ECOVISION

Deliverable 3.1

Norbert Krüger and Florentin Wörgötter
Department of Psychology
University of Stirling

Markus Lappe
Psychologisches Institut II
Westf. Wilhelms-Universität of Münster

Silvio P. Sabatini, Fabio Solari and G.M. Bisio
Department of Biophysical and Electronic Engineering
University of Genoa

Remark: This deliverable 3.1 contains also aspects of the deliverables 3.2 and 3.3 that are due at later stages of the project. Since these aspects touch 3.1 as well as 3.2 and 3.3 we have included them already here to support understanding.

Contents

Chapter 1

Introduction

In this report we address the problem of defining and detecting spatial Gestalts in dynamic visual scenes. A Gestalt can be defined as a group of pixels characterized by strong interdependencies in the feature space. Given that, different approaches can be considered to detect a Gestalt, according to the higher-level properties they are supposed to represent, e.g., being orientation-based, motion-based, stereo-based, color-based.

Cortical visual pathway adopts two parallel pathways: the parvocellular pathway and the magnocellular pathway, corresponding to the "what" and "where" streams, respectively. Both rely on elementary information such as orientation, spatial frequency, velocity, binocular disparity, and color. While the magno system is sensitive primarily to moving objects and carries information about the overall organization of the visual world, the parvo system seems to be important for analyzing the scene in much greater detail. More specifically, electro-physiological studies, suggest that the magno system is responsible for carrying information about movement and depth, and, more generally, may have a global function of interpreting spatial organization. Magno functions may include deciding which visual elements, such as edges and discontinuities, belong to and define individual objects in the scene, as well as determine the overall 3-D organization of the scene, and the positions of objects in space and movements of objects. The parvo system, which is well developed only in primates, seems to have added the ability to scrutinize in much more detail the shape, color, and surface properties of objects, creating the possibility of assigning multiple visual attributes to a single object, and correlating its parts.

In this perspective, we conceived two complementary approaches. The first (see part A), mainly intensity-based (cf. parvo system), detects Gestalts as highly meaningful image tokens as 2-D spatial oriented discontinuities (borders, edges) enriched with information from color, optic flow, contrast (see section 3) and stereo correspondences (see section 4). In this way, by example, using color information enables to see borders that might be camouflaged to a color-blind system (e.g. magno system). The second (see part B), mainly motion-based (cf. magno system), detect Gestalts as regions with spatially-extended coherent 1st-order motion differentials (uniformity, monotonic/non-monotonic gradients).

The approaches in Part A and Part B are complementary. There exist large synergistic potentials that we are already (or will be soon) using in the ongoing project. Here are three examples:

- While in Part A Gestalts basically describe line like structures, the motion Gestalts in part B are areas with an associated common (spatio-temporal) cause in 3D. Therefore, the Gestalts of part A can be used to describe the borders of the motion Gestalts of Part

B.

- The visual Primitives described in chapter 3 essentially condense information while in part B the full optic flow is preserved. The flow can potentially be improved by the motion Gestalts and can then be used to improve feature extraction (see section 3.3.4) of the Primitives.
- In Part A chapter 6 only one RBM is supposed to occur. However, in part B (chapter 8) we are able to segment a scene into different common motion causes. Therefore, by making use of the segmentation described in chapter 8, we can extend the approach described in chapter 6 to multiple motions (e.g., multiple cars driving independently in a scene) by taking the results from the segmentation into account and applying the method separately to different scene segments.

Part A

Local Visual Multi-modal Primitives and their Contextual Relations

Chapter 2

Motivation

We address work package 3.1 by defining *local multi-modal visual Primitives* that represent scene information in a condensed and sparse way (see chapter 3). The Primitives (see figure 2.1B) comprise the modalities 2D position, 2D orientation, contrast transition (phase), colour and optic flow. Our Primitives are motivated by cortical feature processing as well as by functional requirements (as discussed in section 3.1). Figure 2.1G, figure 2.3 and figure 2.4 show the extracted Primitives from a stereo image sequence (see figure 2.2) recorded with Hella in Lippstadt. We use standard algorithms (e.g., [54]) as well a new algorithms [17] to extract features in these modalities. Moreover, we *condense and stabilize* pixel groups to achieve a more efficient and less redundant representation (WP1,1A,B). For edges, condensation takes place along local collinear structures.

By defining stereo correspondences making use of the modalities 2D orientation, contrast transition (phase) and colour information about *3D position and 3D orientation can be associated to our Primitives* (see chapter 4). We show that it is the combination of modalities that gives the best results. The orientations and positions of the corresponding entities in the right image have been drawn in figure 2.1D,F.

We further have investigated the statistical interdependencies of our Primitives (WP1,2A,B) in chapter 5. We introduced a statistical measure that codes the interdependency between events, called 'Gestalt Coefficient'. Using this measure we could show that classical Gestalt laws like collinearity and parallelism can be linked to the statistics of natural images: Going beyond the 1/f law formulated by Field [18] we could show that collinearity is the most dominant spatial second order context for edge like features in natural images. Moreover, we could show that these interdependencies increase significantly when we take also the other modalities that are coded in our Primitives into account. Based on the Gestalt coefficient we have formulated a grouping process in section 5.2 and figure 2.1E,G.

The Primitives are subject to spatial contextual modification. To all aspects of the Primitives *confidences are associated* that are modifiable according to the context. We have drawn a distinction between two principally different types on contexts that, in our view, have to be treated in different ways: We distinguish between deterministic and statistical interdependencies [47]. Deterministic Regulartities allow for deterministic predictions and are often based on geometric laws (such as for stereo and Rigid Body Motion). Statistical regularities allow for probablistic predictions (e.g., the existence of a certain edge makes a collinear edge *more likely*). In our work we make use of both kind of interdependencies (see chapter 6).

