seeks the minimum difference between the measured motion energy and the predicted energy. As with any practical non-linear optimization, recovering the global minimum is not guaranteed.

In contrast to the above energy-based approaches that discard the phase information, Fleet and Jepson (Fleet & Jepson, 1989; Fleet & Jepson, 1993) recover the estimate of velocity using only the local phase component. The motivation for using the phase component is their claim that the phase is more stable than the amplitude component when the inevitable small deviations from purely translational motion occur (Fleet & Jepson, 1993). The initial step consists of extracting the local phase by convolving the image sequence by a set of complex bandpass filter (e.g., Gabor filter,  $n$ th derivative of a Gaussian) tuned to a narrow range of orientation, speed and scale. The filter response is given by:

$$R(x, y, t) = \rho(x, y, t)e^{i\phi(x, y, t)} \quad (3.37)$$

where  $\rho(x, y, t)$  denotes a measure of the local amplitude and  $\phi(x, y, t)$  the local phase. Assuming local constancy of the phase,

$$\frac{d\phi(x, y, t)}{dt} = 0, \quad (3.38)$$

yields the phase-based analog to the optical flow constraint (2.21),

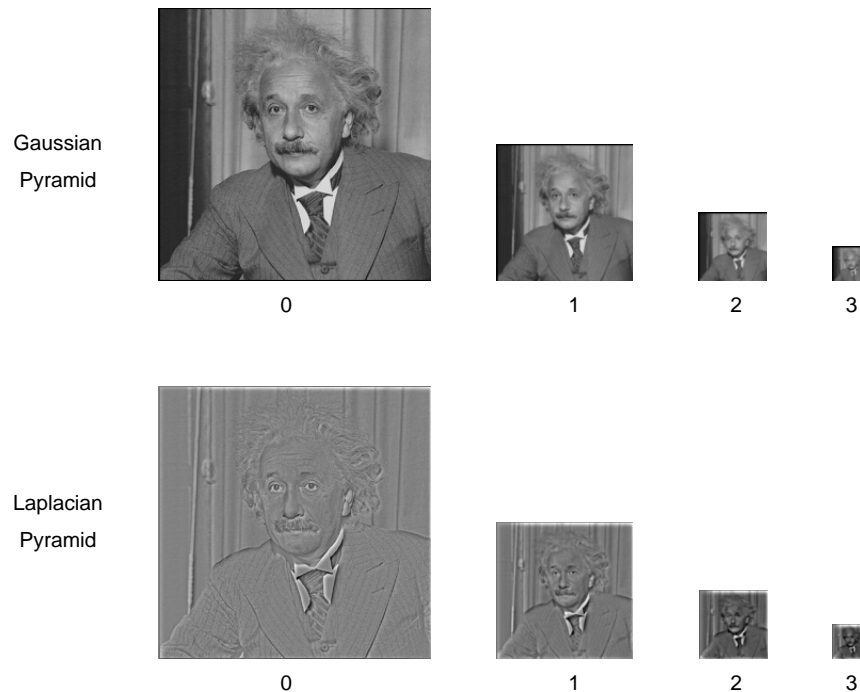
$$\phi_x(x, y, t)u + \phi_y(x, y, t)v + \phi_t(x, y, t) = 0. \quad (3.39)$$

Like the optical flow constraint, the aperture problem is still present when considering the output of a single filter at a point. Possible solutions include pooling, in a (total) least-squares sense, the phase outputs from multiple orientations and/or the phase within a local patch. This is akin to the approach of Weber and Malik (Weber & Malik, 1995), with the exception that Fleet and Jepson consider only the local phase signal, whereas Weber and Malik consider the joint local amplitude and phase signal. Advantages of the phase-based approach claimed by the authors is that it is relatively insensitive to variations in illumination, contrast and perspective projection deformations. The main disadvantage of the approach is its high computational cost due to the bandpass filtering preprocessing step.

### 3.4 Coarse-to-fine processing

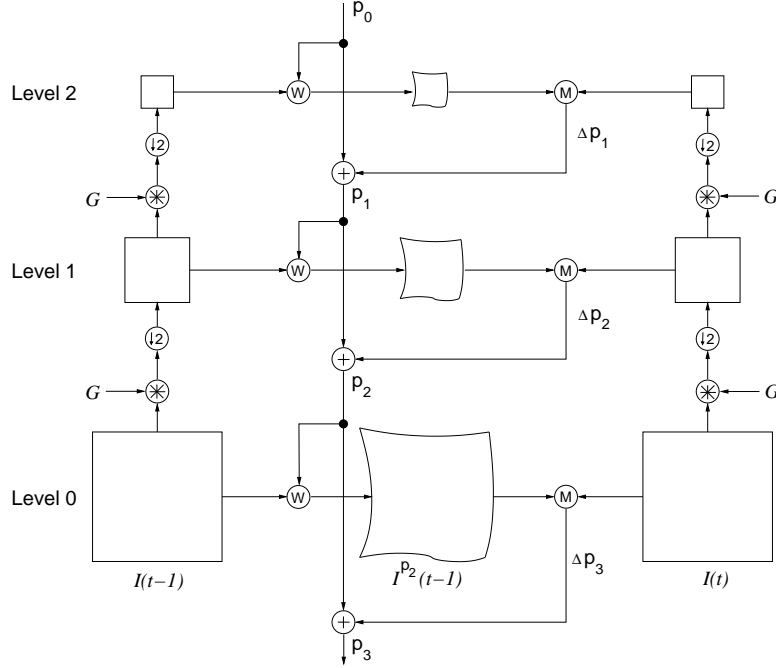
A key element shared by numerous motion estimation algorithms (e.g., (Wong & Hall, 1977; Glazer et al., 1983; Burt et al., 1983; Rosenfeld, 1984; Glazer, 1987; Anandan, 1989; Bergen et al., 1992; Spetsakis, 1997)) is coarse-to-fine processing within an image pyramid data structure such as a Gaussian (Burt, 1981) or Laplacian pyramid (Burt & Adelson, 1983; Crowley & Stern, 1984) (cf. multigrid approaches, e.g., (Glazer, 1984; Terzopoulos, 1986; Enkelmann, 1988; Memin & Perez, 1998)). Image pyramids consist of multiresolution (i.e., multiple frequency passbands) and multiple pixel density (i.e., downsampled) representations of the original image. The Gaussian pyramid represents the input image in terms of successive low-pass versions, whereas the Laplacian pyramid consists of bandpass copies of the input image (see Fig. 3.8 for an example Gaussian and Laplacian pyramid). The basic steps of motion estimation between two images, denoted  $I(t-1)$  and  $I(t)$ , within a pyramid structure are as follows (see Fig. 3.9 for a diagrammatic summary of the basic steps of pyramid processing):

1. Start at the top most level (i.e., lowest resolution) of the pyramid and find the motion estimate.
2. Propagate the motion estimate down to the next level of the pyramid.
3. (a) *Using gradient-based estimator:* Warp (i.e., transform)  $I(t-1)$  towards  $I(t)$  using the previous motion estimate.  
 (b) *Using region-based estimator:* Update search region.
4. Estimate the residual motion as in Step 1 and combine the previous estimate and the estimate of the residual motion.
5. Iterate between Steps 2-4 until the bottom of the pyramid is reached. The estimate at the bottom of the pyramid is taken as the motion estimate.



**Figure 3.8.** Gaussian and Laplacian pyramid examples. The first four levels of the Gaussian (top) and Laplacian (bottom) pyramids for the “Einstein” image are depicted.

Common justifications for embedding the estimation within a coarse-to-fine pyramid scheme are that it allows for the estimation of significant image displacements by reducing the number of local minima that the estimate can be potentially trapped in and computationally efficiency. Sources of these local minima, include, temporal aliasing caused by a displacement greater than half the period of the highest frequency component and noise. Additionally, when gradient-based motion estimators are utilized within a pyramid framework on regions containing multiple motions, they tend to “lock-onto” a single motion as the estimator progresses down the pyramid and report one of the motions present (Burt, 1991; Burt et al., 1991). This property was utilized in (Burt et al., 1989; Bergen et al., 1990; Bergen et al., 1991; Burt, 1991; Burt et al., 1991; Irani et al., 1994). A drawback of the coarse-to-fine pyramid scheme is that it often produces incorrect motion estimates when large estimation errors in coarser scales cannot be corrected at finer scales. This may happen when regions of low texture become flat at the coarsest level of resolution. To address the problem propagating incorrect estimates from coarser scales, Simoncelli (Simoncelli, 1993a) includes the knowledge of the uncertainty of coarse scale estimates. This is done using using a standard Kalman filter where the time variable is replaced by scale.



**Figure 3.9.** Hierarchical motion estimation. Given two temporally ordered images, the first step consists of constructing Gaussian pyramids (Burt, 1981) (or Laplacian (Burt & Adelson, 1983) pyramids formed by differences between successive Gaussian pyramid levels) of each image to level 2, denoted by  $I(t-1)$  and  $I(t)$  respectively. This is accomplished by a series of convolutions (asterisk symbol) by the kernel  $G$  and downsampling by 2 (represented by the symbol with an arrow point down and the number 2). Starting from level 2, each level  $i \in \{0, 1, 2\}$  of pyramid  $I(t-1)$  is warped (represented by symbol  $W$ ) by the previous estimate of the motion model's parameters (warp)  $p_{i-1}$  plus the residual motion estimate  $\Delta p_i$ . This is followed by the estimation (represented by  $M$ ) of the residual motion  $\Delta p_{i+1}$  between the warped image and level  $i$  of  $I(t)$ . This diagram is adapted from (Bergen et al., 1992).

### 3.5 Equivalences

Though motivated differently, the three main groups of optical flow methods, namely, matching, differential and frequency-based methods are broadly equivalent. In this section the analytic equivalence between certain versions of the matching (Reichardt model (Reichardt, 1961) and sum of square differences), differential (Lucas and Kanade algorithm (Lucas & Kanade, 1981a; Lucas & Kanade, 1981b)) and spatio-temporal energy (Adelson and Bergen model (Adelson & Bergen, 1985; Adelson & Bergen, 1986)) approaches are outlined. Interestingly, these equivalences between the three models have posed problems in declaring from psychophysical experiments any one of these models as the model for the human motion sensor (Derrington et al., 2004). However, physiological experimental evidence on motion-selective neurons in the cat striate cortex exists that favours the motion energy model (Emerson et al., 1992).

First, we examine the equivalence between the differential approach (i.e., Lu-

cas and Kanade algorithm) and the spatio-temporal energy approach<sup>6</sup> (Adelson and Bergen model; see Section 3.3). For ease of exposition we limit the discussion to the one-dimensional version of the algorithms; Heeger and Simoncelli (Heeger & Simoncelli, 1992; Simoncelli, 1993b) demonstrates how the two-dimensional differential approach can be interpreted as a spatio-temporal energy approach, in that the differential solution is computed from the opponent combinations of squared oriented filter responses.

The one-dimensional velocity estimate  $v$  using the Lucas and Kanade (Lucas & Kanade, 1981a; Lucas & Kanade, 1981b) algorithm is given as follows,

$$v = \frac{\sum_x w(x) I_x I_t}{\sum_x w(x) I_x^2} \quad (3.40)$$

where  $w(\cdot)$  is a weighting function (e.g., Gaussian) and  $I_x, I_t$  represent the spatial and temporal derivatives of brightness function  $I$ , respectively. A summary of the implementation of the Lucas and Kanade algorithm used in the following sketch of equivalence is given in Algorithm 1 on page 64.

Rather than computing the blurred spatio-temporal derivatives in the Lucas and Kanade algorithm (Algorithm 1 - Step 2) in two steps the operations can be combined into single filters,

$$K_x(x, t) = \frac{\partial G(x, t)}{\partial x} = \frac{\partial}{\partial x} * G(x, t) \quad (3.49)$$

$$K_t(x, t) = \frac{\partial G(x, t)}{\partial t} = \frac{\partial}{\partial t} * G(x, t). \quad (3.50)$$

Thus the numerator calculation in Algorithm 1 - Step 3 can be rewritten as

$$I_x I_t = (I * K_x)(I * K_t) \quad (3.51)$$

which may be rewritten as follows,

$$I_x I_t = \frac{(I * K_t + I * K_x)^2 - (I * K_t - I * K_x)^2}{4}. \quad (3.52)$$

Further manipulation yields,

$$I_x I_t = c \left( \left[ I * \left( \frac{K_t + K_x}{\sqrt{2}} \right) \right]^2 - \left[ I * \left( \frac{K_t - K_x}{\sqrt{2}} \right) \right]^2 \right) \quad (3.53)$$

where  $c = 1/2$ . The manipulation yields two new filters,

$$K_R = \frac{K_t - K_x}{\sqrt{2}} = \frac{1}{\sqrt{2}} \left( \frac{\partial G(x, t)}{\partial t} - \frac{\partial G(x, t)}{\partial x} \right) \quad (3.54)$$

$$K_L = \frac{K_t + K_x}{\sqrt{2}} = \frac{1}{\sqrt{2}} \left( \frac{\partial G(x, t)}{\partial t} + \frac{\partial G(x, t)}{\partial x} \right). \quad (3.55)$$

---

<sup>6</sup>This derivation is adapted from (Adelson & Bergen, 1986).



**Algorithm 1** Lucas and Kanade algorithm (one-dimensional version)

- 1: Convolve  $I(x, t)$  with a spatio-temporal Gaussian  $G(x, t)$  to remove the high spatiotemporal frequencies,

$$I_G(x, t) = I(x, t) * G(x, t) \quad (3.41)$$

where  $*$  denotes convolution.

- 2: Compute the spatial derivative  $I_x(x, t)$  and temporal derivative  $I_t(x, t)$ ,

$$I_x(x, t) = \frac{\partial I_G(x, t)}{\partial x} = \frac{\partial (I(x, t) * G(x, t))}{\partial x} = \frac{\partial}{\partial x} * (I(x, t) * G(x, t)) \quad (3.42)$$

$$I_t(x, t) = \frac{\partial I_G(x, t)}{\partial t} = \frac{\partial (I(x, t) * G(x, t))}{\partial t} = \frac{\partial}{\partial t} * (I(x, t) * G(x, t)) \quad (3.43)$$

- 3: Compute the numerator of (3.40) (denoted  $I_n(x, t)$ ) by multiplying the local spatial and temporal derivatives,

$$I_n(x, t) = I_x(x, t) I_t(x, t) \quad (3.44)$$

- 4: Compute denominator of (3.40) (denoted  $I_d(x, t)$ ) by squaring the local spatial derivative,

$$I_d(x, t) = I_x(x, t)^2 \quad (3.45)$$

- 5: Compute the sums of the numerator and denominator by convolving (3.44) and (3.45) by a spatial window function  $w(x)$  (e.g., a rectangular function with unit response or as in (Adelson & Bergen, 1986) a Gaussian),

$$I'_n = I_n(x, t) * w(x) \quad (3.46)$$

$$I'_d = I_d(x, t) * w(x) \quad (3.47)$$

- 6: Compute the velocity estimate  $v_{est}$  as a ratio of the weighted ratios,

$$v = \frac{\sum_x w(x) I_x I_t}{\sum_x w(x) I_x^2} = \frac{I'_n}{I'_d} \quad (3.48)$$

Thus simplifying (3.53) as follows,

$$I_x I_t = c[(I * K_L)^2 - (I * K_R)^2]. \quad (3.56)$$

Inspection of the filters  $K_R$  and  $K_L$  reveals that they correspond to oriented derivative of Gaussians matched to rightward and leftward movement, respectively. Furthermore, their combination corresponds to an opponent mechanism. Thus the numerator corresponds to a local weighted sum over the opponent channel. Likewise, the denominator corresponds to a local weighted sum over the static energy channel.

Putting everything together demonstrates the mathematical equivalence<sup>7</sup> between

<sup>7</sup>Notice that the quadrature counterparts of  $R$ ,  $L$  and  $S$  do not appear. The phase is eliminated through summation over the window  $\sum_x w(x)$ .

the gradient and energy methods for velocity estimation, formally,

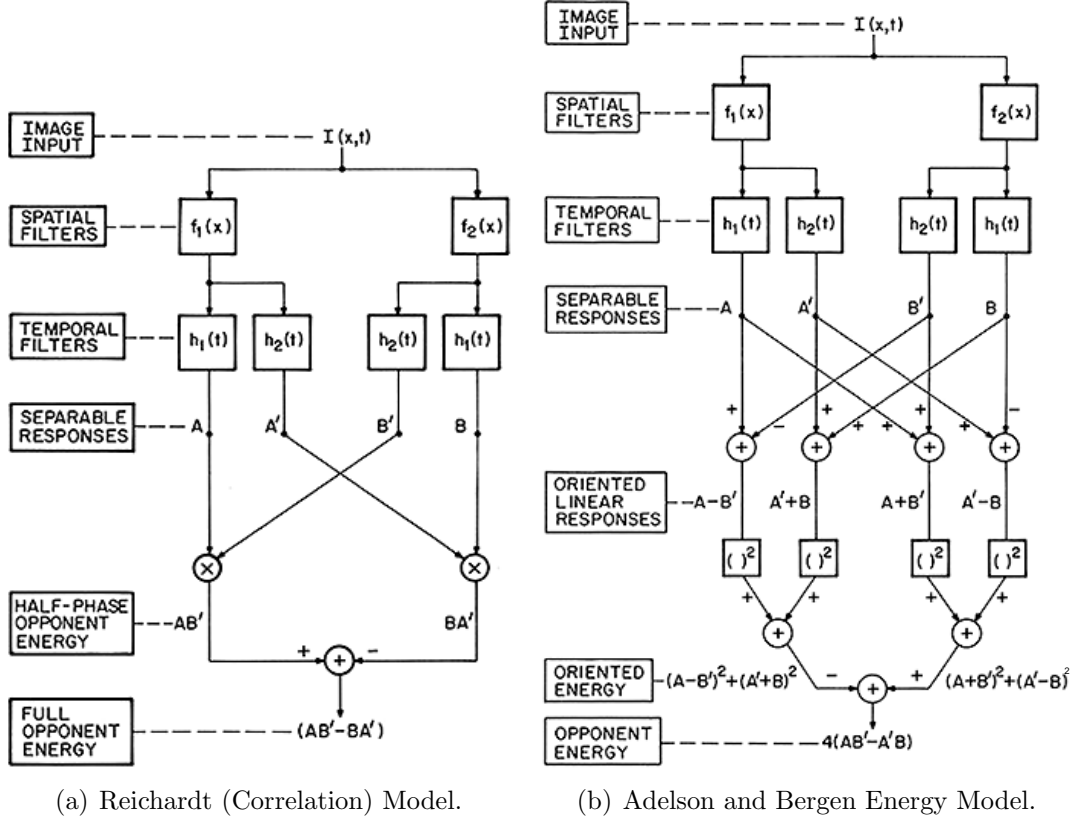
$$\frac{\sum_x w(x)(I * K_t)(I * K_x)}{\sum_x w(x)(I * K_x)^2} = \frac{c \sum_x w(x)(R - L)}{\sum_x w(x)S}. \quad (3.57)$$

Next we turn our attention to the equivalence between an instance of a matching approach in the form of a modified Reichardt model, and an energy-based approach, specifically Adelson and Bergen's energy model (Adelson & Bergen, 1985); the presentation of this equivalence is adapted from (Adelson & Bergen, 1985). The Reichardt model is a classic matching (correlation) approach proposed to model the motion sensor of flies (Reichardt, 1961) and later extended to model the human motion sensor (e.g., (van Santen & Sperling, 1984)). A single Reichardt detector is tuned to speed and therefore one would need many detectors tuned to different speeds to encode the true speed of the pattern. Figure 3.10 (a) depicts a modified version of van Santen and Sperling's Reichardt-type model. An input signal  $I(x, t)$  is fed into two spatial filter units  $f_1(x)$  and  $f_2(x)$  that represent receptive fields that are displaced in terms of position (or equivalently in phase in cases where the displacement of the spatial filter units are less than the largest period). The outputs are fed into two temporal filters  $h_1(t)$  and  $h_2(t)$  where  $h_2(t)$  includes a time delay. Pairs of the output of these units denoted by  $A, A', B'$  and  $B$  are then combined multiplicatively, yielding the outputs  $AB'$  and  $BA'$ . The final step consists of taking the difference of the outputs yielding  $AB' - BA'$ .

Now we consider Adelson and Bergen's energy model depicted in Fig. 3.10 (b). As in the initial stages of the Reichardt model (Fig. 3.10 (a)) the input signal passes through the spatial and temporal filter units with their outputs denoted by  $A, A', B'$  and  $B$ . Sums and differences of the output units are combined yielding spatiotemporally oriented responses selective for rightward and leftward motion as well as their respective quadrature outputs (differing in phase by  $90^\circ$ ). Response quadrature pairs selective for rightward and leftward are combined by summing the squared responses of their respective pairings which yields a measure of the oriented energy (i.e., the phase has been removed). The final stage consists of combining the oriented rightward and leftward energy responses in opponent fashion leading to the final output of  $4(AB' - A'B)$ . Importantly, though the formulations are motivated differently, the outputs of the two models turn out to be identical up to a multiplicative factor.

Alternatively, one can establish a relationship between the differential and matching techniques through the differential-based least-squares method and the matching-based sum of square differences (SSD). Beginning with the SSD formulation,

$$\epsilon(u, v) = \sum_{x, y} [I(x, y, t) - I(x + u, y + v, t + 1)]^2 \quad (3.58)$$



**Figure 3.10.** Equivalence between correlation and energy model. The visual input  $I(x, t)$  is convolved in parallel by two spatial bandpass functions  $f_1(x)$  and  $f_2(x)$  that differ in phase. Each output is then convolved by temporal functions  $h_1(t)$  and  $h_2(t)$  that differ in phase. The operations  $+$ ,  $-$ ,  $(\cdot)^2$  are taken pointwise. Reprinted from (Adelson & Bergen, 1985) with permission from the *Journal of the Optical Society of America*.

an approximation of (3.58) is its first-order Taylor series,

$$\epsilon(u, v) \approx \sum_{x, y} (I(x, y, t) - (I(x, y, t) + uI_x(x, y, t) + vI_y(x, y, t) + I_t(x, y, t)))^2 \quad (3.59)$$

$$= \sum_{x, y} (uI_x(x, y, t) + vI_y(x, y, t) + I_t(x, y, t))^2 \quad (3.60)$$

which yields the least-squares formulation. Therefore, both the SSD and least-squares methods minimize approximately the same error but differ in the way the optimization problem is solved.

By transitivity, the energy, matching and differential-based methods are broadly equivalent (at least among certain formulations of these approaches).

### 3.6 Discussion

This chapter reviewed literature directed at the estimation of optical flow. Though many of the approaches can be shown to be similar if not equivalent, a key advantage of having diverse formulations is that they may provide unique insights that may not be evident in a particular formulation. In practice, the differences in their respective implementations can result in dramatic differences in performance as measured by accuracy (Barron et al., 1994a), density of estimates (Barron et al., 1994a) and computational efficiency (Liu et al., 1998).

The quantitative comparison of flow algorithms is an important issue that has been given limited attention due to the difficulty in obtaining ground truth from real image sequences. Most comparisons rely on synthetic data sets where ground truth is available by construction. The *Yosemite* sequence has emerged as the de facto standard synthetic test sequence for the quantitative comparison of optical flow algorithms (see Fig. 3.11 for an example frame and its corresponding motion field). The sequence was generated by a fly through of the Yosemite valley texture mapped onto a depth map of the valley. In the upper right the flow is mainly divergent, the clouds translate to the right at 1 pixel/frame, while the velocities in the lower left are about 4 pixels/frame. The challenging aspects of this sequence include: the range of velocities, the violation of brightness constancy in the clouds, and the severe spatial aliasing in the lower portion of the images. Table 3.2 summarizes the quantitative performance of several motion estimators based on the Yosemite sequence as test data. The error between the ground truth velocity  $\mathbf{v}_g = (u_g, v_g, 1)^\top$  and the estimate  $\mathbf{v}_e = (u_e, v_e, 1)^\top$  is given by the angular error (Fleet & Jepson, 1990; Fleet, 1992),

$$\epsilon = \arccos\left(\frac{\mathbf{v}_g^\top \mathbf{v}_e}{\|\mathbf{v}_g\| \|\mathbf{v}_e\|}\right). \quad (3.61)$$

This error metric simultaneously measures error in the direction and magnitude. A problem with this measure (or a feature, if you prefer) is that symmetric deviations from the true value result in different angular errors (Otte & Nagel, 1994). Also, when comparing large velocities the differences correspond to relatively small angular errors (Otte & Nagel, 1994). Density is associated with the percentage of estimates that are deemed reliable. For local estimation algorithms, corresponding confidence measures are available for each flow estimate; if the confidence is below a certain threshold, the flow estimate is discarded. There are two main observations that can be made: (1) the average error has generally decreased (see Fig. 3.12) and (2) the global formulation of Bruhn et al. (Bruhn et al., 2005) achieves the least average angular error. The trade-off for accuracy in Bruhn et al.'s formulation is computational speed, requiring on the order of seconds to compute the flow in a single greyscale image. Thus, Bruhn et al.'s formulation, and more generally global methods, are precluded from usage in current real-time applications. The comparative results should be taken with a degree

of skepticism since most reported results for this sequence rely on tuning parameters so that one obtains the best results on a particular frame pair (Roth & Black, 2005).

Given the difficulty with obtaining ground truth for real-world images, Lin and Barron (Lin & Barron, 1994) examine the suitability of using reconstructed real-images from optical flow estimates as an error metric. Beginning with an image and its optical flow, the next image in the sequence is generated by forward (or backward) image reconstruction. The error metric proposed is the root mean square (RMS) error between the actual images and their reconstructed companions. The authors report that the RMS reconstruction error is well correlated to the angular error 3.61.

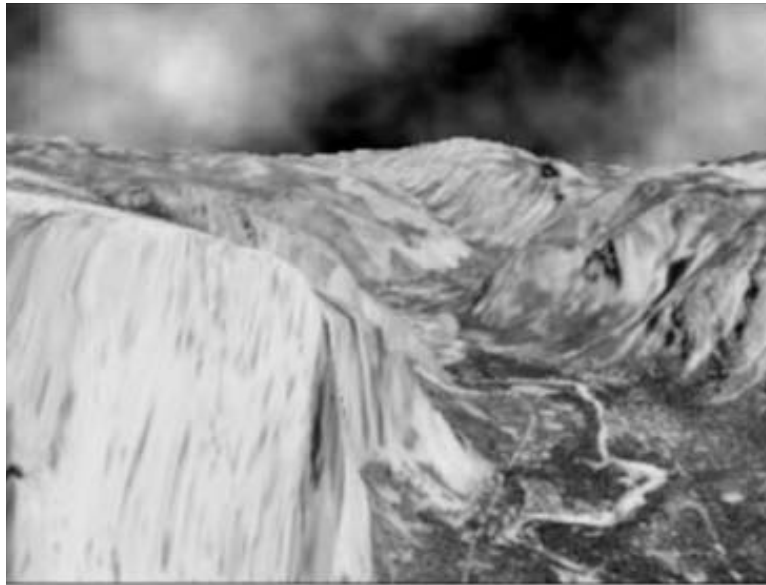
Kearney et al. (Kearney et al., 1987) presented an error analysis of local gradient-based approaches. They concluded that large errors are made in highly textured regions with significant flow; where highly textured regions are defined as regions where the second-order spatial derivative is significant. This has led many to falsely conclude (e.g., (Heeger, 1987)) that local gradient-based approaches are unreliable. The confusion is resolved by realizing that the source of Kearney et al.’s estimation error is directly related to their poor choice of differential operator in their analysis (Jähne, 2005), namely forward differences. Accurate optical flow estimates require care in the estimation of the derivatives; the design of optimized derivative filters for optical flow estimation and more generally multi-dimensional data represents an active area of research (see (Simoncelli, 1994; Scharr, 2005) for example contributions).

For many applications it has been traditionally assumed that the full optical flow estimate should represent an input for further stages of processing. This begs the question, does this really have to be the case? For the problem of *structure from motion*, the recovered optical flow is ill-suited due the extreme sensitivity of the solution to the inevitable imperfections in the flow (e.g., (Daniilidis & Spetsakis, 1997)). Fermüller (Fermüller, 1993) demonstrates that forgoing the full optical flow estimate and instead using partial information in the form of the normal flow in a qualitative manner can yield a reliable estimate of the three-dimensional motion relative to the scene. Others, such as Horn and Weldon (Horn & Weldon, 1988), have proposed direct methods that forgo flow estimates altogether by relating scene parameters directly to image structure.

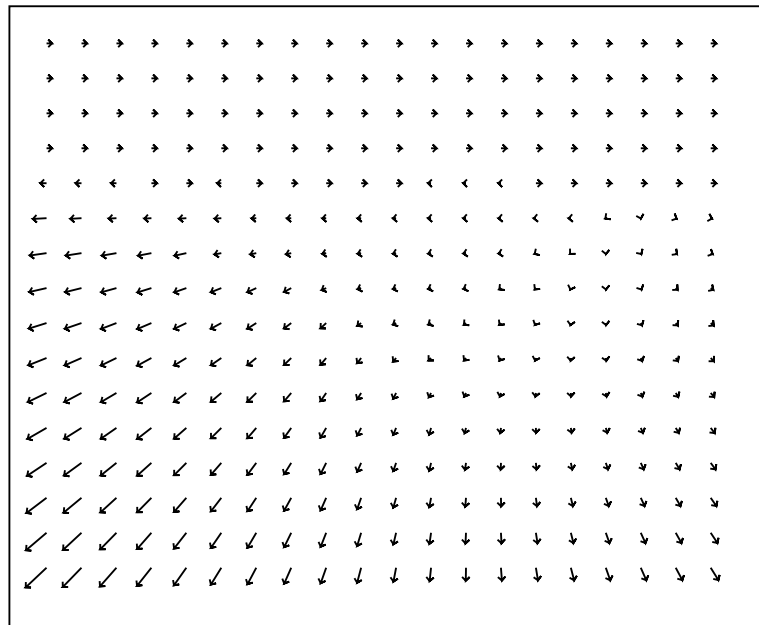
As can be seen in this chapter, a tremendous amount of research in the study of image motion has been based on a very local view of coherence both spatially and temporally. An interesting open problem put forth recently by Nagel (Nagel, 2000) is “what is assumed to remain constant becomes inexorably intertwined with the question over which extent in space and time one assumes the expected constancy to hold”. As pointed out by Nagel (Nagel, 2000) a theoretical foundation is required for this extended view. In the next chapter, approaches that extend the spatial and temporal dimensions for the purpose of making richer regional descriptions of motion are considered.

Technique	Global	Average Error	Standard Deviation	Density
Horn & Schunck ( <a href="#">Horn &amp; Schunck, 1981</a> )	✓	32.43°	30.28°	100%
Lucas & Kanade ( <a href="#">Lucas &amp; Kanade, 1981b</a> )		4.10°	9.58°	35.10%
Nagel ( <a href="#">Nagel, 1983b</a> )	✓	11.71°	10.59°	100%
Heeger ( <a href="#">Heeger, 1987</a> )		10.51°	12.11°	15.2%
Anandan ( <a href="#">Anandan, 1989</a> )	✓	15.84°	13.46°	100%
Fleet & Jepson ( <a href="#">Fleet &amp; Jepson, 1989</a> )		4.29°	11.24°	34.10%
Weber & Malik ( <a href="#">Weber &amp; Malik, 1995</a> )		4.31°	8.66°	64.2%
Black & Anandan ( <a href="#">Black &amp; Anandan, 1996</a> )	✓	4.46°	4.21°	100%
Farneback ( <a href="#">Farneback, 2001</a> )		1.14°	2.14°	100%
Bruhn et al. ( <a href="#">Bruhn et al., 2005</a> )	✓	1.02	N/A	100%
Roth & Black ( <a href="#">Roth &amp; Black, 2005</a> )	✓	1.47°	1.54°	100%

**Table 3.2.** Summary of *Yosemite* (Fig. 3.11) velocity results. The sky region is excluded for all results. The **Average Error** measures the average of (3.61) in the image (excluding velocities categorized as unreliable). **Standard Deviation** measures the standard deviation of (3.61) in the image (excluding velocities categorized as unreliable). **Density** summarizes the percentage of flow measurements in the image that were considered reliable based on the formulations confidence measure. **Global** indicates whether the corresponding estimator is based on a global regularization or a local formulation. N/A  $\equiv$  not available. The results are compiled from ([Barron et al., 1994b](#); [Farneback, 2001](#); [Bruhn et al., 2005](#); [Roth & Black, 2005](#)). Note that the results of Roth and Black ([Roth & Black, 2005](#)) report results on a version of the Yosemite sequence without the presence of the clouds.

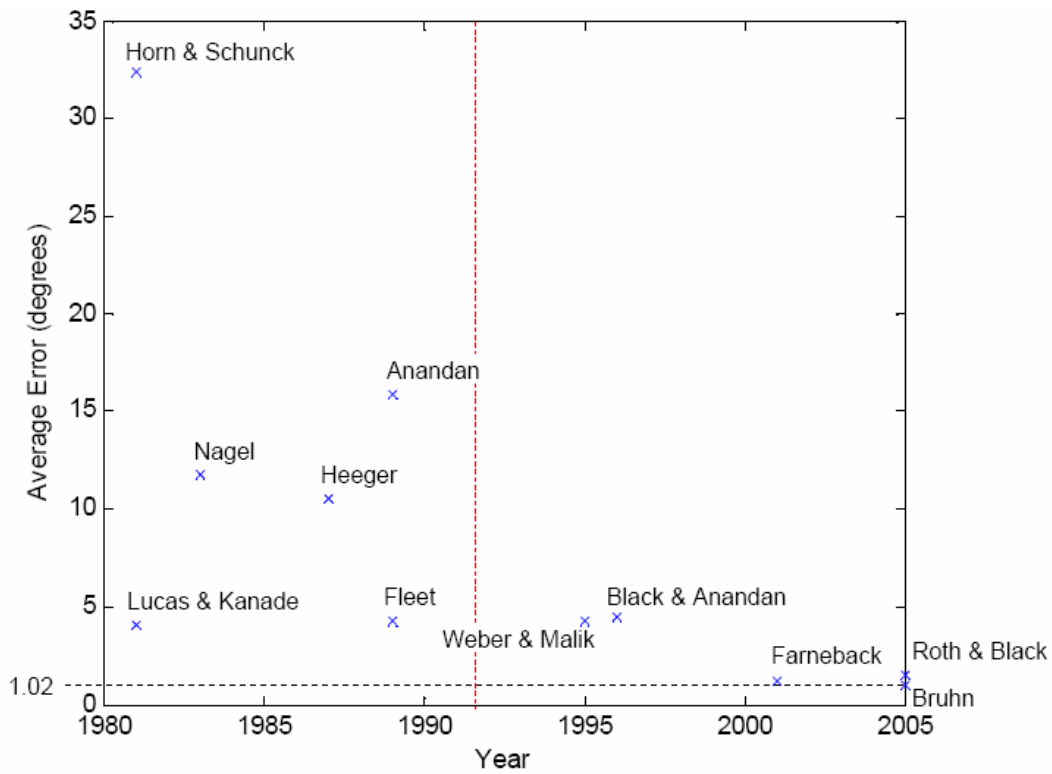


(a)



(b)

**Figure 3.11.** *Yosemite* sequence. (a) depicts a single frame from the *Yosemite* sequence and (b) the correct velocity field for this frame. The sequence was synthetically generated by Lynn Quam at SRI.



**Figure 3.12.** Yosemite sequence performance. Chronological performance results of optical flow approaches (summarized in Table 3.2) on the *Yosemite* sequence as measured by the average angular error (3.61) in the image. Results to the left of the vertical dotted line are those reported by Barron et al. (Barron et al., 1994a) based on their own independent implementations. A general chronological trend in the reduction of the angular error is evident.