We have implemented a modulation according to the spatial context in two ways:

Enhancement of features according to their spatial 2D context: We define groups

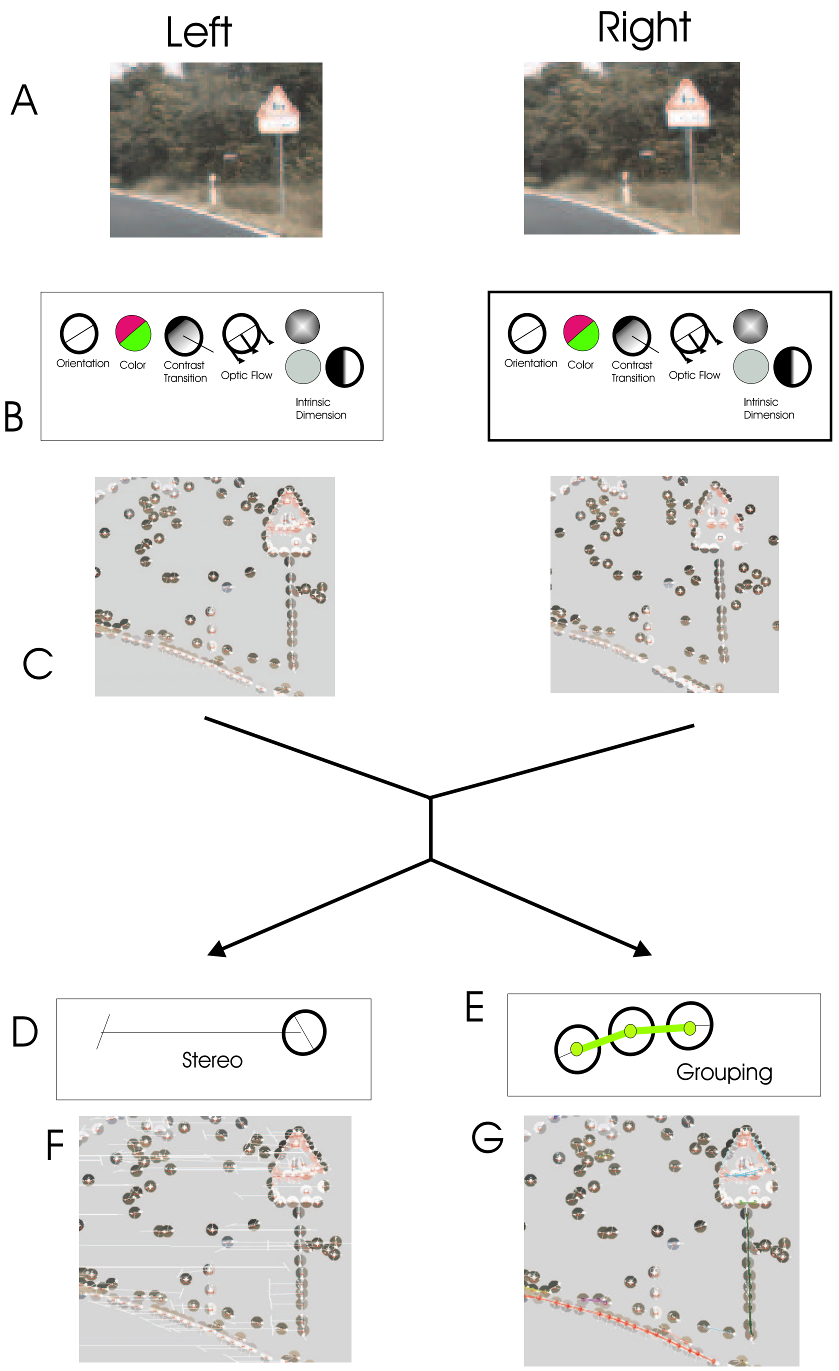


Figure 2.1: Overview of the different stages of feature processing.

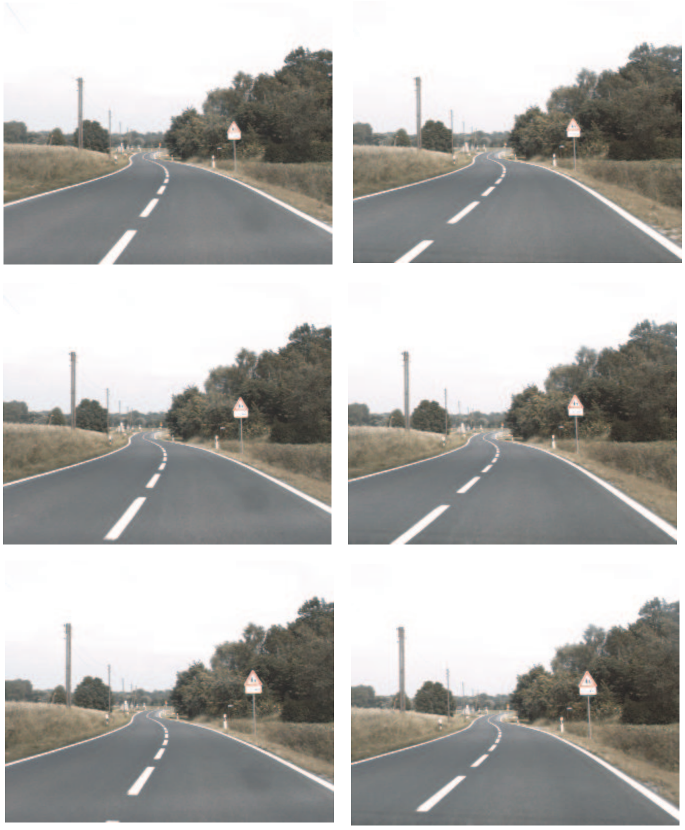


Figure 2.2: Stereo sequence recorded with Hella.

of entities by a purely statistical criterion in section 5.2. Once these groups are defined, we modulate the confidences of our Primitives: confidences become increased if the Primitives are part of a bigger group, otherwise the confidences are decreased.

Elimination of outliers according to the 3D context applied to the task RBM estimation: It turned out, that modulation according to spatial and temporal context is highly interconnected. A collinearity criterion in 3D is used to modify confidences. We demonstrate that in this way we can stabilize RBM estimation. Once the RBM has been computed we can stabilize stereo processing by modifying the confidences according to the temporal context (see chapter 6). In this way, we address already aspects of work package 3.2 and 3.3 in chapter 6.

For a more detailed description we refer to the following publications [48, 47, 45, 44, 42, 43].