## Chapter 4

# Beyond Infinitesimal Descriptors: Regional Descriptors of Motion

RECENTLY, there has been a surge in the general use of video due to increases in computational power and storage, the low-cost of image capturing devices and the interconnected world that facilitates rapid exchange of data. With this availability there is a growing need for the automatic organization (interpretation) of this data into semantically meaningful categories. Image motion is a key element of video that can provide such an organization. This chapter reviews approaches that relax the consideration of infinitesimal region-based descriptors, as considered in Chapter 3, in favour of regional descriptors of motion that have the potential to provide this organization.

Approaches for regional descriptors of motion considered in this chapter are broadly categorized as: *parametric motion models* (Section 4.1) for describing the motion within a given image region, *layered motion representations* (Section 4.2) that attempt to recover the coherent regions (layers) in the imaged scene, *temporal textures* (Section 4.3) that characterize a restricted class of naturally occurring motions based on statistical considerations, analytic considerations of a given optical flow field (Section 4.4) and methods that use qualitative means to characterize the spatiotemporal structure of the image sequence (Section 4.5).

### 4.1 Parametric motion models

In Chapter 3, the description of motion was limited to the analysis of infinitesimal regions of spatiotemporal extent. By relaxing the infinitesimal restriction to regional considerations, the main focus of this chapter, we can now move to describing the structure of a flow field within a region, as well as the velocity at points. This section is concerned with reviewing parametric motion models that describe the structure of

flow fields within a non-infinitesimal region of space and an infinitesimal region of time. The problem of estimating the model parameters can be solved by assuming brightness constancy and using the local regression methods (e.g., (Bergen et al., 1992)) considered in Chapter 3.2.1 and is not considered further.

Numerous parametric models have been proposed in the literature for the purpose of describing motion. Many of these models may be brought within the following common framework of understanding: an  $n$ th-order (Taylor) series expansion of the velocity field  $\mathbf{v}$  at a point  $(x_0, y_0)$ ,

$$\mathbf{v}(x, y, t) = \sum_{j=0}^p \sum_{k=0}^q \frac{\partial^{j+k} \mathbf{v}(x, y, t)}{\partial x^j \partial y^k} \frac{x^j y^k}{j!k!} \quad \text{where } p + q = n, \quad (4.1)$$

with specialization of the coefficients.

The translational model represents the zeroth-order expansion of the flow field,

$$u(x, y, t) = a_0 \quad (4.2)$$

$$v(x, y, t) = a_3 \quad (4.3)$$

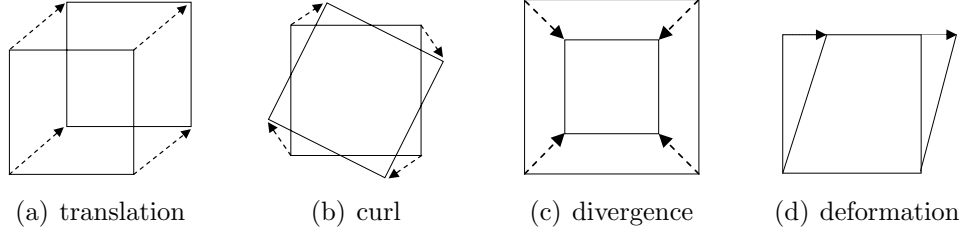
where  $a_0 = u_0$  and  $a_3 = v_0$  represent the instantaneous flow (i.e., velocity) at the point  $(x_0, y_0, t_0)^\top$ . Note the purpose for the non-sequential labeling of the zeroth-order terms  $a_0$  and  $a_3$  is for consistency in exposition with the higher-order expansions to follow. This model was assumed throughout the presentation of the least-squares regression approaches in Chapter 3.2.1. It is important to make a distinction between the use of zeroth-order models in the context of local methods (the topic of Chapter 3) and regional approaches which are the focus of this chapter. In the local context we sought a small region about a pixel for the purpose of stabilizing the flow estimate at a point, in contrast in this section we seek the description of a region with extended support.

Considering a first-order spatial expansion of the velocity field, yields,

$$u(x, y, t) = a_0 + a_1 x + a_2 y$$

$$v(x, y, t) = a_3 + a_4 x + a_5 y$$

where  $a_1 = u_x$ ,  $a_2 = u_y$ ,  $a_4 = v_x$  and  $a_5 = v_y$ . This model is commonly termed the *affine flow* model. A great deal of work has been devoted to the affine model (e.g., (Fuh & Maragos, 1991; Bergen et al., 1992; Campani & Verri, 1992; Irani et al., 1994; Tomasi & Shi, 1994; Derpanis et al., 2004)) since it provides a reasonable approximation of the motion of planar and smooth surfaces within a small region of interest (see Chapter 2.2.1 for details). Equivalently, the affine model may be reexpressed as the sum of the following kinematic quantities (Koenderink & van Doorn, 1975; Koenderink & van Doorn, 1976), *translation*, a rotational component



**Figure 4.1.** Kinematic motion. Transformations of a square element by the kinematic motions are depicted.

termed *curl*, an isotropic expansion/contraction termed *divergence* (div) and an area conserving oriented shear, termed *deformation* (def), formally,

$$\mathbf{v} = \mathbf{T} + \frac{1}{2} \left( \text{curl} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} + \text{div} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \text{def} \mathbf{S} \right) \mathbf{x} \quad (4.4)$$

where

$$\text{translation} := \mathbf{T} = (a_0, a_3)^\top \quad (4.5)$$

$$\text{curl} := a_4 - a_2 \quad (4.6)$$

$$\text{div} := a_1 + a_5 \quad (4.7)$$

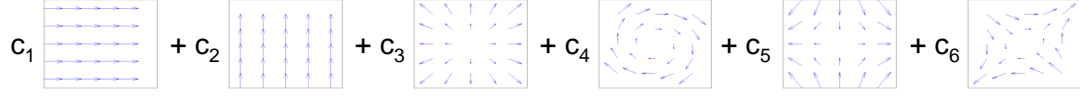
$$\text{def} := ([a_1 - a_5]^2 + [a_2 + a_4]^2)^{1/2} \quad (4.8)$$

and

$$\mathbf{S} = \mathbf{Q}^{-1} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{Q} \quad (4.9)$$

describes an expansion in a certain direction and a contraction in the orthogonal direction; the direction is encoded in the rotation matrix  $\mathbf{Q}$ . Fig. 4.1 illustrates representative transformations of an image by the kinematic quantities. An important feature of this representation is that the curl, divergence and deformation, are invariant to rigid transformations of the image coordinate frame. Owing to the kinematic quantities' semantic description, several authors have reported mappings of the kinematic quantities to semantically meaningful human movements, such as facial expressions (Black & Yacoob, 1997) and hand gestures (Derpanis et al., 2004). From a biological perspective, there is experimental evidence that such a decomposition may be utilized in the primate visual system for processing motion (Graziano et al., 1994; Martinez-Trujillo et al., 2005). Specializations of first-order motion models can be found by restricting consideration to single or combinations of the kinematic quantities in (4.4).

$\mathbf{u}(\mathbf{x}; \mathbf{c}) =$



**Figure 4.2.** Affine motion basis. The representation of an arbitrary affine velocity field as expressed as a sum of basis flows, where  $c_i, i = 1, \dots, 6$  represent scalar coefficients.

The final expansion considered here is the second-order spatial expansion of  $\mathbf{u}(x, y, t)$ , yields,

$$u(x, y, t) = a_0 + a_1x + a_2y + a_6x^2 + a_7xy \quad (4.10)$$

$$v(x, y, t) = a_3 + a_4x + a_5y + a_8xy + a_9y^2 \quad (4.11)$$

where  $a_6 = \frac{1}{2}u_{xx}$ ,  $a_7 = \frac{1}{2}u_{xy}$ ,  $a_8 = \frac{1}{2}v_{yx}$  and  $a_9 = \frac{1}{2}v_{yy}$ . Assuming that  $a_6 = a_8$  and  $a_7 = a_9$ ,

$$u(x, y, t) = a_0 + a_1x + a_2y + a_6x^2 + a_7xy \quad (4.12)$$

$$v(x, y, t) = a_3 + a_4x + a_5y + a_6xy + a_7y^2 \quad (4.13)$$

the resulting flow field description corresponds exactly to that of a moving planar surface (for details see Chapter 2.2.1). This model is usually termed the *quadratic flow model* (e.g., (Horn, 1986; Bergen et al., 1992; Irani et al., 1994; Black & Yacoob, 1997)). In relatively small regions of analysis, the affine model is preferred over the quadratic model since the second order coefficients, being small in magnitude, are unreliable to estimate due to image noise (Negahdaripour & Lee, 1991). From a kinematic perspective, Black and Yacoob (Black & Yacoob, 1997) extend the affine-based kinematic description by attributing the parameters  $a_6$  and  $a_7$  approximately to the motions of “yaw” and “pitch”.

Alternatively, the polynomial-based motion models detailed above can be considered as a linear combination of basis flows (Fleet et al., 2000),

$$\mathbf{u}(x, y, t; \mathbf{c}) = \sum_{i=1}^n c_i \mathbf{b}_i(x, y) \quad (4.14)$$

where  $\{\mathbf{b}_i(x, y)\}$  is the basis set and  $\{c_i\}$  are scalar coefficients,  $i = 1, \dots, n$ . Fig. 4.2 depicts the affine model as expressed by a linear combination of basis flows. With this linear basis-set interpretation in mind one can now ask: what other basis flows are useful?

Hoey and Little (Hoey & Little, 2000) introduce the use of Zernike polynomials (Zernike, 1934) to model the flow field induced by facial expressions. Zernike polynomials are an orthogonal basis set of complex polynomials defined within the

unit circle. The lowest order Zernike polynomials correspond to the standard affine basis. The next order polynomials correspond approximately to yaw, pitch and roll and following orders represent motions with higher spatial “frequency”. Hoey and Little argue that the utility of the Zernike polynomials lies in their simplicity and high expressiveness while providing a general basis for a broad range of motions.

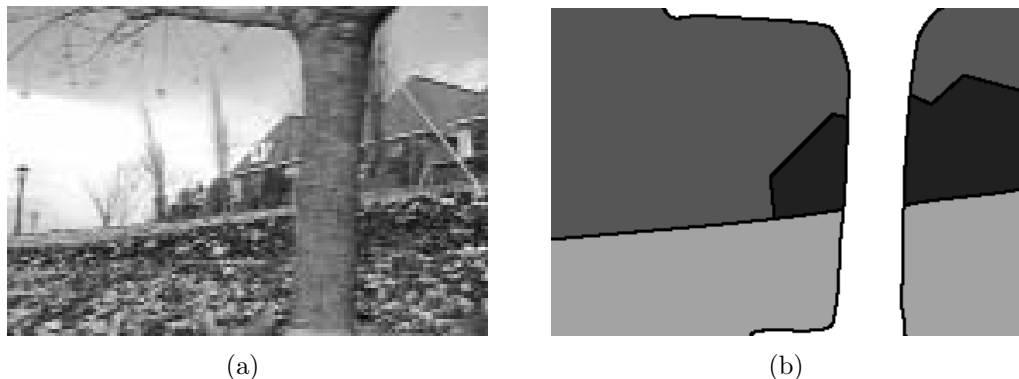
Fleet et al. (Fleet et al., 1998; Fleet et al., 2000) argue that the above models have limited applicability to the analysis of image regions from complex natural scenes. Fleet et al. (Fleet et al., 1998; Fleet et al., 2000) extend the use of linear parameterized models to natural complex models (e.g., mouth movements and human gait) while forgoing geometric modeling. Instead, the basis set is “learnt” off-line using principle components analysis (Oja, 1983) on a set of exemplar flows of specific classes of motions.

## 4.2 Layered motion model representations

Layered models attempt to identify regions in an image that exhibit “coherent” motion. Introduced formally by Adelson (Adelson & Anandan, 1990; Adelson, 1991; Adelson, 1995), layered motion representations are inspired by traditional cel animation (Solomon, 1994). In brief, cel animation consists of a series of images painted on sheets of clear celluloid (i.e., the cels) where the depth ordering of the cels determines the occlusion relationships amongst the cels and images in each layer are restricted to a common motion. Given this assumed representation of a video sequence, approaches based on layered representations attempt to invert this formation process by extracting the constituent layers and their respective coherent (parametric) motions (see Fig. 4.3 for an example layered decomposition of a scene).

Probably the earliest incarnation of a layered model approach appeared in the work of Fennema and Thompson (Fennema & Thompson, 1979). The approach uses the Hough transform (Hough, 1962) to cluster points in the scene that exhibit coherent translation and estimate the motion of the respective clusters.

Sequential application of dominant motion estimators have been proposed for extracting multiple motions. In the first pass, the global dominant motion present in the image is estimated. Once the dominant motion is estimated, the image pixels consistent with the dominant motion are removed from further consideration and the estimation process is iterated to identify the remaining motions. Black and Anandan (Black & Anandan, 1993; Black & Anandan, 1996), Odobez and Bouthemy (Odobez & Bouthemy, 1995) and Darrell and Pentland (Darrell & Pentland, 1995) incorporate M-estimators (discussed in Chapter 3.2: Robust methods) as the dominant motion estimator which treats the non-dominant motions as outliers. Bober and Kittler (Bober & Kittler, 1994) combine a Hough-based approach with an M-estimator to extract successive dominant motions. M-estimator-based approaches fail when two



**Figure 4.3.** Layered model decomposition. (a) Frame 1 of the **Garden Sequence** and (b) depicts a potential layered (coherent regions highlighted in different grayscales) decomposition of (a).

regions are at the same scale. This is due to the fact that the support of the respective regions exceeds the breakdown point of the robust estimator and thus neither region can be considered an outlier (see Chapter 3.2.4 for details). Several contributions (Burt et al., 1989; Bergen et al., 1990; Burt, 1991; Burt et al., 1991; Bergen et al., 1991; Irani et al., 1994) employ pyramid-based motion estimators due to the dominant motion “lock-on” characteristics of these estimators (see Chapter 3.4).

Wang and Adelson (Wang & Adelson, 1993; Wang & Adelson, 1994) propose an approach consisting of two stages: (1) compute motion estimates using a least-squares approach (Lucas & Kanade, 1981b) within non-overlapping square image patches and (2) use K-means clustering (Bishop, 1995) to group motion estimates into regions exhibiting consistent affine flow.

A significant amount of recent research cast the extraction of layers within a probabilistic framework (Jepson & Black, 1993a; Jepson & Black, 1993b; Ayer & Sawhney, 1995; Weiss & Adelson, 1996; Weiss, 1997; Jojic & Frey, 2001; Wong et al., 2004). The scene is modeled by a parametric mixture model consisting of latent variables (the layer assignment of the pixels) and unknown parameters (the layer motions). The original solution to this problem, first proposed by Jepson and Black (Jepson & Black, 1993a; Jepson & Black, 1993b), computes the maximum likelihood estimate using the *expectation maximization* (EM) algorithm (Dempster et al., 1977). K-means clustering may be considered as a special case of the EM algorithm. Specifically, the EM algorithm and K-means approach coincide in the case where the underlying parametric model is a Gaussian mixture model with  $K$  Gaussian components, the covariances are equal and the means are unknown (Mitchell, 1997). The basic idea behind EM-based approaches is to estimate both the unknown parameters of the layer distributions and the layer assignments by iterating between the following two steps until convergence:

- **Expectation Step (E-Step):** Calculate the expected value of the layer assignments based on the current estimates of the parameters of the layer distributions.
- **Maximization Step (M-Step):** Calculate the new maximum likelihood estimate for the layer parameters based on the current expected values of the layer assignments.

The procedure is iterated until a small change in the parameters is realized. This iterative process is guaranteed to increase the log likelihood (Dempster et al., 1977), though the solution may converge to a sub-optimal local maximum and/or convergence may be slow.

A shortcoming of the approach as presented in (Jepson & Black, 1993a; Jepson & Black, 1993b) is that the number of layers are assumed to be known. Several authors (Ayer & Sawhney, 1995; Weiss & Adelson, 1996) have partially addressed this issue by proposing extensions that incorporate the estimation of the number of layers (i.e., models) into the basic EM framework. A further drawback is that the choice of initial parameters may lead to a suboptimal solution (i.e., local maxima) due to the local search nature of the algorithm. To ameliorate this problem in practice, the EM algorithm is usually run several times using random initial parameters.

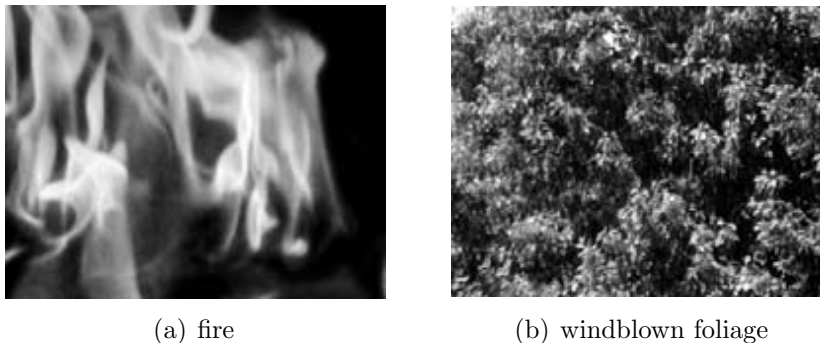
## 4.3 Temporal textures

In this section the recognition of a restricted class of naturally occurring motions that have garnered special attention in the literature is reviewed, these motions are commonly referred to as *temporal textures* or *dynamic textures*. These motions are characterized by regions in space-time that exhibit statistical regularity and have indeterminate extents. Examples of temporal textures include (see Fig. 4.4): windblown foliage, turbulent flow in cloud patterns, ripples on water and falling snow.

The study of temporal textures can be traced back to the seminal work of Nelson and Polana (Nelson & Polana, 1992; Polana & Nelson, 1997). Temporal texture analysis shares a great deal of similarity with classical gray-level texture analysis (e.g., (Rosenfeld & Troy, 1970; Haralick, 1971; Haralick, 1979; Connors & Harlow, 1980)) in that it is concerned with identifying invariances of a region of indeterminate extent. Given these similarities, the work of Nelson and Polana (Nelson & Polana, 1992; Polana & Nelson, 1997) as well as subsequent work (e.g. (Bouthemy & Fablet, 1998; Peh & Cheong, 2002; Rahman & Murshed, 2004; Lu et al., 2005)) centre around adapting existing statistical techniques used for gray-level texture analysis to analyze temporal textures.