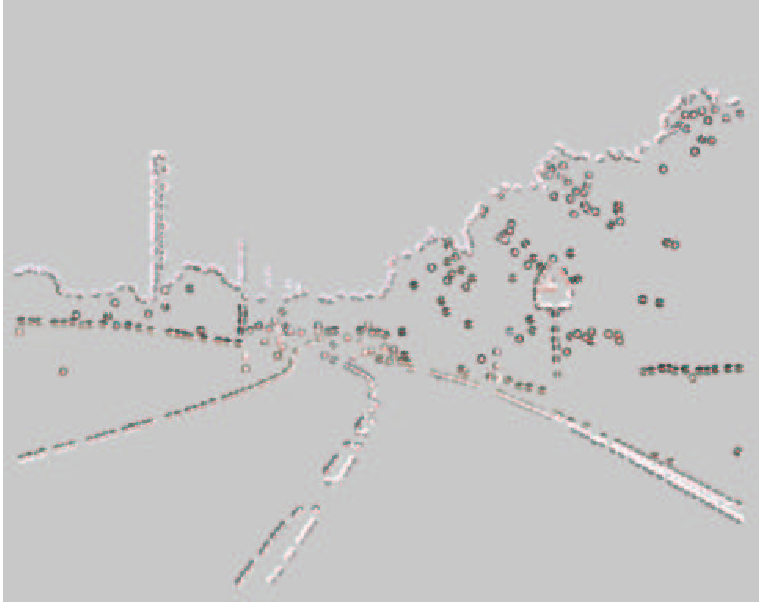


Figure 2.3: Primitives derived from image sequence in figure 2.2.

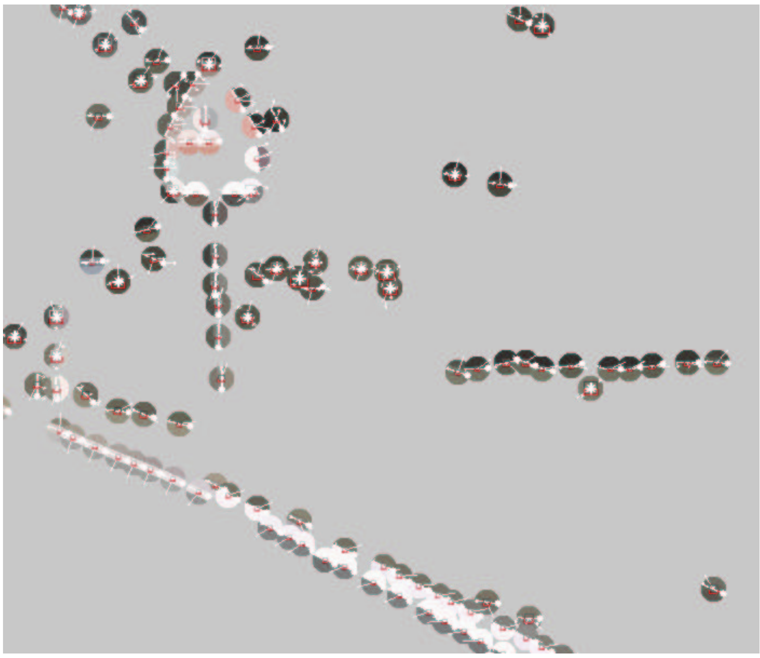


Figure 2.4: Frame from figure 2.3.

Chapter 3

Defining local multi–modal visual Primitives

A local edge can be analyzed by local feature attributes (see figure 3.1) such as orientation or energy in certain frequency bands (see, e.g., [35]). Furthermore, color can be associated to the edge. This image patch also describes a certain region of the 3D space and therefore 3D attributes can be associated such as a 3D-position or a 3D-direction (see chapter 4). Moreover, the edge changes in time due to ego-motion or object motion. Therefore time specific features such as a 2D velocity vector can be associated to this image patch [4] and 3D motion can be associated to the underlying 3D-structure (see [43]). Going again one step further the 2D image patch as well as the underlying 3D structure has a statistical relation to its neighbors commonly referred to as Gestalt principles (see chapter 5).

Therefore, this edge is an entity that has expressions in many domains (2D, 3D and time) and is involved in a lot of relations to other entities. In this work we define local multi–modal Primitives that realize these multi-modal relations. The modalities in addition to the usually applied semantic parameters position and orientation (see, e.g., [35]) are contrast transition, color, optic flow and 3D information. All these modalities are derived and coded in accordance to the edge structure, e.g., color values are coded as 'left of' and 'right of' the edge, optic flow is averaged along the edge and 3D information is coded such it allows to reconstruct a 3D edge. In addition, to each semantic parameter of the Primitives confidence information about its presence and reliability is associate that is subject to contextual modulation.

3.1 Biological Motivation

Our Primitives are motivated by processing in the human visual system as well as by functional considerations. For example, for the task object recognition it is essential to store perceptually relevant object information in a *condensed and sparse way.* We therefore intend to reduce the amount of information within an edge image patch while preserving perceptual relevant information. Furthermore, condensation is also important to reduce the cost of communication between visual entities (see below). Sparse coding has also been discussed in terms of efficient memorizing of information [57].

The definition of our Primitives involves a certain amount of prestructuring: semantic attributes are basically predetermined. However, they also *adapt* to attributes of the signal, e.g., they move their position according to the intrinsic dimension of the energy and orientation

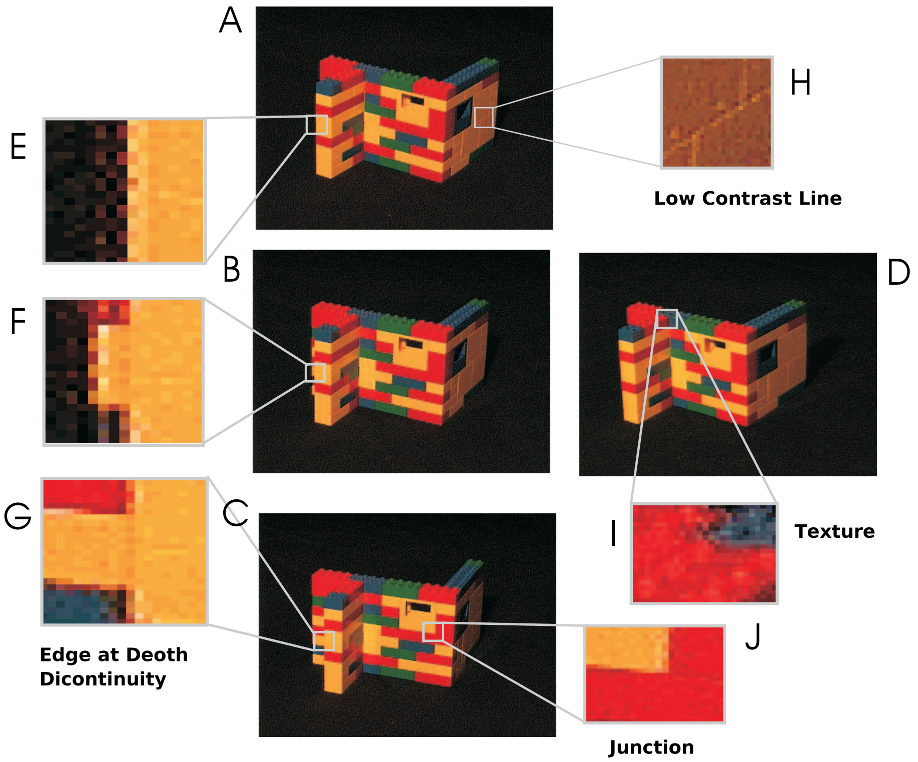


Figure 3.1: Examples of image structures in an image sequence

distribution of the image patch they describe. Also spatial and temporal relations, such as the coding of Gestalt principles, allows for a process of adaptation (see chapter 5 and 6).