The majority of proposed approaches to temporal texture recognition (Nelson & Polana, 1992; Polana & Nelson, 1997; Bouthemy & Fablet, 1998; Peh & Cheong,





**Figure 4.4.** Example frames of temporal textures (Courtesy of Martin Szummer's temporal texture database (Szummer, 1996)).

2002; Rahman & Murshed, 2004; Lu et al., 2005), rely on the extraction of highly discriminable features from the normal flow (2.26) that are combined together to form signatures for each of the temporal textures under consideration. The normal flow field is selected in favor of the optical flow field based on the following arguments (Nelson & Polana, 1992): the extraction of optical flow is time consuming, the quality of the extracted optical flow is quite low for such complex dynamic scenes and one should throw out as much information as possible on the grounds that limited computational resources should be concentrated on the essential information. Nelson and Polana (Nelson & Polana, 1992; Polana & Nelson, 1997) extract several statistical features from the normal flow field, these include: the mean magnitude, estimates of the expansion or contraction (divergence), estimates of the rotation (curl) and directional difference statistics extracted from a cooccurrence matrix<sup>1</sup>. Cooccurrence matrices are used to measure the spatial dependence among features. Historically, Julesz introduced cooccurrence matrix statistics in the context of human texture discrimination experiments (Julesz, 1962), while Rosenfeld and Troy (Rosenfeld & Troy, 1970) and Haralick (Haralick, 1971) introduced them for machine vision analysis of textures. The classification of the features is based on a nearest neighbour classifier. Nelson and Polana report 100% success in the classification of seven different texture samples.

A shortcoming of the Nelson and Polana work is that the temporal evolution of the textures is ignored since only spatial interactions of the normal flow computed within a single frame are considered. To address this issue, Bouthemy and Fablet (Bouthemy & Fablet, 1998) transfer the study of cooccurrence relationships to the temporal domain to analyze purely temporal interactions while ignoring spatial interactions.

---

<sup>1</sup>A cooccurrence matrix contains the relative frequencies  $P(i, j)$  with which two neighboring resolution cells separated by a fixed distance  $d$  (or at fixed angle  $\theta$  relative to each other), one with feature  $i$  and the other with feature  $j$ ; these matrices are symmetric (Haralick, 1979).

More recently, Peh and Cheong (Peh & Cheong, 2002) analyze joint statistics of the spatial and temporal aspects of temporal textures.

While the studies reviewed above demonstrate impressive recognition rates, they also share several shortcomings. A main limitation of these approaches, is that the problem of determining the region on analysis is assumed away by using presegmented data. The extracted normal flow is highly dependent on the appearance of the temporal texture (Polana & Nelson, 1997), this limits the ability of normal flow-based temporal texture approaches to abstract prototypical descriptors for a class of textures. For example, in the case of modeling a class of flags blowing in the wind, the extracted temporal texture descriptors will be highly dependent on the flag’s appearance. There is also the open question of selecting the discriminative features in a principled manner; the approaches above select features based on intuition and trial and error. Finally, the reported recognition rates are based on a small set of temporal textures (typically 10). Thus, reported results shed little light on the general applicability of these approaches.

Several recent studies (Saisan et al., 2001; Doretto et al., 2003) have presented model-based approaches for the recognition of temporal textures. Each temporal texture is assumed to be the result of a second-order stationary<sup>2</sup> process that is modeled as the output of a stochastic linear-dynamical system, formally,

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{v}(t) \quad \leftarrow \text{state model} \quad (4.15)$$

$$I(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{w}(t) \quad \leftarrow \text{observation model} \quad (4.16)$$

where  $\mathbf{x}(t) \in \mathbb{R}^n$  represents the (hidden) state of the model,  $I(t) \in \mathbb{R}^m$  the observed image (in lexicographic order),  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $\mathbf{C} \in \mathbb{R}^{m \times n}$  represents the static model parameters and the noise processes  $\mathbf{v}(t)$  and  $\mathbf{w}(t)$  are i.i.d. distributed as  $\mathcal{N}(0, \mathbf{Q})$  and  $\mathcal{N}(0, \mathbf{R})$ , respectively. The recognition of a texture consists of the following two steps:

1. fit the model (Eqs. (4.15) and (4.16)) to the texture
2. use the “learned” model parameters as a signature to index into a database of known temporal textures.

Notice here that we are dealing with spatiotemporal structure directly, as opposed to building the representation on top of the recovered flow, as considered above. Saisan et al. (Saisan et al., 2001) report that using this model they were able to achieve approximately 90% correct recognition on a database of 200 temporal textures.

Drawbacks that limit the applicability of the approach include the need for pre-segmented inputs, the potential failure of the assumed dynamical model to generally

---

<sup>2</sup>A stochastic process of order  $k$  is stationary if the joint statistics up to order  $k$  are time-invariant. For example a process  $I(t)$  is second-order stationary if its mean  $E\{I(t)\}$  is constant and its covariance  $E\{(I(t_i))(I(t_j))\}$  only depends on  $i - j$  (Papoulis & Pillai, 2002).

characterize the class of temporal textures and the unwieldy computational requirements. In terms of computational requirements, Saisan et al. (Saisan et al., 2001) report that it takes about 5 minutes to fit a model to a texture consisting of 150 color frames at a resolution of  $320 \times 220$  coded in Matlab using a 1GHz desktop PC.

## 4.4 Optical flow-based reasoning

Given an extracted optical flow field we next consider methods for interpreting the flow field based on analytic considerations.

Some of the earliest work focused on the detection of *dynamic occlusion boundaries* in the optical flow field. Dynamic occlusion boundaries are defined as regions where the flow properties (direction and/or magnitude) differ on either side of the boundary. Knowledge of the location of dynamic occlusion boundaries play the following useful roles:

- provide shape information that may be absent from other sources of information (e.g., intensity edges)
- facilitate object segmentation
- improve optical flow estimates around boundaries.

Generally, these early attempts (Nakayama & Loomis, 1974; Clocksin, 1980; Thompson et al., 1985) centre on the use of local edge operators to detect discontinuities in the optical flow field. Thompson et al. (Thompson et al., 1985) extend this work with an approach to identify the side of the occlusion boundary that corresponds to the occluding surface. This is based on the principle that the occlusion boundary moves with the image region of the occluding surface. Spoerri and Ullman (Spoerri & Ullman, 1990) detect occlusion boundaries based on the local distribution of flow vectors. These methods have proven unreliable due to their dependence on accurate flow estimates around the discontinuities; exactly where obtaining reliable optical flow estimates are difficult.

Other flow field analyses have relied on differentiating the flow to extract the first-order terms  $u_x, u_y, v_x$  and  $v_y$  encompassed in the Jacobian of the series expansion of the flow field. For instance, Subbarao (Subbarao, 1990) assuming that the local surface structure is approximately planar and by extension the sufficiency of an affine representation of the flow-field (see Chapter 2.2.1), demonstrated that the Jacobian of the flow field is sufficient to determine: the maximum and minimum time-to-collision of the observer to the object, and the maximum and minimum angular velocity of the object along the direction of view.

Beyond the difficulty of extracting good flow estimates, there is the issue that the optical flow extracted is generally quantitatively different from the quantity it

is intended to measure, the motion field, unless very special conditions hold (Verri & Poggio, 1987a; Verri & Poggio, 1987b; Verri & Poggio, 1989) (see Chapter 2.2.2 for the list of conditions). Consequently, several studies have investigated extracting qualitative types of information from the optical flow (eg., (Verri & Poggio, 1987a; Verri & Poggio, 1987b; Verri & Poggio, 1989; Wildes, 1993; Cohen & Herlin, 1996; Cohen & Herlin, 1999)). Many of these studies are based on *dynamical systems theory* (Hirsch & Smale, 1974) or *singularity theory* (Arnold, 1991).

Dynamical systems theory uses the notion of a phase portrait to represent geometrically the solution of a differential equation. The basic idea behind phase portrait methods in the context of optical flow representation (Cohen & Herlin, 1996; Cohen & Herlin, 1999; Koenderink & van Doorn, 1975; Koenderink & van Doorn, 1976; Verri & Poggio, 1987a; Verri & Poggio, 1987b; Verri & Poggio, 1989) is to locally approximate the flow pattern around singular points<sup>3</sup> by an affine model (i.e., two-dimensional linear differential equations), formally,

$$\begin{aligned} u(x, y, t) &= a_0 + a_1x + a_2y \\ v(x, y, t) &= a_3 + a_4x + a_5y \end{aligned}$$

or written more compactly in matrix notation,

$$\mathbf{u} = \mathbf{t} + \mathbf{J}\mathbf{x} \quad (4.18)$$

where,

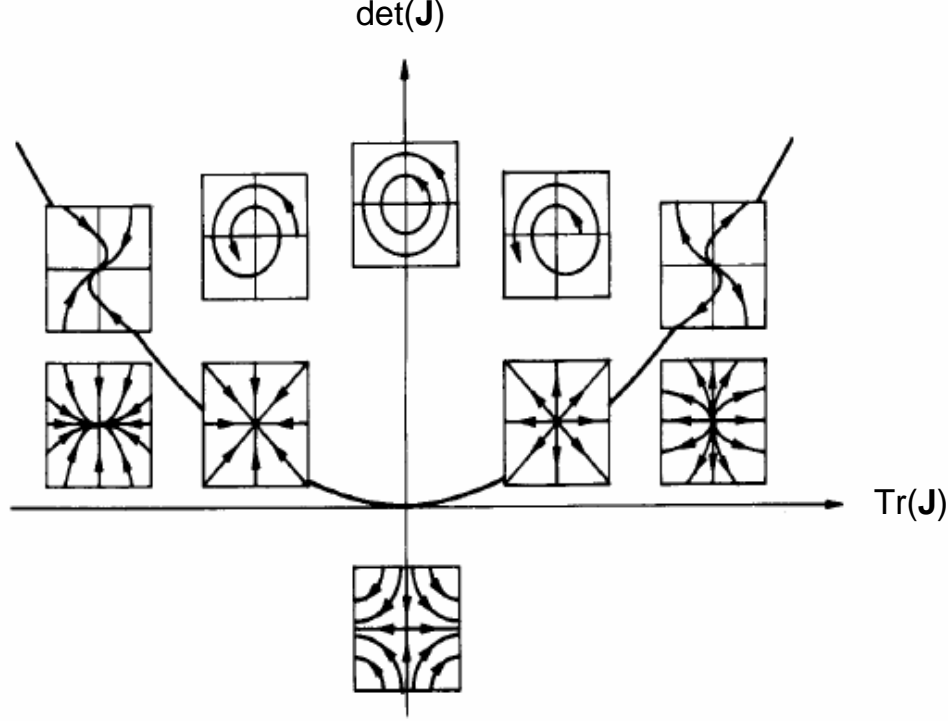
$$\mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} a_0 \\ a_3 \end{pmatrix}, \quad \mathbf{J} = \begin{bmatrix} a_1 & a_2 \\ a_4 & a_5 \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix} \quad (4.19)$$

followed by qualitative classification. Notice that this locally affine consideration of the flow field implies that this analysis is limited to the flow field induced by locally smooth (planar-like) surface structures. Furthermore, the structure of the Jacobian matrix  $\mathbf{J}$  is a function of the motion and surface parameters of the surface in the world (see Chapter 2.2.1 for details). For a non-singular (i.e., invertible) Jacobian matrix  $\mathbf{J}$  there are only a finite number of possible qualitatively different descriptions of the neighbourhood of a singular point, these include (see columns 1 and 2 of Table 4.1): node, saddle, star, improper, centre and spiral. Classification into one of the prototypical patterns is based on the eigenvalues of matrix  $\mathbf{J}$  (summarized in Table 4.1 and pictorially in Fig. 4.5) or in terms of the kinematic decomposition of  $\mathbf{J}$  (see Section 4.1 Eqs. (4.4)-(4.8) for the definition of the kinematic decomposition and

---

<sup>3</sup>Singular points (also known as critical points, equilibrium solutions, steady state solutions and fixed points) are defined as the solution(s)  $\mathbf{x}$  to

$$\mathbf{u} = \mathbf{t} + \mathbf{J}\mathbf{x} = \mathbf{0} \quad (4.17)$$



**Figure 4.5.** Phase portrait classification. A geometric representation is depicted of the linear phase portrait in the determinant (det) - trace (tr) space of  $\mathbf{J}$ , where  $\det(\mathbf{J}) = \lambda_1 \lambda_2$ ,  $\text{tr}(\mathbf{J}) = \lambda_1 + \lambda_2$  and  $\lambda_1, \lambda_2$  are the eigenvalues of  $\mathbf{J}$ .

Table 4.2 for a summary of the kinematic-based phase portrait classification) (Shu & Jain, 1994). In contrast to the general quantitative difference between the optical and motion flow descriptions, the qualitative description of the singular points between the two flows will remain the same as long as they are “close” in the topological sense (Verri & Poggio, 1989). Besides describing image motion, this theory has been used to analyze flow fields in diverse contexts, including, fluid flow and texture analysis (Kass & Witkin, 1987; Rao & Jain, 1990; Helman & Hesselink, 1990; Helman & Hesselink, 1991; Rao & Jain, 1992; Shu & Jain, 1992; Shu & Jain, 1994).

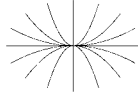
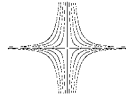
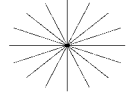
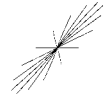
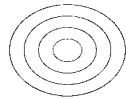
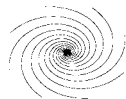
Wildes (Wildes, 1993) utilizes *singularity theory* to qualitatively describe the *visual motion field* (2.9) induced by a moving surface. Singularity theory is concerned with the study of points and sets of *singular points*. *Singular points* are defined as image points where the determinant of the Jacobian matrix  $\mathbf{J}$  (encompassing the first-order terms of the spatial structure of the velocity field  $u_x, u_y, v_x$  and  $v_y$ ) vanish,

$$\det(\mathbf{J}) = \det \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} = 0. \quad (4.20)$$

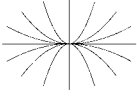
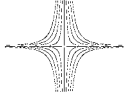
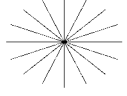

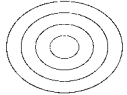
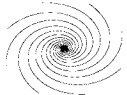
Wildes’ analysis is restricted to surfaces moving with either a pure three-dimensional

rotational or translational motion. By considering purely rotational three-dimensional motion, the surface structure component  $Z$  which is coupled exclusively with the translational parameters of the motion field (2.9) vanishes. In this case, Wildes' demonstrates that singularity analysis yields: a signature indicative of the presence of three-dimensional rotational motion, the axis of angular rotation and the relative magnitude of the angular and radial rotations. In the case of pure translation, the surface parameterization must be considered. Wildes' assumes that the surface is represented by a second-order series expansion (a *Monge patch*):  $Z(X, Y) = \frac{1}{2}\kappa_1 X^2 + \frac{1}{2}\kappa_2 Y^2 + \kappa_3 XY + pX + qY + r$ . In this case, the singularity analysis yields: signatures of the qualitative three-dimensional surface shape, specifically, whether the shape is locally elliptic, hyperbolic or parabolic, the major and minor axes of the surface and constraints on the direction of the angular translation and surface gradient.

A general criticism of the approaches discussed above is that limited (if any) empirical validation has been reported to validate their respective claims.

Flow Pattern	Phase Portrait	Jordan Form	Eigenvalue-Based Classification
node		$\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$ $\lambda_1 \times \lambda_2 > 0$	Real Eigenvalues $\lambda_1 < \lambda_2$
saddle		$\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$ $\lambda_1 \times \lambda_2 < 0$	$\lambda_1 \neq 0$ and $\lambda_2 \neq 0$
star		$\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$	Real Equal Eigenvalues $\lambda_1 = \lambda_2$
improper		$\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$	$\lambda_1 \neq 0$ and $\lambda_2 \neq 0$
centre		$\begin{bmatrix} \alpha & \beta \\ \beta & \alpha \end{bmatrix}$ $\alpha = 0$	Complex Eigenvalues
spiral		$\begin{bmatrix} \alpha & \beta \\ \beta & \alpha \end{bmatrix}$ $\alpha \neq 0$	$\lambda_1 = \alpha + i\beta$ $\lambda_2 = \alpha - i\beta$ $\beta \neq 0$

**Table 4.1.** Eigenvalue-based classification of first-order planar phase portraits.

Flow Pattern	Phase Portrait	Kinematic-Based Classification
node		$(\text{def}(\mathbf{v}))^2 > (\text{curl}(\mathbf{v}))^2$ and $ \text{div}(\mathbf{v})  > \sqrt{(\text{def}(\mathbf{v}))^2 - (\text{curl}(\mathbf{v}))^2}$
saddle		$(\text{def}(\mathbf{v}))^2 > (\text{curl}(\mathbf{v}))^2$ and $ \text{div}(\mathbf{v})  < \sqrt{(\text{def}(\mathbf{v}))^2 - (\text{curl}(\mathbf{v}))^2}$
star		$(\text{def}(\mathbf{v}))^2 = (\text{curl}(\mathbf{v}))^2 = 0$ and $\text{div}(\mathbf{v}) \neq 0$
improper		$(\text{def}(\mathbf{v}))^2 = (\text{curl}(\mathbf{v}))^2 \neq 0$ and $\text{div}(\mathbf{v}) \neq 0$
centre		$(\text{def}(\mathbf{v}))^2 < (\text{curl}(\mathbf{v}))^2$ and $\text{div}(\mathbf{v}) = 0$
spiral		$(\text{def}(\mathbf{v}))^2 < (\text{curl}(\mathbf{v}))^2$ and $\text{div}(\mathbf{v}) \neq 0$

**Table 4.2.** Kinematic-based classification of first-order planar phase portraits (Shu & Jain, 1994).



Spatiotemporal gray value structure	Rank(S)	Eigenvalues
<i>homogeneous region</i>	0	$\lambda_1 = \lambda_2 = \lambda_3 = 0$
<i>aperture problem</i>	1	$\lambda_1 \geq 0$ and $\lambda_2 = \lambda_3 = 0$
<i>coherent motion</i>	2	$\lambda_1, \lambda_2 \geq 0$ and $\lambda_3 = 0$
<i>incoherent motion</i>	3	$\lambda_1, \lambda_2, \lambda_3 \geq 0$

**Table 4.3.** Regional classification by eigenvalue analysis.

## 4.5 Spatiotemporal structure-based reasoning

In the previous section the primary focus was on the qualitative interpretation of the structure of a given optical flow field. In contrast, the following section is concerned with qualitative approaches that make semantically meaningful abstractions of the local spatiotemporal structure, while foregoing explicit computation of optical flow.

### 4.5.1 Structure tensor methods

Structure tensor-based methods have been demonstrated to provide precise quantitative estimates of the optical flow field from image sequences (Jähne, 1990; Bigün et al., 1991; Haußecker et al., 1998; Middendorf & Nagel, 2001; Mota et al., 2001; Spies & Jähne, 2001; Liu et al., 2003). As noted in Chapter 3.2.1, the tensor-based approach for optical flow estimation is equivalent to the total least-square approach. In this section, qualitative aspects of the structure tensor  $\mathbf{S}$  (3.12) for categorizing a spatiotemporal region are examined. Interestingly, the structure tensor and its qualitative analysis have also appeared in the context of corner detection in the spatial domain (Förstner & Gülch, 1987; Harris & Stephens, 1988) and texture-based segmentation (Zhang & Nagel, 1994).

A qualitative description of motion may be extracted through an eigenvalue analysis (or equivalently matrix rank analysis) of the structure tensor  $\mathbf{S}$  (3.12) (Jähne, 2005). The eigenvalue analysis yields the following qualitative descriptions:

- a) *homogeneous*: gray-level region, where no velocity estimate is possible
- b) *aperture problem*: the normal velocity can only be estimated
- c) *coherent motion*: both components of the velocity vector can be estimated
- d) *incoherent motion*: the translation model fails to hold.

Table 4.3 summarizes an idealistic classification of the possible patterns. In practice

due to the effects of local windowing and noise these ideal signatures can not be recovered, this necessitates the introduction of user-defined thresholds to differentiate the various classes. It is instructive to point out that equivalently the qualitative analysis can be made by considering the energy spectrum of a local patch (Jähne, 2005). In this case, the frequency space analog of the structure tensor, the *inertial tensor* (Bigün et al., 1991), is qualitatively analyzed. A drawback of the structure tensor analysis is that only a small number of categories can be distinguished due to the finite number of eigenvalues; the *non-coherent motion* category represents a default catch-all for more complex motions, such as transparency motion.

Next, let us consider a visualization of the structure tensor and its qualitative analysis. Recall from our discussion in Chapter 3.2.1 that the reformulated optical flow constraint with the extra degree of freedom  $w$  used to define the structure tensor is given as,

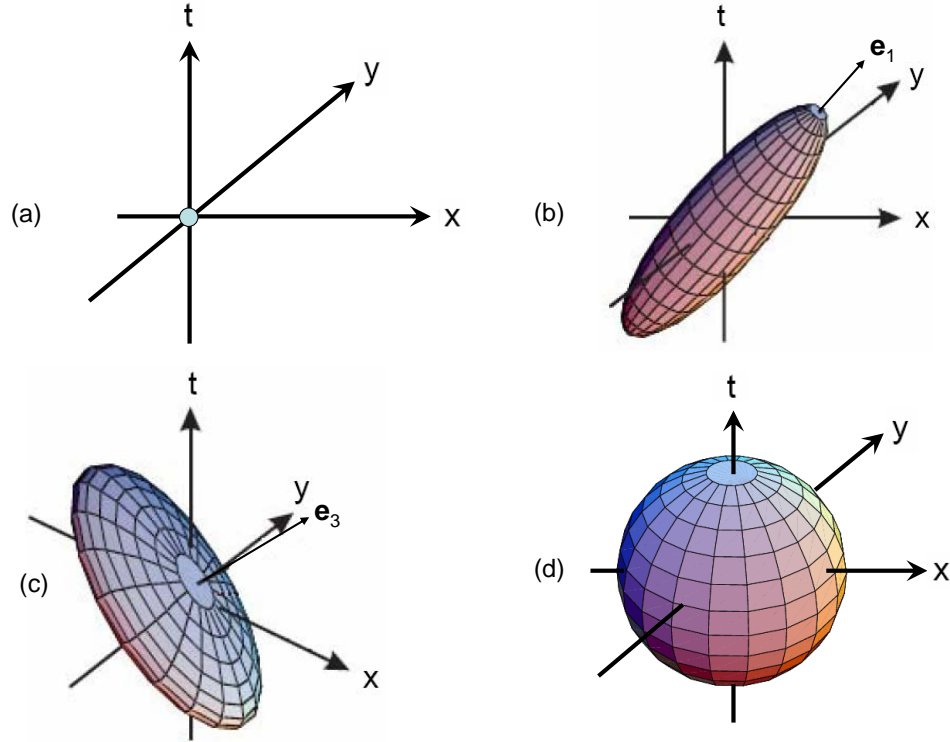
$$\nabla I \cdot \mathbf{u} = 0, \quad (4.21)$$

where  $\nabla I = (I_x, I_y, I_t)^\top$ ,  $\mathbf{u} = (u, v, w)^\top$  and  $\|\mathbf{u}\| = 1$ . Given a particular  $\mathbf{u}$ , Eq. (4.21) constrains the gradient in gradient space  $I_x$ - $I_y$ - $I_t$  to a plane through the origin with normal  $(u, v, w)^\top$ . The structure tensor may be visualized by an ellipsoid (or equivalently a covariance) that measures the dispersion of the points in gradient space. The magnitudes of the (three) eigenvalues of  $\mathbf{S}$  correspond to the length of the principle axes of the ellipsoid. In the ideal case where the points in gradient space lie on a plane, the smallest principle axis will be zero. Thus, a unique normal to the plane exists to describe the solution. In the case where the points in gradient space lie along a line, as occurs when the *aperture problem* is present, there is only one non-zero principle axis in the direction of the line. Thus, rather than a unique normal there are a family of (infinitely) many normals consistent with the linear structure. In the case where the local image structure is homogeneous, the principle axes of the ellipsoid vanish (point structure at origin). Figure 4.6 depicts the ellipsoidal visualization of the structure tensor.

The qualitative analysis of the structure tensor can be considered as a measure of the *intrinsic dimensionality*<sup>4</sup> of the gradient data or the energy spectrum (Krüger & Felsberg, 2003). Zetsche and Barth (Zetsche & Barth, 1991) proposed an operator that exclusively responds to intrinsically three-dimensional spatiotemporal image structures which they suggest to be a reliable indicator of motion discontinuities. The approach relies on regarding the spatiotemporal image as a hypersurface  $\mathbf{H}(x, y, t)$  in four-dimensional space, in the form of a Monge patch  $\mathbf{H}(x, y, t) = (x, y, t, I(x, y, t))$ , and uses the three-dimensional curvature of the hypersurface to identify the presence of an intrinsically three-dimensional signal.

---

<sup>4</sup>A data set in  $n$  dimensions is said to have an intrinsic dimensionality equal to  $n'$  if the data lies entirely within  $n'$ -dimensional space (Bishop, 1995).

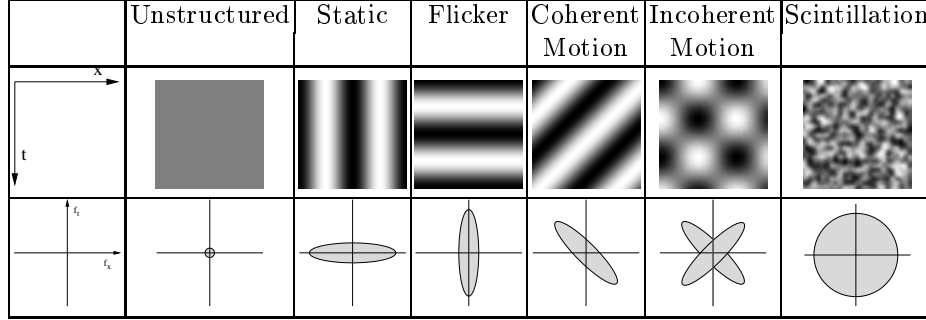


**Figure 4.6.** Principle axes classification of structure tensor. (a) plot of gradient space  $I_x-I_y-I_t$  of a region in space-time that exhibits a negligible amount of variation in all directions. This is indicative a region devoid of any structure (i.e., homogeneous). (b) plot of gradient space representing a region where significant change in only one direction,  $e_1$ , is present. This is indicative of a region where the aperture problem is present. (c) plot of gradient space representing a region where the gray values exhibit significant variation in two orthogonal directions and are relatively constant in the remaining orthogonal direction,  $e_3$ . This is indicative of a region where a single motion is present and the full flow can be recovered. (d) plot of gradient space representing a region where there is significant variation in all directions. This indicative of non-coherent motion (e.g., motion discontinuity).

### 4.5.2 Spectral analysis-based approaches

In this section, approaches that leverage the spectral analysis for single (discussed in Chapter 2.2.2) and multiple motions (discussed in Chapter 2.2.5) are reviewed.

Wildes and Bergen ([Wildes & Bergen, 2000](#)) propose a small set of primitive categories that qualitatively describe the orientation (or lack thereof) of a local region in space-time based on an oriented energy representation. The authors limit the analysis to one spatial dimension and time, thus constant velocity ideally manifests in the energy spectrum as a line through the origin. Note that in addition to the use of oriented energy filters for optical flow estimation (see Chapter 3.3), oriented energy filters have been utilized in the context of spatial texture analysis (e.g., ([Knutsson & Granlund, 1983](#); [Bergen & Adelson, 1986](#); [Bovik et al., 1990](#); [Landy & Bergen,](#)



**Figure 4.7.** Primitive spatiotemporal patterns. The top row depicts prototypical structural patterns that comprise the qualitative categorizations proposed in (Wildes & Bergen, 2000). The bottom row depicts the distribution of energy in the frequency domain for the corresponding categories. Reprinted from (Wildes & Bergen, 2000) with kind permission of Springer Science and Business Media.

1991)). The motion classes considered are as follows:

**Unstructured:** no discernible orientation, the energy is concentrated at the origin

**Static:** the energy resides along the spatial axis

**Flicker:** the energy resides along the temporal axis

**Coherent:** the energy resides along a line through the origin oriented approximately  $45^\circ$

**Incoherent:** the energy resides along several oriented lines through the origin

**Scintillation:** no discernible orientation but the energy is not concentrated at the origin.

Each of these categories have distinctive prototypical patterns of distribution of oriented energy in the frequency domain (see Fig. 4.7). Results, though limited, on natural image examples demonstrate a correlation between qualitative descriptors and spatiotemporal patches. Wildes and Bergen’s approach can be seen as alternative to the structure tensor approach for quantifying the intrinsic dimensionality of space-time images; for a discussion of the structure tensor approach and its relationship to quantifying the intrinsic dimensionality of space-time images (Section 4.5.1). Furthermore, the approach offers finer grained distinctions of patterns of the same intrinsic dimensionality through the consideration of absolute orientation (e.g., flicker versus coherent) and the distribution of energy (e.g., incoherent versus scintillation).

Chomat and Crowley (Chomat & Crowley, 1999) sample the energy spectrum using a set of multi-scale oriented energy filters (12 motion energy receptive fields, 4 orientations at 3 scales) and map the energy outputs to high-level interpretations,

specifically, human movements. These categorizations are based on joint statistics of the spatiotemporal filter responses “learnt” from a set of training image sequences.

Yu et al. (Yu et al., 1999; Yu et al., 2003) propose an approach for detecting (locally) the presence of *occlusion* and *transparency* motion, and estimating the velocity of the constituent layers. The approach proceeds in three stages:

1. Determine the number of motions present.
2. Estimate the motion of the layers.
3. Classify the motion type, occlusion or transparency, and extract layers.

Assuming that there are at most two motions present, the first stage samples the local energy spectrum uniformly in all directions (parameterized by angles  $\theta$  and  $\phi$ ). The presence of two motions is indicated by a signature in the distribution of the energy. The second stage utilizes an EM-based algorithm in the local energy spectrum to simultaneously estimate the normals of the respective dominant planes. The authors treat the cases of transparency and occlusion in a uniform fashion by assuming that two dominant planes are present and the additional distortion term in the case of occlusion is negligible. In the final stage, the spatiotemporal organization of the local image patch is considered for the purposes of classifying the patch as either occlusion or transparency.

Langer and Mann (Langer & Mann, 2001; Langer & Mann, 2003) focus on a special case of *optical snow* where the observer (camera) moves parallel to a static scene (see Chapter 2.2.5 for details on the definition of optical snow). In this case, the authors demonstrate, by specializing the motion field equations (2.9) to parallel translation exclusively, the set of image velocities of the scene  $\{\mathbf{v}_i\}$  have a common direction  $(\tau_x, \tau_y)$  and the speeds vary inversely with depth, formally,

$$\mathbf{v}_i = \alpha_i(\tau_x, \tau_y) \quad \text{where } \alpha_i \in \mathbb{R} \quad (4.22)$$

Langer and Mann formulate an optimization problem within the local Fourier domain to measure the direction and range of speeds that parameterize the “bowtie” structure.

# Chapter 5

## Open problems

THIS chapter concludes this paper with a discussion of a number of open problems in motion analysis.

### 5.1 Modeling theory

Much of the research in motion analysis has been driven from the motion field. In contrast, the study of the spatiotemporal image sequence structure is still in its infancy stages with most research centred around incremental extensions to the structure tensor (Chapter 4.5.1). A potential starting point for future research is the work of Wildes and Bergen ([Wildes & Bergen, 2000](#)) (Chapter 4.5.2). In their work operations were limited to a single spatiotemporal scale. As discussed by Wildes and Bergen, multiscale extensions to this approach may yield finer distinctions between their proposed prototypical categories of structures by way of signatures that may manifest across scales. Note that there are two potential scale extensions to be explored. The first scale extension is concerned with the frequency tuning used for sampling the local spectrum. The second, is concerned with varying the (spatiotemporal) region of aggregation. More generally, little work has explored the role of scale in regional analysis of visual space-time.

The problem of motion boundaries has received periodic attention in the literature, however to date it still remains an open problem. Early analyses of motion boundaries largely centered around differentiating the recovered flow field (see Chapter 4.4). Several recent approaches have attacked this problem via learning methods. Fleet et al. ([Fleet et al., 2000](#)) propose a model of motion boundaries based on learning a basis of flow fields from an idealized (analytical) generative model. Apostoloff and Fitzgibbon ([Apostoloff & Fitzgibbon, 2005](#)) learn the appearance of T-junctions from hand-labelled data.

In Chapter 4.1, several popular representations of motion in the computer vision

literature were outlined. Interestingly, the human visual system has the remarkable ability to accurately recognize a wide variety of motions. An interesting unresolved question is whether the human visual system relies on a highly general representation of motion, such as the affine representation, while giving up some accuracy in the process or (at least in part) on class specific representations. Models of this process could provide direction for the development of robust machine-based solutions. More generally, the wide variety of computational analyses given Chapters 2-4 could motivate corresponding studies for biological visual systems.

There has been a great deal of work on the statistics of natural images (Ruderman, 1994; Olshausen & Field, 1996; Huang & Mumford, 1999) due to the availability of large databases. In comparison there has been very little study of the statistics of flow fields from natural scenes due to the unavailability of a representative database. Recently, Roth and Black (Roth & Black, 2005) addressed this issue by introducing a database of flow fields recovered from natural sequences. The construction of the database relies on the combination of accurate depth data of natural scenes and movements representative of the imaging scenario. The scene depth information is adapted from the Brown range image database (Lee & Huang, 2000) recovered with a laser range finder, that provides the accurate scene depth for a set of 197 complex indoor and outdoor scenes. The camera movement data is based on the movement of a hand-held or car-mounted video camera. Given the flow database, Roth and Black (Roth & Black, 2005) develop an optical flow recovery algorithm based on flow field priors recovered from the database. They report competitive flow field estimates on the Yosemite sequence (the reader is referred to Chapter 3.6 Table 3.2 for a comparison of results). Potential avenues for future work based on the availability of natural flow field databases, include the development of richer priors for the purpose of regularizing estimation problems, scene motion classification and motion boundary detection.

## 5.2 Estimation

It is important to continue the exploration of alternative estimation frameworks in application to image motion. Such considerations may yield improved accuracy in recovered motion estimates and even alter the way we think about what can be recovered. One potentially interesting avenue to explore are set-based estimation methods. Set-based methods (e.g., interval analysis (Moore, 1966)) combine data under the assumption that the sensor error is bounded and focus on establishing bounds on the maximal parameter error. Set-based methods are useful for modeling measurement errors if the true underlying noise distribution cannot be faithfully modeled. The application of set-based methods has appeared in a variety of literature, such as, symbolic reasoning (Brooks, 1981), robot mapping (Engelson & McDermott, 1992) and localization (Atiya & Hager, 1993; Hager et al., 1993). Interestingly, no such

contribution has appeared in the area of motion analysis. This begs the question: is there a role for set-based methods in the analysis of motion?