In the human visual system beside local orientation also other modalities such as color and optic flow (that are also part of our Primitives) are computed (see, e.g. [21]). All these low level processes face the problem of an extremely high degree of vagueness and uncertainty [1]. This arises from a couple of factors. Some of them are associated with image acquisition and interpretation: Owing to noise in the acquisition process along with the limited resolution of cameras, only erroneous estimates of semantic information (e.g., orientation) are possible. Furthermore, illumination variation heavily influences the measured grey level values and is hard to be modeled analytically [33]. Information extracted across image frames, e.g., in stereo and optic flow estimation, faces (in addition to the above mentioned problems) the correspondence and aperture problem which interfere in a fundamental and especially difficult way (see, e.g., [2, 37]). However, the human visual systems acquires visual representations which allow actions with high precision and certainty within the 3D world under rather uncontrolled conditions. The human visual system can achieve the needed certainty and completeness by integrating visual information across modalities (see, e.g., ([58, 32]). This integration is manifested in the huge connectivity across brain areas in which the different visual modalities are processed as well as in the large number of feedback connections from higher to lower cortical areas (see, e.g., [21]). The power of modality fusion arises from the *huge intrinsic relations* given by regularities within and across visual modalities. The essential need for integrating visual information in addition to optimizing single modalities to design efficient artificial visual systems has also been recognized in the vision community after a long period of work on improving single modalities

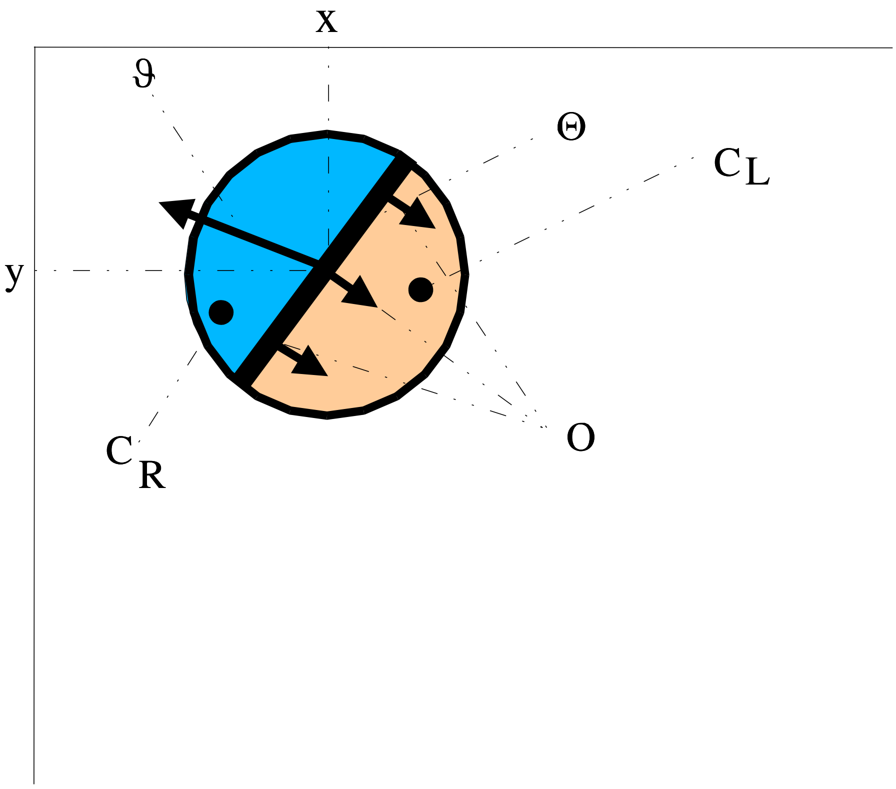


Figure 3.2: Schematic representation of a basic feature vector.

[1].

However, integration of information makes it necessary that local feature extraction is subject to modification by contextual influences and this communication has necessarily to be paid for with a certain cost. This cost can be reduced by limiting the amount of information transferred from one place to the other, i.e. by reducing the bandwidth. This is an additional reason why we are after a *condensed* description of a local edge patch, which however contains the relevant information. Here relevance has to be understood not only in an information theoretical sense, but in a global sense (it has to be subject to modifications by global interdependencies, in particular it has to be connectable with other entities) and action oriented sense (the transfered information has to be relevant for the actions the individium has to perform).

3.2 Local Descriptors of Edges

In the following we describe the coding of semantic associated to our Primitives. We compute the following semantic attributes and associate them to our Primitives (see also figure 3.2).

Position: Each Primitive has a position $\mathbf{x}$. This position reflects an *interpretation* of a whole image patch and is adapted to the spatial structure of the patch.

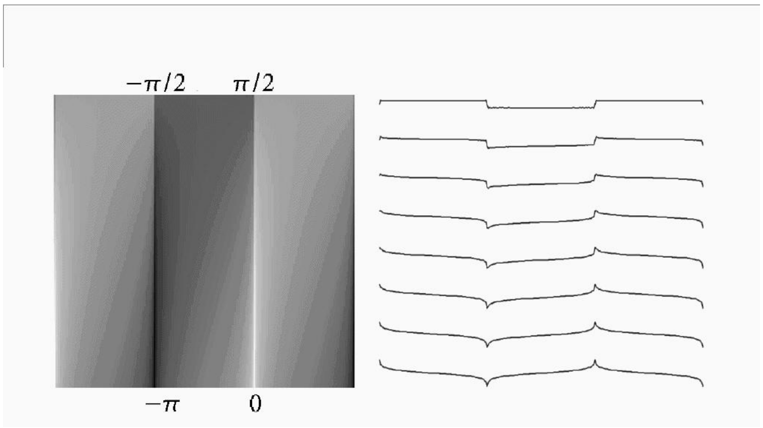


Figure 3.3: **Left:** Variation of contrast transition according to phase variation. **Right:** Luminance profiles corresponding to left image.