## 5.3 Experimental evaluation

Despite the tremendous amount of research dedicated to optical flow recovery, empirical evaluation has been relegated to a small number of image sequences, such as the Yosemite sequence (discussed in Chapter 3.6). In the case of the Yosemite sequence, there has been a progressive improvement in reported results<sup>1</sup> but also a recent trend in diminishing returns (see Chapter 3.6, Table 3.2 and Fig. 3.12). A major problem with judging performance of optical flow approaches based on the Yosemite sequence is that it represents a relatively simple test case that contains very little occlusion, specularities and multiple motions etc. (Barron et al., 1994b); problems that remain to be addressed in a satisfactory manner. So although the field has nearly solved the recovery of motion for the Yosemite sequence, the general problem of optical flow recovery remains to be solved. In addition, the Yosemite sequence is synthetic, which brings into question the degree in which one can extrapolate to real world performance. An obvious solution would be the introduction of a battery of publicly available challenging real benchmark sequences. However, the collection of real image sequences with ground truth is non-trivial. As an example, one of the few real image sequences with ground truth to appear in the literature is the *Marbled-Block* sequence by Otte and Nagel (Otte & Nagel, 1994) (Fig. 5.1). They measured the ground truth based on an accurate three-dimensional model of the scene, a calibrated camera, and knowledge of the trajectory of the camera.

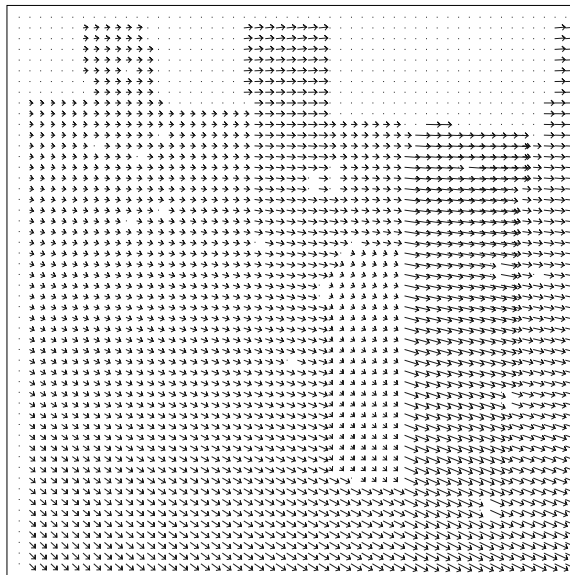
---

<sup>1</sup>It can be argued that the reported improvements with the Yosemite sequence by the most recent methods are due in part to parameter tuning, improvements in filter design and the manner in which results are reported (e.g., not considering the clouds and picking the frame that yields the best results).





(a)



(b)

**Figure 5.1.** Otte and Nagel's *Marbled-Block* sequence (Otte & Nagel, 1994). (a) depicts a frame from Otte and Nagel's *Marbled-Block* sequence and (b) the correct velocity field for this frame.

# Bibliography

- Adelson, E. (1991). Layered representations for motion sequences. Technical report, MIT Media Lab, TR-181.
- Adelson, E. (1995). Layered representations for vision and video. In *IEEE Workshop on Representation of Visual Scenes*.
- Adelson, E. & Anandan, P. (1990). Ordinal characteristics of transparency. Technical report, MIT Media Lab, TR-150.
- Adelson, E. & Bergen, J. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America - A*, 2(2), 284–299.
- Adelson, E. & Bergen, J. (1986). The extraction of spatio-temporal energy in human and machine vision. In *IEEE Workshop on Motion: Representation and Analysis* (pp. 151–155).
- Adelson, E. & Bergen, J. (1991). The plenoptic function and the elements of early vision. In M. Landy & J. Movshon (Eds.), *Computational Models of Visual Processing* (pp. 3–20). Cambridge, MA: MIT Press.
- Adelson, E. & Movshon, J. (1982). Phenomenal coherence of moving visual patterns. *Nature*, 300, 523–525.
- Adiv, G. (1983). Recovering motion parameters in scenes containing multiple moving objects. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 399–400).
- Aggarwal, J. (1986). Motion and time-varying imagery: An overview. In *Workshop on Motion: Representation and Analysis* (pp. 1–6).
- Aggarwal, J. & Badler, N. (Eds.). (1979). *Abstracts for the Workshop on Computer Analysis of Time-Varying Imagery*, University of Pennsylvania, Moore School of Electrical Engineering, Philadelphia, PA.

- Aggarwal, J. & Nandhakumar, N. (1988). On the computation of motion from sequences of images: A review. *Proceedings of the IEEE*, 76, 917–935.
- Anandan, P. (1989). A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2(3), 283–310.
- Andrews, R. & Lovell, B. (2003). Color optical flow. In *Workshop on Digital Image Computing* (pp. 135–139).
- Apostoloff, N. & Fitzgibbon, A. (2005). Learning spatiotemporal t-junctions for occlusion detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. II: 553–559).
- Aris, R. (1989). *Vectors, Tensors, and the Basic Equations of Fluid Mechanics*. New York: Dover Publications.
- Arking, A., Lo, R. & Rosenfeld, A. (1978). A Fourier approach to cloud motion estimation. *Journal of Applied Meteorology*, 17, 735–744.
- Arnold, V. (1991). *The theory of singularities and its applications*. New York: Cambridge University Press.
- Arnsperg, J. (1988). Optic acceleration. In *IEEE International Conference on Computer Vision* (pp. 364–373).
- Aschwandten, P. & Guggenbuhl, W. (1993). Experimental results from a comparative study on correlation-type registration algorithms. In Förstner & Ruwiedel (Eds.), *Robust Computer Vision* (pp. 268–289). Wickmann.
- Atiya, S. & Hager, G. (1993). Real-time vision-based robot localization. *IEEE Transactions on Robotics and Automation*, 9, 785–800.
- Ayer, S. & Sawhney, H. (1995). Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding. In *IEEE International Conference on Computer Vision* (pp. 777–784).
- Bab-Hadiashar, A. & Suter, D. (1996). Robust optic flow estimation using least median of squares. In *IEEE International Conference on Image Processing* (pp. 513–516).
- Bainbridge-Smith, A. & Lane, R. (1997). Determining optical-flow using a differential method. *Image and Vision Computing*, 15(1), 11–22.
- Balanza, M. & Cortelazzo, G. (1989). Frequency domain interpretation of accelerated translations. In *IEEE Multidimensional Signal Processing Workshop* (pp. 36–37).

- Bar-Shalom, Y. & Li, X. (1993). *Estimation and Tracking: Principles, Techniques and Software*. Artech House.
- Barnard, S. & Thompson, W. (1980). Disparity analysis of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(4), 333–340.
- Barnea, D. & Silverman, H. (1972). A class of algorithms for fast digital image registration. *IEEE Transactions on Computers*, 21(2), 179–186.
- Barron, J., Beauchemin, S. & Fleet, D. (1994a). On optical flow. In *International Conference on Artificial Intelligence and Information-Control Systems of Robots* (pp. 3–14).
- Barron, J., Fleet, D. & Beauchemin, S. (1994b). Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1), 43–77.
- Barron, J. & Klette, R. (2002). Quantitative color optical flow. In *IEEE International Conference on Pattern Recognition*.
- Beauchemin, S. & Barron, J. (1994). On optical flow. In *International Conference on Artificial Intelligence and Information-Control Systems of Robots* (pp. 377–386).
- Beauchemin, S. & Barron, J. (1995). The computation of optical-flow. *ACM Computing Surveys*, 27(3), 433–467.
- Beauchemin, S. & Barron, J. (2000a). The frequency structure of 1D occluding image signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(2), 200–206.
- Beauchemin, S. & Barron, J. (2000b). On the Fourier properties of discontinuous motion. *Journal of Mathematical Imaging and Vision*, 13(3), 155–172.
- Beaudet, P. (1978). Rotationally invariant image operators. In *International Conference on Pattern Recognition* (pp. 579–583).
- Bergen, J. & Adelson, E. (1986). Visual texture segmentation based on energy measures. *Journal of the Optical Society of America*, 3, 98.
- Bergen, J., Anandan, P., Hanna, K. & Hingorani, R. (1992). Hierarchical model-based motion estimation. In *European Conference on Computer Vision* (pp. 237–252).
- Bergen, J., Burt, P., Hanna, K., Hingorani, R., Jeanne, P. & Peleg, S. (1991). Dynamic multiple-motion computation. In *Artificial Intelligence and Computer Vision: Proceedings of the Israeli Conference* (pp. 147–156).

- Bergen, J., Burt, P., Hingorani, R. & Peleg, S. (1990). Computing two motions from three frames. In *IEEE International Conference on Computer Vision* (pp. 27–32).
- Bertero, M., Poggio, T. & Torre, V. (1988). Ill-posed problems in early vision. *Proceedings for the IEEE*, 76(8), 869–889.
- Bigün, J., Granlund, G. & Wiklund, J. (1991). Multidimensional orientation estimation with applications to texture analysis and optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8), 775–790.
- del Bimbo, A., Nesi, P. & Sanz, J. (1996). Optical flow computation using extended constraints. *IEEE Transactions on Image Processing*, 5(5), 720–739.
- Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.
- Black, M. & Anandan, P. (1993). A framework for the robust estimation of optical flow. In *IEEE International Conference on Computer Vision* (pp. 231–236).
- Black, M. & Anandan, P. (1996). The robust estimation of multiple motions: Parametric and piecewise-smooth flow-fields. *Computer Vision and Image Understanding*, 63(1), 75–104.
- Black, M. & Jepson, A. (1998). EigenTracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1), 63–84.
- Black, M. & Yacoob, Y. (1997). Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1), 23–48.
- Bober, M. & Kittler, J. (1994). Estimation of complex multimodal motion: An approach based on robust statistics and Hough transform. *Image and Vision Computing*, 12(10), 661–668.
- Bolles, R. & Baker, H. (1985). Epipolar-plane image analysis: A technique for analyzing sequences. In *Image Understanding Workshop* (pp. 137–148).
- Bolles, R., Baker, H. & Marimont, D. (1987). Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1), 7–56.
- Bouthemy, P. & Fablet, R. (1998). Motion characterization from temporal cooccurrences of local motion-based measures for video indexing. In *IEEE International Conference on Pattern Recognition* (pp. Vol I: 905–908).

- Bovik, A., Clark, M. & Geisler, W. (1990). Multichannel texture analysis using localized spatial filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1), 55–73.
- Brooks, R. (1981). Symbolic reasoning among 3-D models and 2-D images. *Artificial Intelligence*, 17(1-3), 285–348.
- Brown, M., Burschka, D. & Hager, G. (2003). Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8), 993–1008.
- Bruhn, A., Weickert, J. & Schnorr, C. (2005). Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61(3), 211–231.
- Burt, P. (1981). Fast filter transforms for image processing. *Computer Graphics and Image Processing*, 16(1), 20–51.
- Burt, P. (1991). Image motion analysis made simple and fast, one component at a time. In *British Machine Vision Conference* (pp. 1–8).
- Burt, P. & Adelson, E. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4), 532–540.
- Burt, P., Bergen, J., Hingorani, R., Kolczynski, R., Lee, W., Leung, A., Lubin, J. & Shvaytser, H. (1989). Object tracking with a moving camera. In *IEEE Workshop on Motion and Video Computing* (pp. 2–12).
- Burt, P., Hingorani, R. & Kolczynski, R. (1991). Mechanisms for isolating component patterns in the sequential analysis of multiple motion. In *Motion Workshop* (pp. 187–193).
- Burt, P., Yen, C. & Xu, X. (1982). Local correlation measures for motion analysis: A comparative study. In *IEEE Conference on Pattern Recognition and Image Processing* (pp. 269–274).
- Burt, P., Yen, C. & Xu, X. (1983). Multi-resolution flow - through motion analysis. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 246–252).
- Buxton, B. & Buxton, H. (1984). Computation of optic flow from the motion of edge features in image sequences. *Image and Vision Computing*, 2(2), 59–75.
- Cafforio, C. & Rocca, F. (1976). Methods for measuring small displacements of television images. *IEEE Transactions on Information Theory*, 22(5), 573–579.

- Campani, M. & Verri, A. (1992). Motion analysis from first-order properties of optical flow. *Computer Vision Graphics and Image Processing*, 56(1), 90–107.
- Camus, T. (1995). Real-time quantized optical flow. In *IEEE Workshop on Computer Architectures for Machine Perception* (pp. 126–131).
- Carneiro, G. & Jepson, A. (2003). Multi-scale phase-based local features. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. I: 736–743).
- Carneiro, G. & Jepson, A. (2005). The distinctiveness, detectability, and robustness of local image features. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. II: 296–301).
- de Castro, E. & Morandi, C. (1987). Registration of translated and rotated images using finite Fourier transforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5), 700–703.
- Chen, W., Giannakis, G. & Nandhakumar, N. (1996). Spatiotemporal approach for time-varying global image motion estimation. *IEEE Transactions on Image Processing*, 5, 1448–1461.
- Chomat, O. & Crowley, J. (1999). Probabilistic recognition of activity using local appearance. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. II: 104–109).
- Chu, C. & Delp, E. (1989). Robust computation of optical-flow in a multiscale differential framework. *Journal of the Optical Society of America - A*, 6(6), 871–878.
- Clocksini, W. (1980). Perception of surface slant and edge labels from optical flow: A computational approach. *Perception* (pp. 253–269).
- Cohen, I. & Herlin, I. (1996). Optical flow and phase portrait methods for environmental satellite image sequences. In *European Conference on Computer Vision* (pp. II:141–150).
- Cohen, I. & Herlin, I. (1999). Non-uniform multiresolution method for optical flow and phase portrait models: Environmental applications. *International Journal of Computer Vision*, 33(1), 29–49.
- Connors, R. & Harlow, C. (1980). A theoretical comparison of texture algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(3), 204–222.
- Cornelius, N. & Kanade, T. (1983). Adapting optical-flow to measure object motion in reflectance and x-ray image sequences. In *Image Understanding Workshop* (pp. 257–265).

- Cremers, D. & Yuille, A. (2003). A generative model based approach to motion segmentation. In *German Pattern Recognition Symposium* (pp. 313–320).
- Crowley, J. & Stern, R. (1984). Fast computation of the difference of low-pass transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2), 212–222.
- Daniilidis, K., Makadia, A. & Bulow, T. (2002). Image processing in catadioptric planes: spatiotemporal derivatives and optical flow computation. In *Workshop on Omnidirectional Vision* (pp. 3–10).
- Daniilidis, K. & Spetsakis, M. (1997). Understanding noise sensitivity in structure from motion. In Y. Aloimonos (Ed.), *Visual Navigation: From Biological Systems to Unmanned Ground Vehicles* (pp. 60–88). Lawrence Erlbaum: Hillsdale, NJ.
- Darrell, T. & Pentland, A. (1995). Cooperative robust estimation using layers of support. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5), 474–487.
- Daugman, J. (1980). Two-dimensional spectral analysis of cortical receptive field profile. *Vision Research*, 20(10), 847–856.
- Dempster, A., Laird, N. & Rubin, D. (1977). Maximal likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39, 185–197.
- Derpanis, K., Wildes, R. & Tsotsos, J. (2004). Hand gesture recognition within a linguistics-based framework. In *European Conference on Computer Vision* (pp. Vol I: 282–296).
- Derrington, A., Harriet, A. & Delicato, L. (2004). Visual mechanisms of motion analysis and motion perception. *Annual Review of Psychology*, 55, 181–205.
- Doretto, G., Chiuso, A., Wu, Y. & Soatto, S. (2003). Dynamic textures. *International Journal of Computer Vision*, 51(2), 91–109.
- DuFaux, F. & Moscheni, F. (1995). Motion estimation techniques for digital TV: A review and a new contribution. *Proceedings of the IEEE*, 83(6), 858–876.
- Duncan, J. & Chou, T. (1992). On the detection of motion and the computation of optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(3), 346–352.
- Egnal, G., Mintz, M. & Wildes, R. (2004). A stereo confidence metric using single view imagery with comparison to five alternative approaches. *Image and Vision Computing*, 22(12), 943–957.



- Elgammal, A., Harwood, D. & Davis, L. (2000). Non-parametric model for background subtraction. In *European Conference on Computer Vision* (pp. II: 751–767).
- Emerson, R., Adelson, E. & Bergen, J. (1992). Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Research*, 32, 203–219.
- Engelson, S. & McDermott, D. (1992). Error correction in mobile robot map learning. In *IEEE International Conference on Robotic Automation* (pp. 2555–2560).
- Enkelmann, W. (1988). Investigation of multigrid algorithms for the estimation of optical flow fields in image sequences. *Computer Vision Graphics and Image Processing*, 43(2), 150–177.
- Fahle, M. & Poggio, T. (1981). Visual hyperacuity: spatio-temporal interpolation in human vision. *Proceedings of the Royal Society of London - B*, 213, 451–477.
- Farneback, G. (2001). Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field. In *IEEE International Conference on Computer Vision* (pp. I: 171–177).
- Faugeras, O. (1990). On the motion of 3D curves and its relationship to optical flow. In *European Conference on Computer Vision* (pp. 107–117).
- Förstner, W. & Güllich, E. (1987). A fast operator for detection and precise location of distinct points, corners and centers of circular features. In *ISPRS Intercommission Workshop* (pp. 149–155).
- Fennema, C. & Thompson, W. (1979). Velocity determination in scenes containing several moving objects. *Computer Graphics Image Processing*, 9(4), 301–315.
- Fermüller, C. (1993). Global 3-D motion estimation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 415–421).
- Fleet, D. (1992). *Measurement of Image Velocity*. Kluwer.
- Fleet, D., Black, M. & Jepson, A. (1998). Motion feature detection using steerable flow fields. In *IEEE International Conference on Pattern Recognition* (pp. 274–281).
- Fleet, D., Black, M., Yacoob, Y. & Jepson, A. (2000). Design and use of linear models for image motion analysis. *International Journal of Computer Vision*, 36(3), 171–193.

- Fleet, D. & Jepson, A. (1989). Computation of normal velocity from local phase information. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 379–386).
- Fleet, D. & Jepson, A. (1990). Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1), 77–104.
- Fleet, D. & Jepson, A. (1993). Stability of phase information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(12), 1253–1268.
- Fleet, D. & Langley, K. (1994). Computational analysis of non-Fourier motion. *Vision Research*, 34(22), 3057–3079.
- Fleet, D. & Weiss, Y. (2005). Optical flow estimation. In N. Paragios, Y. Chen & O. Faugeras (Eds.), *Handbook of Mathematical Methods in Computer Vision* (pp. 241–260). Springer.
- Francois, E. & Bouthemy, P. (1993). Motion segmentation and qualitative dynamic scene analysis from an image sequence. *International Journal of Computer Vision*, 10(2), 157–182.
- Freeman, W. & Adelson, E. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9), 891–906.
- Fuh, C.-S. & Maragos, P. (1991). Affine models for image matching and motion detection. In *International Conference on Acoustics, Speech, and Signal Processing* (pp. 2409–2412).
- Fujita, T. (1969). Present status of cloud velocity computations from the ATS I and ATS III satellites. In *COSPAR Space Research IX* (pp. 557–570).
- Fusiello, A., Roberto, V. & Trucco, E. (1997). Efficient stereo with multiple windowing. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 858–863).
- Gabor, D. (1946). Theory of communication. *Journal of Institute of Electrical Engineers*, 93(3), 429–459.
- Galvin, B., McCane, B., Novins, K., Mason, D. & Mills, S. (1998). Recovering motion fields: An evaluation of eight optical flow algorithms. In *British Machine Vision Conference* (pp. 1563–1567).
- Gennert, M. & Negahdaripour, S. (1987). Relaxing the brightness constancy assumption in computing optical flow. In *MIT AI Memo*.

- Gibson, J. (1950). *The Perception of the Visual World*. New York: Houghton Mifflin.
- Girosi, F., Verri, A. & Torre, V. (1989). Constraints for the computation of optical flow. In *IEEE Proceedings of Visual Motion Workshop* (pp. 116–124).
- Glazer, F. (1981). Computing optic flow. In *International Joint Conference on Artificial Intelligence* (pp. 644–647).
- Glazer, F. (1984). Multilevel relaxation in low-level computer vision. In *Multilevel Image Processing and Analysis* (pp. 312–330).
- Glazer, F. (1987). Hierarchical gradient-based motion detection. In *DARPA Proceedings of Image Understanding* (pp. 733–748).
- Glazer, F., Reynolds, G. & Anandan, P. (1983). Scene matching by hierarchical correlation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 432–441).
- Goldberg, M. & Kourtz, P. (1977). The use of LANDSAT imagery for forestry mapping: Description of a proposed operational system. In *IEEE Workshop on Picture Data Description and Management* (pp. 61–63).
- Golland, P. & Bruckstein, A. (1997). Motion from color. *Computer Vision and Image Understanding*, 68(3), 346–362.
- Gong, S. (1989). Curve motion constraint equations and its applications. In *Workshop on Visual Motion* (pp. 73–80).
- Granlund, G. (1978). In search of a general picture processing operator. *Computer Graphics Image Processing*, 8(2), 155–173.
- Graziano, M., Andersen, R. & Snowden, R. (1994). Tuning of mst neurons to spiral motions. *Journal of Neuroscience*, 14(1), 54–67.
- Guichard, F. & Rudin, L. (1996). Accurate estimation of discontinuous optical flow by minimizing divergence related functionals. In *IEEE International Conference on Image Processing* (pp. 497–499).
- Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13, 49–52.
- Hager, G. & Belhumeur, P. (1996). Real-time tracking of image regions with changes in geometry and illumination. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 403–410).

- Hager, G. & Belhumeur, P. (1998). Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(10), 1025–1039.
- Hager, G., Engelson, S. & Atiya, S. (1993). On comparing statistical and set-based methods in sensor data fusion. In *IEEE International Conference on Robotic automation* (pp. 352–358).
- Hampel, F., Ronchetti, E., Rousseeuw, P. & Stahel, W. (1986). *Robust Statistics: An Approach Based on Influence Functions*. Wiley.
- Haralick, R. (1971). A texture-context feature extraction algorithm for remotely sensed imagery. In *IEEE Decision and Control Conference* (pp. 650–657).
- Haralick, R. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5), 786–804.
- Haralick, R. & Lee, J. (1983). The facet approach to optic flow. In *Proceedings of Image Understanding Workshop* (pp. 84–93).
- Harris, C. & Stephens, M. (1988). A combined corner and edge detector. In *Alvey Vision Conference* (pp. 147–152).
- Haskell, B. (1974). Frame-to-frame coding of television pictures using two-dimensional Fourier transforms. *IEEE Transactions Information Theory*, 20(1), 119–120.
- Haußecker, H. & Fleet, D. (2000). Computing optical flow with physical models of brightness variation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. II: 760–767).
- Haußecker, H. & Fleet, D. (2001). Computing optical flow with physical models of brightness variation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6), 661–673.
- Haußecker, H. & Spies, H. (1999). Motion. In B. Jähne, H. Haußecker & P. Geißler (Eds.), *Handbook of Computer Vision and Applications* (pp. 309–396). Academic Press.
- Haußecker, H., Spies, H. & Jähne, B. (1998). Tensor-based image sequence processing techniques for the study of dynamical processes. In *International Symposium On Real-time Imaging and Dynamic Analysis* (pp. 704–711).
- Heeger, D. (1987). Model for the extraction of image flow. *Journal of the Optical Society of America-A*, 2(2), 1455–1471.

- Heeger, D. (1988). Optical flow from spatiotemporal filters. *International Journal of Computer Vision*, 1(4), 279–302).
- Heeger, D. & Pentland, A. (1986). Seeing structure through chaos. In *Workshop on Motion: Representation and Analysis* (pp. 131–136).
- Heeger, D. & Simoncelli, E. (1992). Model of visual motion sensing. In L. Harris & M. Jenkin (Eds.), *Spatial Vision in Humans and Robots*. Cambridge University Press.
- Helman, J. & Hesselink, L. (1990). Surface representations of two- and three-dimensional fluid flow topology. In *IEEE Conference on Visualization* (pp. 6–13).
- Helman, J. & Hesselink, L. (1991). Visualizing vector field topology in fluid flows. *IEEE Computer Graphics and Applications*, 11(3), 36–46.
- Hildreth, E. (1984). The computation of the velocity field. *Proceedings of the Royal Society of London*, 221, 189–220.
- Hildreth, E. & Koch, C. (1987). The analysis of visual motion: from computational theory to neuronal mechanisms. *Annual Review of Neuroscience*, 10, 477–533.
- Hirsch, M. & Smale, S. (1974). *Differential equations, dynamical systems and linear algebra*. New York: Academic Press.
- Hirschmüller, H., Innocent, P. & Garibaldi, J. (2002). Real-time correlation-based stereo vision with reduced border errors. *International Journal of Computer Vision*, 47(1-3), 229–246.
- Hoey, J. & Little, J. (2000). Representation and recognition of complex human motion. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. I: 752–759).
- Horn, B. (1986). *Robot Vision*. Cambridge, MA: MIT Press.
- Horn, B. & Schunck, B. (1981). Determining optical flow. *Artificial Intelligence*, 17(1-3), 185–203.
- Horn, B. & Schunck, B. (1993). Determining optical flow: A retrospective. *Artificial Intelligence*, 59(1-2), 81–87.
- Horn, B. & Weldon, Jr., E. (1988). Direct methods for recovering motion. *International Journal of Computer Vision*, 2(1), 51–76.
- Hough, P. (1962). Method and means for recognizing complex patterns. U.S. Patent 3,069,654.

- Huang, J. & Mumford, D. (1999). Statistics of natural images and models. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. I: 541–547).
- Huang, T. & Netravali, A. (1994). Motion and structure from feature correspondences: A review. *Proceedings for the IEEE*, 82(2), 252–268.
- Huang, T. & Tsai, R. (1981). Image sequence analysis: Motion estimation. In T. Huang (Ed.), *Image Sequence Analysis* (pp. 1–18). Springer.
- Huber, P. (1981). *Robust Statistics*. Wiley.
- Irani, M. & Anandan, P. (1998). Robust multi-sensor image alignment. In *IEEE International Conference on Computer Vision* (pp. 959–966).
- Irani, M., Rousso, B. & Peleg, S. (1994). Computing occluding and transparent motions. *International Journal of Computer Vision*, 12(1), 5–16.
- Isard, M. & Blake, A. (1998). CONDENSATION – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1), 5–28.
- Jähne, B. (1990). Motion determination in space-time images. In *European Conference on Computer Vision* (pp. 161–173).
- Jähne, B. (2005). *Digital Image Processing, sixth edition*. Springer-Verlag.
- Jähne, B., Haußecker, H., Spies, H., Schmundt, D. & Schurr, U. (1998). Study of dynamical processes with tensor-based spatiotemporal image processing techniques. In *European Conference on Computer Vision* (p. II: 322).
- Jepson, A. & Black, M. (1993a). Mixture models for optical flow computation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 760–761).
- Jepson, A. & Black, M. (1993b). Mixture models for optical flow computation. Technical report, University of Toronto, RBCV-TR-93-44.
- Jojic, N. & Frey, B. (2001). Learning flexible sprites in video layers. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. I:199–206).
- Julesz, B. (1962). Visual pattern discrimination. *IRE Transactions on Information Theory*, 8(2), 84–92.
- Julier, S., Uhlmann, J. & Durrant-Whyte, H. (1995). A new approach for filtering nonlinear systems. In *American Control Conference* (pp. 1628–1632).
- Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, 82, 35–45.

- Kass, M. & Witkin, A. (1987). Analyzing oriented patterns. *Computer Vision, Graphics, and Image Processing*, 37(3), 362–385.
- Kearney, J., Thompson, W. & Boley, D. (1987). Optical flow estimation: An error analysis of gradient-based methods with local optimization. *Pattern Analysis and Image Intelligence*, 9(2), 229–244.
- Kitchen, L. & Rosenfeld, A. (1982). Gray level corner detection. *Pattern Recognition Letters*, 1(2), 95–102.
- Knutsson, H. & Granlund, G. (1983). Texture analysis using two-dimensional quadrature filters. In *IEEE Workshop on Computer Architecture for Pattern Analysis and Image Database Management* (pp. II: 768–784).
- Koenderink, J. & van Doorn, A. (1975). Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer. *Optica Acta*, 22(9), 773–791.
- Koenderink, J. & van Doorn, A. (1976). Local structure of movement parallax of the plane. *Journal of the Optical Society of America*, 66(7), 717–723.
- Koller, D., Heinze, H. & Nagel, H. (1991). Algorithmic characterization of vehicle trajectories from image sequences by motion verbs. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 90–95).
- Kollnig, H., Nagel, H. & Otte, M. (1994). Association of motion verbs with vehicle movements extracted from dense optical flow fields. In *European Conference on Computer Vision* (pp. B:338–347).
- Konrad, J. & Dubois, E. (1992). Bayesian estimation of motion vector fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(9), 910–927.
- Korn, G. & Korn, T. (1968). *Mathematical Handbook for Scientists and Engineers*. Mineola, NY: Dover Publications.
- Krüger, N. & Felsberg, M. (2003). A continuous formulation of intrinsic dimension. In *British Machine Vision Conference*.
- Kretzmer, E. (1952). Statistics of television signals. *Bell System Technical Journal*, 31(4), 7551–763.
- Kuglin, C. & Hines, D. (1975). The phase correlation image alignment method. In *IEEE Conference on Cybernetics and Society* (pp. 163–165).

- Lai, S. & Fang, M. (1999). Robust and efficient image alignment with spatially-varying illumination models. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. II: 167–172).
- Landy, M. & Bergen, J. (1991). Texture segregation and orientation gradient. In *Vision Research* (pp. 679–691).
- Langer, M. & Mann, R. (2001). Dimensional analysis of image motion. In *IEEE International Conference on Computer Vision* (pp. I: 155–162).
- Langer, M. & Mann, R. (2003). Optical snow. *International Journal of Computer Vision*, 55(1), 55–71.
- Langley, K., Fleet, D. & Atherton, T. (1992). Multiple motions from instantaneous frequency. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 846–849).
- Lee, A. & Huang, J. (2000). Brown range image database .  
<http://www.dam.brown.edu/ptg/brid/index.html>.
- Lim, J. (1990). *Two-Dimensional Signal and Image Processing*. Prentice-Hall.
- Limb, J. & Murphy, J. (1975a). Estimating velocity of moving images in television signals. *Computer Graphics, Image Processing*, 4(3), 311–327.
- Limb, J. & Murphy, J. (1975b). Measuring the speed of moving objects from television signals. *IEEE Transactions on Communications*, 23(4), 474–478.
- Lin, T. & Barron, J. (1994). Image reconstruction error for optical flow. In *Vision Interface* (pp. 73–80).
- Liou, S. & Jain, R. (1989). Motion detection in spatio-temporal space. *Computer Vision Graphics and Image Processing*, 45(2), 227–250.
- Little, J. & Verri, A. (1989). Analysis of differential and matching methods for optical flow. In *IEEE Workshop on Visual Motion* (pp. 173–180).
- Liu, H., Chellappa, R. & Rosenfeld, A. (2003). Accurate dense optical flow estimation using adaptive structure tensors and a parametric model. *IEEE Transactions on Image Processing*, 12(10), 1170–1180.
- Liu, H., Hong, T., Herman, M., Camus, T. & Chellappa, R. (1996). Accuracy vs. efficiency trade-offs in optical flow algorithms. In *European Conference on Computer Vision* (pp. II:174–183).



- Liu, H., Hong, T., Herman, M., Camus, T. & Chellappa, R. (1998). Accuracy vs. efficiency trade-offs in optical flow algorithms. *Computer Vision and Image Understanding*, 72(3), 271–286.
- Longuet-Higgins, H. & Prazdny, K. (1980). The interpretation of a moving retinal image. *Proceedings of the Royal Society of London - B*, 208, 385–397.
- Lowe, D. (1999). Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision* (pp. 1150–1157).
- Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110.
- Lu, Z., Xie, W., Pei, J. & Huang, J. (2005). Dynamic texture recognition by spatio-temporal multiresolution histograms. In *Workshop on Motion* (pp. II: 241–246).
- Lucas, B. & Kanade, T. (1981a). An iterative image registration technique with an application to stereo vision. In *DARPA Image Understanding Workshop* (pp. 121–130).
- Lucas, B. & Kanade, T. (1981b). An iterative registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligences* (pp. 674–679).
- Malik, J. & Perona, P. (1990). Preattentive texture discrimination with early vision mechanism. *Journal of the Optical Society of America-A*, 7(5), 923–932.
- Mann, R. & Langer, M. (2005). Spectrum analysis of motion parallax in a 3d cluttered scene and application to egomotion. *Journal of the Optical Society of America-A*, 22(9), 1717–1731.
- Marburger, H. & Neumann, B. and Novak, H. (1981). Natural language dialogue about moving objects in an automatically analyzed traffic scene. In *International Joint Conferences on Artificial Intelligence* (pp. 49–51).
- Markandey, F. & Flinchbaugh, B. (1990). Multispectral constraints for optical flow computation. In *IEEE International Conference on Computer Vision* (pp. 38–41).
- Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman.
- Marr, D. & Ullman, S. (1979). Directional selectivity and its use in early visual processing. In *MIT AI Memo*.

- Martinez-Trujillo, J.C., Tsotsos, J., Simine, E., Pomplun, M., Wildes, R., Treue, S., Heinze, H.-J. & Hopf, J.-M. (2005). On benchmarking optical flow. *NeuroReport*, 16(5), 205–219.
- McCane, B., Novins, K., Crannitch, D. & Galvin, B. (2001). On benchmarking optical flow. *Computer Vision and Image Understanding*, 84(1), 126–143.
- McCurdy, P. (Ed.). (1944). *Manual of Photogrammetry* (first Ed.). New York: Pitman Publishing Corporation.
- Meer, P. (2004). Robust techniques for computer vision. In G. Medioni & S. Kang (Eds.), *Emerging Topics in Computer Vision* chapter 4. Prentice Hall.
- Meer, P., Stewart, C. & Tyler, D. (2000). Robust computer vision: An interdisciplinary challenge. *Computer Vision and Image Understanding*, 78(1), 1–7.
- Memin, E. & Perez, P. (1998). A multigrid approach for hierarchical motion estimation. In *IEEE International Conference on Computer Vision* (pp. 933–938).
- Middendorf, M. & Nagel, H. (2001). Estimation and interpretation of discontinuities in optical flow fields. In *IEEE International Conference on Computer Vision* (pp. I: 178–183).
- Mikolajczyk, K. & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10), 1615–1630.
- Mintz, D. & Meer, P. (1991). Robust estimators in computer vision: An introduction to least median of squares regression. In Y. Feldman & A. Bruckstein (Eds.), *Artificial Intelligence and Computer Vision* (pp. 61–70). Elsevier.
- Mitchell, T. (1997). *Machine Learning*. McGraw Hill.
- Mitiche, A. & Bouthemy, P. (1996). Computation and analysis of image motion: A synopsis of current problems and methods. *International Journal of Computer Vision*, 19(1), 29–55.
- Mitiche, A., Wang, Y. & Aggarwal, J. (1987). Experiments in computing optical flow with the gradient-based, multiconstraint method. *Pattern Recognition*, 20(2), 173–179.
- Moore, R. (1966). *Interval Analysis*. Englewood Cliffs, N.J.: Prentice Hall.
- Moravec, H. (1977). Towards automatic visual obstacle avoidance. In *International Joint Conference on Artificial Intelligence* (p. 584).

- Mota, C., Stuke, I. & Barth, E. (2001). Analytic solutions for multiple motions. In *IEEE International Conference on Image Processing* (pp. II: 917–920).
- Mounts, F. (1969). A video encoding system with conditional picture-element replenishment. *The Bell Systems Technical Journal*, 48(7), 2545–2554.
- Mukawa, N. (1989). Motion field estimation for shaded scenes. In *IEEE International Conference on Image Processing* (pp. 796–800).
- Mulligan, J. (1992). Motion transparency is restricted to two planes. *Investigative Ophthalmology and Visual Science Supplemental (ARVO)*, 33, 1049.
- Muybridge, E. (1887). Animal locomotion. Technical report, University of Pennsylvania.
- Nagel, H. (1978). Analysis techniques for image sequences. In *International Conference on Pattern Recognition* (pp. 186–211).
- Nagel, H. (1981). Image sequence analysis: What can we learn from applications? In *Image Sequence Analysis* (pp. 19–213).
- Nagel, H. (1983a). Constraints for the estimation of displacement vector fields from image sequences. In *International Joint Conference on Artificial Intelligence* (pp. 945–951).
- Nagel, H. (1983b). Displacement vectors derived from second-order intensity variations in image sequences. *Computer Vision Graphics and Image Processing*, 21(1), 85–117.
- Nagel, H. (1986). Image sequences - ten (octal) years - from phenomenology towards a theoretical foundation. In *International Conference on Pattern Recognition* (pp. 1174–1185).
- Nagel, H. (1987). On the estimation of optical flow: Relations between different approaches and some new results. *Artificial Intelligence*, 33(3), 299–324.
- Nagel, H. (1989). On a constraint equation for the estimation of displacement rates in image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(1), 13–30.
- Nagel, H. (1990). Extending the ‘oriented smoothness constraint’ into the temporal domain and the estimation of derivatives of optical flow. In *European Conference on Computer Vision* (pp. 139–148).