Frequency: We describe the signal on different frequency levels f independently (see figure 3.4). Often the decision in which frequency band the relevant information does occur is a difficult one, therefore we leave this decision open. It may be even that for the same position on different frequency levels there occur different kind of semantic information.

Orientation: The local orientation associated to the image patch is described by θ.

Contrast transition: The contrast transition is coded in the phase φ of the applied filter (see figure 3.3) Note that the phase changes under change of background (in contrast to color, see below). In case of boundaries of objects it rather represents a description of the transition between object background than it represents a description of the object.

Color: Color $(\mathbf{c}_l, \mathbf{c}_m, \mathbf{c}_r)$ is processed by integrating over image patches in coincidence with their edge structure (e.g., , integrating separately over the left and right side of the edge as well as a middle strip). In case of a boundary edge of a moving object at least the color at one side of the edge is expected to be stable (see figure 3.1E,F,G). In this way stable object attributes can be accumulated over time (see chapter 6).

Optic Flow: Local displacements $\mathbf{o}$ is computed by a well known optical flow technique [54] and is averaged across a local area defined by pixel points with high magnitude.

Furthermore, we represent the system's confidence that the entity e does exist by c. We end up with a parametric description of an Primitive as

$$E = (\mathbf{x}, f, \theta, \varphi, (\mathbf{c}_l, \mathbf{c}_m, \mathbf{c}_r), \mathbf{o}; c).$$

Furthermore, to each of the parameters $\varphi, (\mathbf{c}_l, \mathbf{c}_m, \mathbf{c}_r), \mathbf{o}$ there exist confidences $c_i, i \in \{\varphi, \mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r, \mathbf{o}\}$ that code the reliability of the specific sub–aspects.

All semantic parameters and confidences are thought to be subject to contextual information (see chapter 5 and 6). Figure 2.3 and figure 2.4 show an example of extracted features.

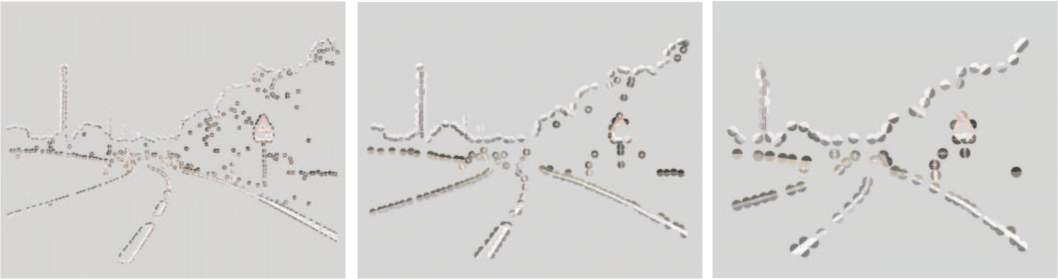


Figure 3.4: Primitives derived for different frequencies.

3.3 Definition of the Primitives

The definition of our Primitives is done in three steps:

Preprocessing: We apply a filtering process to extract magnitude, orientation and phase for each pixel position (see section 3.3.1).

Sparsification: We sparsify our representation to a small set of discrete pixels (see section 3.3.3). For sparsification the concept of intrinsic dimensionality (see section 3.3.2) is vital.

Feature Extraction: We extract features for the above mentioned modalities (see section 3.3.4).

3.3.1 The monogenic Signal

Our Primitives are to a large degree based on a rotation invariant quadrature filter, which is derived from the concept of the *monogenic signal* ([17]). Considered in polar coordinates, the monogenic signal performs a *split of identity* [17]: it decomposes an intrinsically one-dimensional signal into intensity information (amplitude), orientation information, and symmetry information (phase information). These features are point-wise mutually orthogonal. The intensity information can be interpreted as an indicator for the likelihood of the presence of a certain structure with a certain orientation and a certain contrast transition (called 'structure' in [16]).

A quadrature filter based on the monogenic signal is rotation invariant, i.e., it commutes with the rotation operator. Hence, for an appropriate choice of polar coordinates, two coordinates do not change under rotations (amplitude and phase), whereas the third coordinate directly reflects the rotation angle. This kind of quadrature filter, which is called *spherical quadrature filter* [15], is formed by a triple of filters: a radial bandpass filter and its two Riesz transforms. As in [15] we construct the bandpass filter from *difference of Poisson* (DOP) filters, in order to have analytic formulations of all filter components in the spatial domain and in the frequency domain. The DOP filter is an even filter (wrt. point reflections in the origin) and its impulse response (convolution kernel) and frequency response (Fourier transform of the kernel) are respectively given by

$$
\begin{aligned}
h_{\mathrm{e}}(\mathbf{x}) &= \frac{s_1}{2\pi(|\mathbf{x}|^2 + s_1^2)^{\frac{3}{2}}} - \frac{s_2}{2\pi(|\mathbf{x}|^2 + s_2^2)^{\frac{3}{2}}} && (3.1) \\
H_{\mathrm{e}}(\mathbf{u}) &= \exp(-2\pi|\mathbf{u}|s_1) - \exp(-2\pi|\mathbf{u}|s_2) \ . && (3.2)
\end{aligned}
$$

For convenience, we combine the two Riesz transforms of the DOP filter in a complex, odd filter, yielding the impulse response and the frequency response

$$\begin{aligned} h_{\mathrm{o}}(\mathbf{x}) &= \frac{x_1 + ix_2}{2\pi(|\mathbf{x}|^2 + s_1^2)^{\frac{3}{2}}} - \frac{x_1 + ix_2}{2\pi(|\mathbf{x}|^2 + s_2^2)^{\frac{3}{2}}} && (3.3) \\ H_{\mathrm{o}}(\mathbf{u}) &= \frac{u_2 - iu_1}{|\mathbf{u}|}(\exp(-2\pi|\mathbf{u}|s_1) - \exp(-2\pi|\mathbf{u}|s_2)) \ , && (3.4) \end{aligned}$$

respectively. The impulse responses of the filters are shown in figure 3.5 for $(s_1, s_2) = (1, 2), (2, 4), (4, 8)$. These filters were used for the feature generation. The corresponding frequency responses are illustrated in figure 3.6.

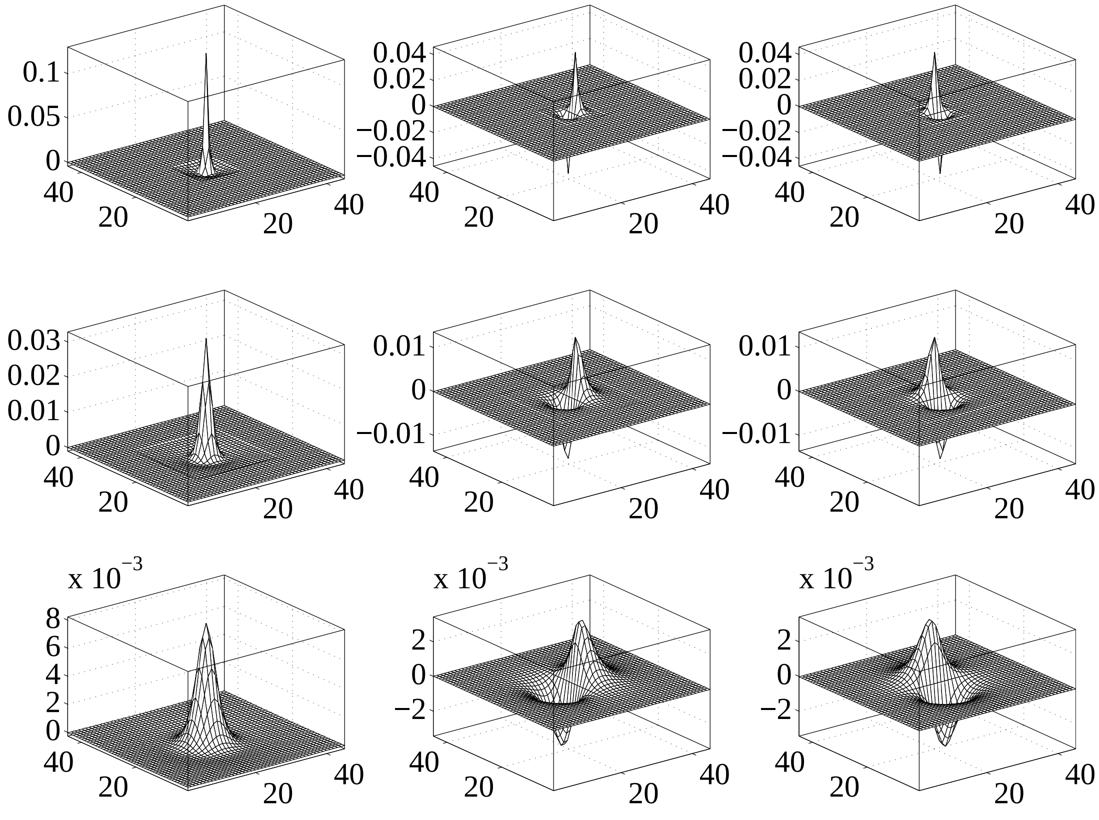


Figure 3.5: Impulse responses of the DOP filter and its Riesz transforms. From left to right: DOP filter, first Riesz transform, second Riesz transform. From top to bottom: scales (1,2), (2,4), (4,8).

In order to illustrate the effect of the quadrature filter to an image I, we applied the three filters to the image in the top left of figure 3.7. The convolution results $I_{\mathrm{e}} = h_{\mathrm{e}} * I$ and $I_{\mathrm{o}} = h_{\mathrm{o}} * I$ (complex) can be found in the same row.

The split of identity is obtained by switching to appropriate polar coordinates. In particular,

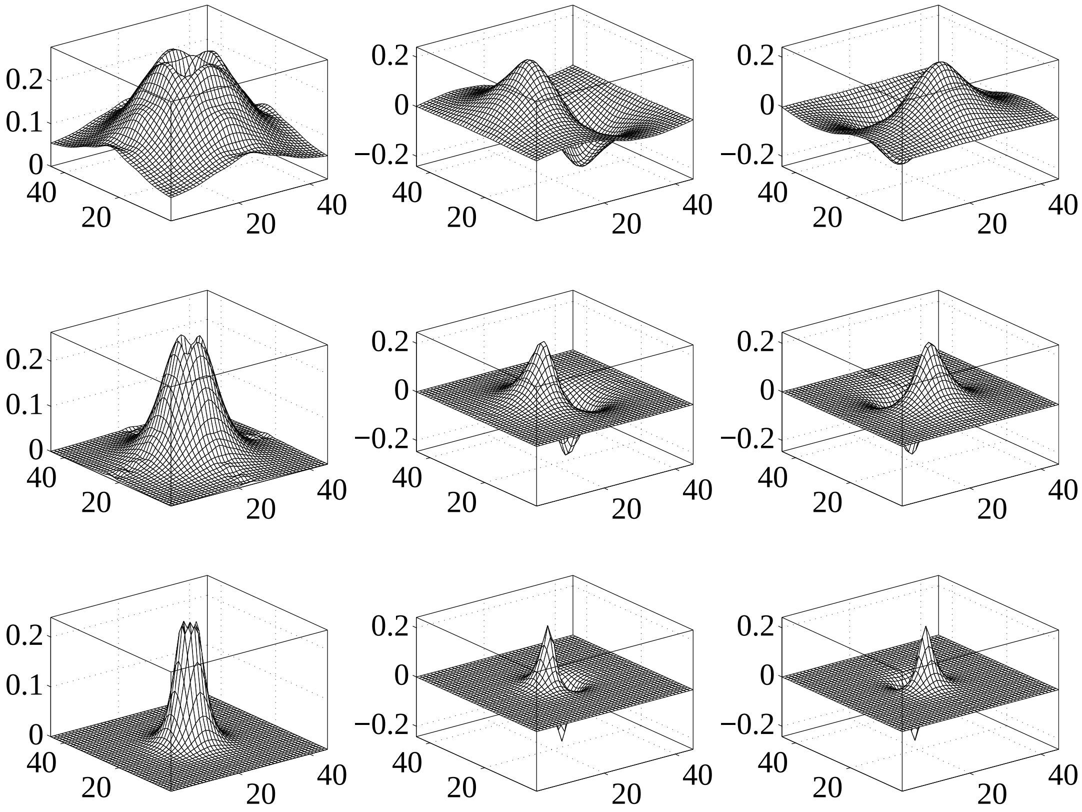


Figure 3.6: Frequency responses of the DOP filter and its Riesz transforms. From left to right: DOP filter, first Riesz transform, second Riesz transform. From top to bottom: scales (1,2), (2,4), (4,8).