- Nagel, H. (1995). Optical-flow estimation and the interaction between measurement errors at adjacent pixel positions. *International Journal of Computer Vision*, 15(3), 271–288.
- Nagel, H. (2000). Image sequence evaluation: 30 years and still going strong. In *IEEE International Conference on Pattern Recognition* (pp. Vol I: 149–158).
- Nagel, H. & Enkelmann, W. (1986). An investigation of smoothness constraints for the estimation of displacement vector fields for image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 565–593.
- Nakayama, K. & Loomis, J. (1974). Optical velocity patterns, velocity sensitive neurons, and space perception: A hypothesis. *Perception* (pp. 63–80).
- Negahdaripour, S. (1992). Motion recovery from image sequences using only first-order optical flow information. *International Journal of Computer Vision*, 9(3), 163–184.
- Negahdaripour, S. & Lee, S. (1991). Motion recovery from image sequences using first-order optical flow information. In *IEEE Workshop on Visual Motion* (pp. 132–139).
- Negahdaripour, S. & Yu, C. (1993). A generalized brightness change model for computing optical flow. In *IEEE International Conference on Computer Vision* (pp. 2–11).
- Nelson, R. & Polana, R. (1992). Qualitative recognition of motion using temporal texture. *Computer Vision, Graphics and Image Processing*, 56(1), 78–89.
- Nesi, P., del Bimbo, A. & Ben-Tzvi, D. (1995). A robust algorithm for optical-flow estimation. *Computer Vision and Image Understanding*, 62(1), 59–68.
- Nestares, O., Fleet, D. & Heeger, D. (2000). Likelihood functions and confidence bounds for total-least-squares problems. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. I: 523–530).
- Netravali, A. & Robbins, J. (1979). Motion-compensated television coding: Part 1. *Bell System Technical Journal*, 58(3), 631–670.
- Odobez, J. & Bouthemy, P. (1995). Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4), 348–365.
- Ohta, N. (1989). Optical flow detection by color images. In *IEEE International Conference on Image Processing* (pp. 801–805).

- Oja, E. (1983). *Subspace Methods of Pattern Recognition*. John Wiley and Sons.
- Okutomi, M. & Kanade, T. (1992). A locally adaptive window for signal matching. *International Journal of Computer Vision*, 7(2), 143–162.
- Okutomi, M., Katayama, Y. & Oka, S. (2002). A simple stereo algorithm to recover precise object boundaries and smooth surfaces. *International Journal of Computer Vision*, 47(1-3), 261–273.
- Oliensis, J. (2002). Exact two-image structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12), 1618–1633.
- Olshausen, B. & Field, D. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(23), 333–339.
- Onoe, M., Hamano, N. & Ohba, K. (1973). Computer analysis of traffic flow observed by subtractive television. *Computer Vision Graphics and Image Processing*, 2, 377–392.
- Oppenheim, A., Willsky, A. & S.H., N. (1997). *Signals and Systems*. Upper Saddle River, NJ: Prentice Hall.
- Otte, M. & Nagel, H. (1994). Optical flow estimation: Advances and comparisons. In *European Conference on Computer Vision* (pp. A:51–60).
- Papoulis, A. & Pillai, S. (2002). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill.
- Peh, C. & Cheong, L. (2002). Synergizing spatial and temporal texture. *IEEE Transactions on Image Processing*, 11(10), 1179–1191.
- Phong, B. (1975). Illumination for computer generated pictures. *Communications of the ACM*, 18(6), 311–317.
- Poggio, T., Torre, V. & Koch, C. (1984). Ill-posed problems and regularization analysis in early vision. In *MIT AI Memo*.
- Poggio, T., Torre, V. & Koch, C. (1985). Computational vision and regularization theory. *Nature*, 317, 314–319.
- Polana, R. & Nelson, R. (1997). Temporal texture and activity recognition. In M. Shah & R. Jain (Eds.), *Motion-based recognition* (pp. 87–115). Kluwer Academic.
- Press, W., Flannery, B., Teukolsky, S. & Vetterling, W. (1992). *Numerical Recipes in C*. Cambridge University Press.

- Rahman, A. & Murshed, M. (2004). Real-time temporal texture characterisation using block based motion co-occurrence statistics. In *IEEE International Conference on Image Processing* (pp. III: 1593–1596).
- Rao, A. & Jain, R. (1990). Analyzing oriented textures through phase portraits. In *International Conference on Pattern Recognition* (pp. Vol-I 336–340).
- Rao, A. & Jain, R. (1992). Computerized flow field analysis: Oriented texture fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(7), 693–709.
- Reichardt, W. (1961). “Autocorrelation, a principle for the evaluation of sensory information by the central nervous system”. New York: in *Sensory Communication*, Wiley.
- Rosenfeld, A. (1984). *Multiresolution Image Processing and Analysis*. Berlin: Springer.
- Rosenfeld, A. & Troy, E. (1970). Visual texture analysis. In *Conference Record for Symposium on Feature Extraction and Selection in Pattern Recognition* (pp. 115–124).
- Roth, S. & Black, M. (2005). On the spatial statistics of optical flow. In *IEEE International Conference on Computer Vision* (pp. 42–49).
- Rousseeuw, P. & Leroy, A. (2003). *Robust Regression and Outlier Detection*. Wiley.
- Ruderman, D. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, 5, 517–548.
- Saisan, P., Doretto, G., Wu, Y. & Soatto, S. (2001). Dynamic texture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. II:58–63).
- van Santen, J. & Sperling, G. (1984). Temporal covariance model of human motion perception. *Journal of the Optical Society of America-A*, 1, 451–473.
- Scharr, H. (2005). Optimal filters for extended optical flow. In *International Workshop on Complex Motion*.
- Schmid, C. & 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. & Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2), 151–172.

- Schnörr, C. (1994). Unique reconstruction of piecewise smooth images by minimizing strictly convex nonquadratic functionals. *Journal of Mathematical Imaging and Vision*, 4, 189–198.
- Schunck, B. (1984). The motion constraint equation for optical flow. In *IEEE International Conference on Pattern Recognition* (pp. 20–22).
- Schunck, B. (1985). Image flow: Fundamentals and future research. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 560–571).
- Schunck, B. (1988). Image flow: Fundamentals and algorithms. In *Motion Understanding: Robot and Human Vision* (pp. 23–80).
- Schunck, B. (1989). Image flow segmentation and estimation by constraint line clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(10), 1010–1027.
- Shizawa, M. & Mase, K. (1990). Simultaneous multiple optical flow estimations. In *International Conference on Pattern Recognition* (pp. Vol-I 274–278).
- Shizawa, M. & Mase, K. (1991a). Principle of superposition: A common computational framework for analysis of multiple motion. In *IEEE Workshop on Motion* (pp. 164–172).
- Shizawa, M. & Mase, K. (1991b). A unified computational theory for motion transparency and motion boundaries based on eigenenergy analysis. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 289–295).
- Shu, C. & Jain, R. (1992). Vector field analysis for oriented patterns. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 673–676).
- Shu, C. & Jain, R. (1994). Vector field analysis for oriented patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9), 946–950.
- Simoncelli, E. (1993a). Coarse-to-fine estimation of visual motion. In *IEEE Workshop on Multidimensional Signal Processing* (pp. 128–129).
- Simoncelli, E. (1993b). Distributed analysis and representation of visual motion. In *MIT Ph.D.*
- Simoncelli, E. (1994). Design of multi-dimensional derivative filters. In *IEEE International Conference on Image Processing* (pp. 790–794).
- Simoncelli, E. (2003). Local analysis of visual motion. In L. Chalupa & J. Werner (Eds.), *The Visual Neurosciences* chapter 109. MIT Press.

- Simoncelli, E., Adelson, E. & Heeger, D. (1991). Probability distributions of optical flow. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 310–315).
- Singh, A. (1990). An estimation-theoretic framework for image-flow computation. In *IEEE International Conference on Computer Vision* (pp. 168–177).
- Singh, A. (1991). Image-flow computation: An estimation-theoretic framework, unification and integration. *Machine Vision and Applications*, 4, 55.
- Sobey, P. & Srinivasan, M. (1991). Measurement of optical flow by a generalized gradient scheme. *Journal of the Optical Society of America-A*, 8(9), 1488–1498.
- Solomon, C. (1994). *The history of animation: Enchanted Drawings*. New York: Wings Books.
- Spetsakis, M. (1994). An optical flow estimation algorithm that uses Gabor filters and affine model for flow. Technical report, Department of Computer Science, York University, CS-94-06.
- Spetsakis, M. (1997). Optical flow estimation using discontinuity conforming filters. *Computer Vision and Image Understanding*, 68(3), 276–289.
- Spies, H. & Jähne, B. (2001). A general framework for image sequence analysis. In *Fachtagung Informationstechnik* (pp. 125–132).
- Spoerri, A. & Ullman, S. (1990). The early detection of motion boundaries. In *MIT AI-TR*.
- Stauffer, C. & Grimson, W. (1999). Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition* (pp. II: 246–252).
- Stewart, C. (1999). Robust parameter estimation in computer vision. *SIAM Review*, 41(3), 513–537.
- Stillér, C. & Konrad, J. (1999). Estimating motion in image sequences: a tutorial on modeling and computation of 2D motion. In *IEEE Signal Processing Magazine* (pp. 70–98).
- Subbarao, M. (1990). Bounds on time-to-collision and rotational component from first-order derivatives of image flow. *Computer Vision, Graphics and Image Processing*, 50(3), 329–341.
- Szumner, M. (1996). Temporal texture sample database .  
<ftp://whitechapel.media.mit.edu/pub/szumner/temporal-texture/raw/>.



- Terzopoulos, D. (1986). Image analysis using multigrid relaxation methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(2), 129–139.
- Thompson, W., Mutch, K. & Berzins, V. (1985). Dynamic occlusion analysis in optical flow fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(4), 374–383.
- Tikhonov, A. (1995). *Numerical Methods for the solution of ill-posed problems*. Dordrecht: Kluwer.
- Tistarelli, M. (1994). Multiple constraints for optical flow. In *European Conference on Computer Vision* (pp. A:61–70).
- Tistarelli, M. (1995). Computation of coherent optical flow by using multiple constraints. In *IEEE International Conference on Computer Vision* (pp. 263–268).
- Tomasi, C. & Shi, J. (1994). Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 593–600).
- Tretiak, O. & Pastor, L. (1984). Velocity estimation from image sequences with second order differential operators. In *IEEE International Conference on Pattern Recognition* (pp. 16–19).
- Tsotsos, J., Mylopoulos, J., Covvey, H. & Zucker, S. (1979). ALVEN: A study on motion understanding by computer. In *International Joint Conference on Artificial Intelligence* (pp. 890–892).
- Tsotsos, J., Mylopoulos, J., Covvey, H. & Zucker, S. (1980). A framework for visual motion understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(6), 563–573.
- Ullman, S. (1979). *The Interpretation of Visual Motion*. MIT Press.
- Uras, S., Girosi, F., Verri, A. & Torre, V. (1989). A computational approach to motion perception. *Biological Cybernetics*, 60, 79–87.
- Van Huffer, S. & Vandewalle, J. (1991). *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM.
- Vega-Riveros, J. & Jabbour, K. (1989). Review of motion analysis techniques. *Communications, Speech and Vision, IEE Proceedings*, 136, 397–404.
- Verri, A., Girosi, F. & Torre, V. (1990). Differential techniques for optical flow. *Journal of the Optical Society of America-A*, 7, 912–922.

- Verri, A. & Poggio, T. (1987a). Against quantitative optical flow. In *IEEE International Conference on Computer Vision* (pp. 171–180).
- Verri, A. & Poggio, T. (1987b). Qualitative information in the optical flow. In *DARPA* (pp. 825–834).
- Verri, A. & Poggio, T. (1989). Motion field and optical flow: Qualitative properties. *Pattern Analysis and Machine Intelligence*, 11(5), 490–498.
- Wang, J. & Adelson, E. (1993). Layered representation for motion analysis. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 361–366).
- Wang, J. & Adelson, E. (1994). Representing moving images with layers. *IEEE Transactions on Image Processing Special Issue: Image Sequence Compression*, 3(5), 625–638.
- Wang, S., Markandey, V. & Reid, A. (1992). Total least squares fitting spatiotemporal derivatives to smooth optical flow fields. In *Proceedings of the SPIE: Signal and Data Processing of Small Targets* (pp. 42–55).
- Watson, A. & Ahumada, A. (1985). Model of human visual-motion sensing. *Journal of the Optical Society of America-A*, 2(2), 322–342.
- Waxman, A. & Ullman, S. (1985). Surface structure and three-dimensional motion from image flow kinematics. *International Journal of Robotics Research*, 4(3), 72–94.
- Waxman, A., Wu, J. & Bergholm, F. (1988). Convected activation profiles and the measurement of visual motion. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 717–723).
- Weber, J. & Malik, J. (1995). Robust computation of optical-flow in a multiscale differential framework. *International Journal of Computer Vision*, 14(1), 67–81.
- Weickert, J., Bruhn, A., Papenberg, N. & Brox, T. (2003). Variational optic flow computation: From continuous models to algorithms. In *International Workshop on Computer Vision and Image Analysis*.
- Weickert, J. & Schnörr, C. (2001a). A theoretical framework for convex regularizers in pde-based computation of image motion. *International Journal of Computer Vision*, 45(3), 245–264.
- Weickert, J. & Schnörr, C. (2001b). Variational optic flow computation with a spatio-temporal smoothness constraint. *Journal of Mathematical Imaging and Vision*, 14, 245–255.

- van de Weijer, J. & Gevers, T. (2004). Robust optical flow from photometric invariants. In *IEEE International Conference on Image Processing* (pp. III: 1835–1838).
- Weinstock, R. (1974). *Calculus of Variations: With Applications to Physics and Engineering*. Dover.
- Weiss, Y. (1997). Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 520–526).
- Weiss, Y. & Adelson, E. (1996). A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 321–326).
- Weiss, Y. & Adelson, E. (1998). Slow and smooth: A Bayesian theory for the combination of local motion signals in human vision. Technical report, MIT Artificial Intelligence Laboratory, TR-1624.
- Weiss, Y. & Fleet, D. (2002). Velocity likelihoods in biological and machine vision. In R. Rao, B. Olshausen & M. Lewicki (Eds.), *Probabilistic Models of the Brain: Perception and Neural Function* (pp. 77–96). MIT Press.
- Wildes, R. (1993). On the qualitative structure of temporally evolving visual motion fields. In *Conference on Artificial Intelligence* (pp. 844–850).
- Wildes, R., Amabile, M., Lanzillotto, A. & Leu, T. (1997). Experiments with an algorithm for recovering fluid flow from video imagery. In *DARPA* (pp. 185–192).
- Wildes, R., Amabile, M., Lanzillotto, A. & Leu, T. (2000). Recovering estimates of fluid flow from image sequence data. *Computer Vision and Image Understanding*, 80(2), 246–266.
- Wildes, R. & Bergen, J. (2000). Qualitative spatiotemporal analysis using an oriented energy representation. In *European Conference on Computer Vision* (pp. II: 768–784).
- Willick, D. & Yang, Y. (1991). Experimental evaluation of motion constraint equations. *Computer Vision Graphics and Image Processing*, 54(2), 206–214.
- Wohn, K., Davis, L. & Thrift, P. (1983). Motion estimation based on multiple local constraints and nonlinear smoothings. *Pattern Recognition*, 16(6).

- Wong, R. & Hall, E. (1977). Hierarchical search for image matching. In *IEEE Conference on Decision and Control*.
- Wong, K, Y., Ye, L. & Spetsakis, M. (2004). EM clustering of incomplete data applied to motion segmentation. In *British Machine Vision Conference* (pp. 237–246).
- Woodham, R. (1990). Multiple light source optical flow. In *IEEE International Conference on Computer Vision* (pp. 42–46).
- Yachida, M. (1981). Determining velocity map by 3-D iterative estimator. In *International Joint Conference on Artificial Intelligence* (pp. 716–718).
- Yagi, Y. (1999). Omnidirectional sensing and its applications. *IEICE Transactions on Information and Systems*, *E82-D(3)*, 568–579.
- Yu, W., Daniilidis, K., Beauchemin, S. & Sommer, G. (1999). Detection and characterization of multiple motion points. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. I: 171–177).
- Yu, W., Sommer, G., Beauchemin, S. & Daniilidis, K. (2002). Oriented structure of the occlusion distortion: Is it reliable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24(9)*, 1286–1290.
- Yu, W., Sommer, G. & Daniilidis, K. (2003). Multiple motion analysis: in spatial or in spectral domain? *Computer Vision and Image Understanding*, *90(2)*, 129–152.
- Zabih, R. & Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *European Conference on Computer Vision* (pp. II: 151–158).
- Zernike, F. (1934). Beugungstheorie des schneidenverfahrens und seiner verbesserten form, der phasenkontrastmethode. *Physica*, *1*, 689–704.
- Zetzsche, C. & Barth, E. (1991). Direct detection of flow discontinuities by 3D curvature operators. *Pattern Recognition Letters*, *12*, 771–779.
- Zhang, J. & Nagel, H. (1994). Texture-based segmentation of road images. In *IEEE Intelligent Vehicles Symposium* (pp. 260–265).