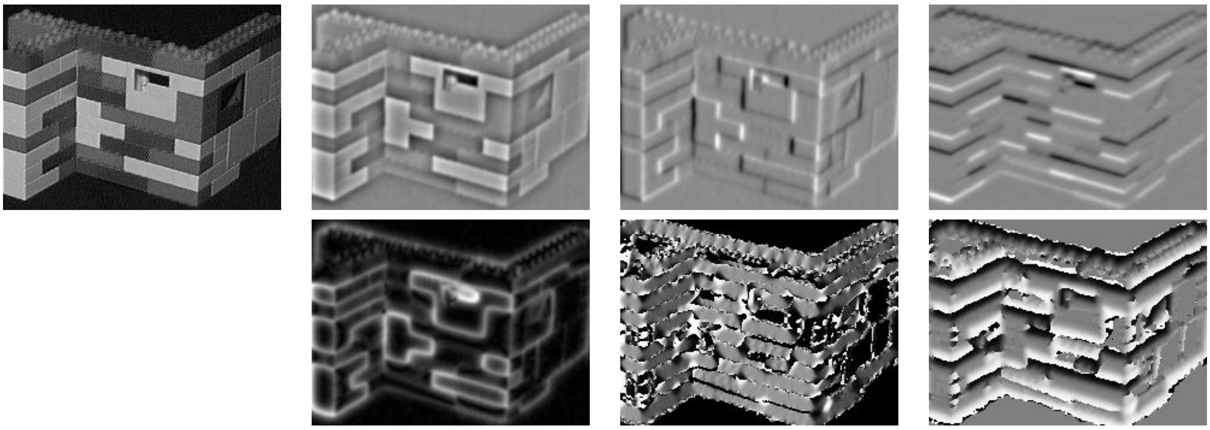


Figure 3.7: Upper row, from left to right: original image, image filtered with even kernel, image filtered with real part of odd kernel, and image filtered with imaginary part of odd kernel. Bottom row: interpretation in polar coordinates. From left to right: amplitude, masked orientation, and masked phase.

we transform the filter responses according to

$$\overline{m}(\mathbf{x}) = \sqrt{I_{\mathrm{e}}(\mathbf{x})^2 + |I_{\mathrm{o}}(\mathbf{x})|^2} \tag{3.5}$$

$$\theta(\mathbf{x}) = \arg I_{\mathrm{o}}(\mathbf{x}) \pmod{\pi} \tag{3.6}$$

$$\varphi(\mathbf{x}) = sign(\Im\{I_{\mathrm{o}}(\mathbf{x})\}) \arg(I_{\mathrm{e}}(\mathbf{x}) + i|I_{\mathrm{o}}(\mathbf{x})|) \ , \tag{3.7}$$

which gives the desired amplitude, orientation, and phase information. In case of our example from figure 3.7, we obtain the results illustrated in the bottom row of figure 3.7.

We want to interpret the amplitude as a confidence for the occurrence of a local line segment. We therefore apply a normalization $m(\mathbf{x}) = N(\overline{m}(\mathbf{x}))$ of the amplitude to the interval $[0,1]$ as already described in [46]. This normalization takes local and global information into account and is able to detect structures also in areas with low energy.

Figure 3.8 shows a radial slice cut through the DOP bandpass filters for a certain range of scales and their superposition demonstrating a homogeneous covering of the frequency domain. For infinite many bandpass filters, the superposition is one everywhere, except for the origin. The applied bandpasses are indicated by the darker color.

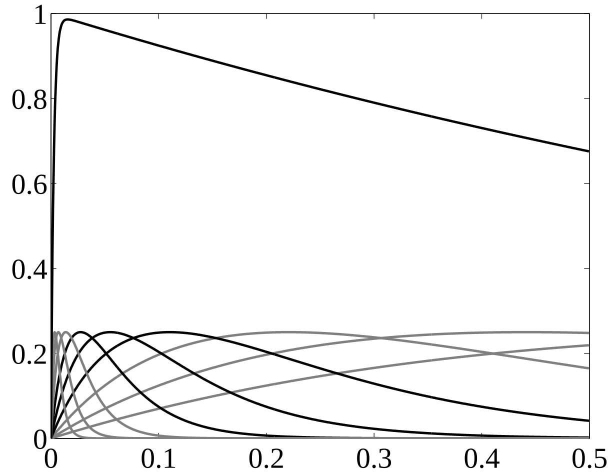


Figure 3.8: DOP bandpass filters and their superposition approaching the identity. The superposition and the filters applied in this paper are indicated by the darker lines.

3.3.2 The Concept of intrinsical Dimension

There exist special subdomains of local image patches that require different treatment since they are involved in different contexts. These subdomains can be characterized by their intrinsic dimensionality (see, e.g., [73, 26, 15]).

Characterization of intrinsic Dimension

The intrinsic dimensionality of a local image patch can be characterized by a term expressing the local energy $\tilde{m}$ in the image patch and a term $\tilde{\sigma}^2(\theta)$ expressing the variance of the orientation in the image patch.

For example, an intrinsically zero dimensional image patch is characterized by low amplitude within the whole patch ($\tilde{m} \approx 0$, then also $\tilde{\sigma}^2(\theta) \approx 0$). In the triangle shown in figure 3.9 they correspond to the coordinate $(0, 0)$. However, although $m \approx 0$ the local image patch can also be a projection of a 3D–edge (that usually corresponds to i1D signals) or junction (that usually corresponds to i2D signals). The the low contrast may be caused by e.g., unlikely background–object constellation (see figure 3.1E) or bad illumination (see figure 3.1H). Therefore we do not want to impose a final decision on the intrinsic dimension but want to allow for diverse interpretations (e.g., as edges, homogeneous image patches, textures (see figure 3.1I) or junctions (see figure 3.1J)). We therefore (and this is a general design principle of our approach) associate confidences to the different interpretations of intrinsic dimensionality (see below).

Intrinsically one-dimensional signals: An intrinsically one dimensional signal is characterized by a high amplitude and a low variance of local orientation within an image patch. In the triangle in figure 3.9 this corresponds to the coordinate (1,0). Orientation can only meaningfully be associated to an intrinsically one–dimensional signal patch. In contrast, for a homogeneous image patch (i0D) or a junction (i2D) the concept of orientation does not make sense at all.

Once a 2D–orientation is associated we can associate also a 3D–orientation by depth cues such as stereo (see chapter 4). With an intrinsically one–dimensional image patch specific problems are associated, for example the aperture problem which is less severe (or non existent) for intrinsically two–dimensional signals.

Intrinsically zero-dimensional signals: Intrinsically 0D signals correspond to the coordinate $(0, 0)$ on the triangle shown in figure 3.9. A mean color-value can be associated to a zero–dimensional image patch (and most patches in natural images have this property). It can be assumed with high likelihood that it is a projection of a 3D–surface (however under certain circumstances in can also be caused by, e.g., a 3D depth discontinuity).

It is a severe problem to extract any kind of 3D information *directly* from this image patch since it is impossible to find correspondences at homogeneous surfaces. However, 3D–information might be acquired indirectly by taking context information into account.

Intrinsically two-dimensional signals: Intrinsically 2D signals are characterized by high energy and high orientation variance. They correspond to the coordinate $(1, 1)$ in the triangle shown in figure 3.9. A parametric description of 2D-image patches is more difficult since there are at least two possible 3D–sources for an intrinsically two–dimensional image patch. First, it may be caused by edges meeting in a point or it may be caused by texture. The underlying 3D–description would be different. A texture is most likely produced by a surface–like structure while a junction most likely is associated to a specific 3D–depth discontinuity.

The aim of the current work is to define a local edge descriptor which analysis intrinsically one–dimensional image patches which are associated to the 3D–space (see chapter 4) and time domain (see chapter 6). It is important to distinguish the 3 cases described above since the concept of 2D and 3D orientation is only applicable to i1D signals. Of course, in a complete system i0D and i2D signals need also be regarded.

Coding intrinsic dimensionality by barycentric coordinates

We code the confidences associated to the different intrinsic dimensions (c_{0D}, c_{1D}, c_{2D}) by barycentric coordinates. Barycentric coordinates describe the area of the triangle opposite to the points spanning the triangle (see figure 3.9). We define the term $\tilde{m}$ expressing the energy and the term $\tilde{\sigma}^2(\theta)$ expressing the variance of local orientation by

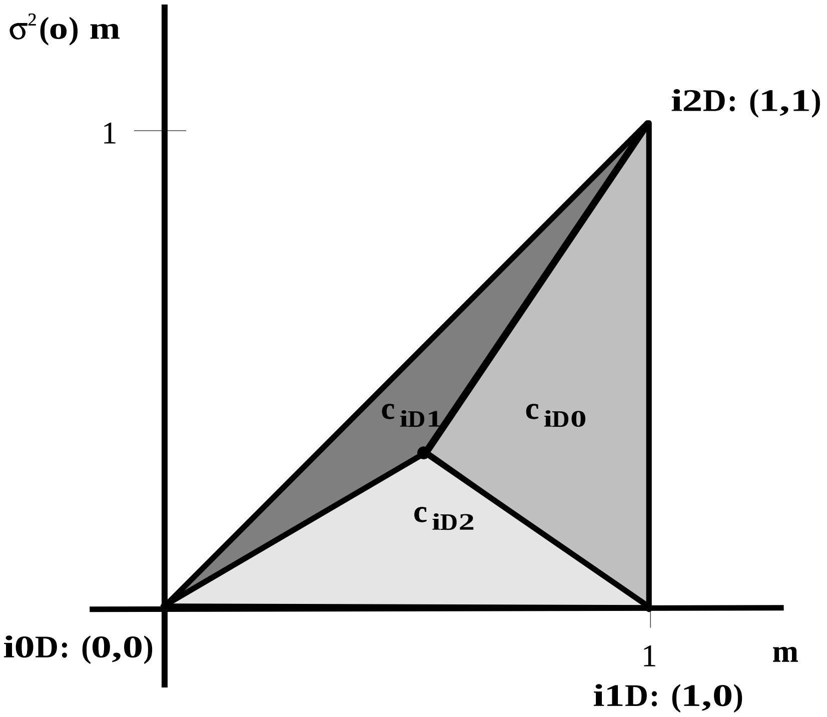


Figure 3.9: Coding of intrinsic Dimensionality using brycentric Coordinates

$$\tilde{m} = m^{n_m}$$
$$\tilde{\sigma}^2(\theta) = (m \cdot \sigma^2(\theta))^{n_{\sigma^2(\theta)}}$$

$\tilde{\sigma}^2(\theta)$ is normalized such that it is in $[0,1]$. $\tilde{m} \in [0,1]$ holds by definition. We define 2D points by

$$\mathbf{p} = (p_1, p_2) = (\tilde{m}, \tilde{\sigma}^2(\theta)).$$

Our confidences then are simply the barycentric coordinates of a point $\mathbf{p}$ in the triangle spanned by $(0,0), (1,0), (1,1)$:

$$c_{0D} = 1 - p_1$$
$$c_{1D} = p_1 - p_2$$
$$c_{2D} = p_2$$

Note that all c_{0D}, c_{1D}, c_{2D} are positive and add up to 1 by definition.

3.3.3 Sparsification and Definition of Position

We want to express a local image patch corresponding to an intrinsically one–dimensional signal patch by a Primitive E. For this we first perform a hexagonal sampling of the image into areas

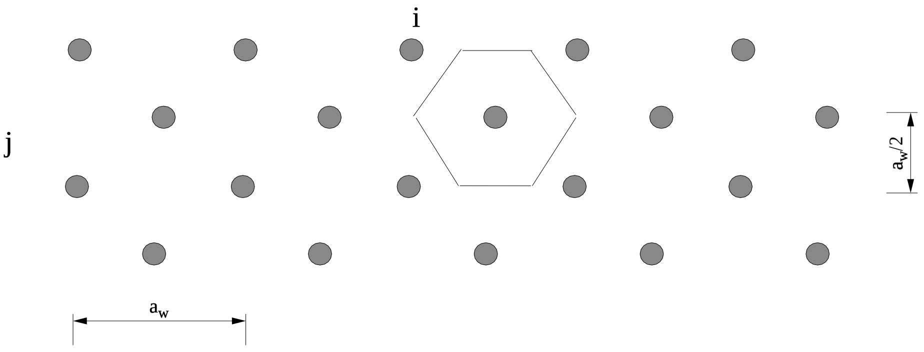


Figure 3.10: Hexagonal sampling

$A^{(i,j)}$ (see figure 3.10) with center $s^{(i,j)}$ defined by

$$\begin{array}{rcll} s_1^{(i,j)} & = & i \cdot a_w & \text{for } i \text{ even} \\ s_1^{(i,j)} & = & a_w/2 + i \cdot a_w & \text{for } i \text{ odd} \\ s_2^{(i,j)} & = & j \cdot \frac{a_w}{2} & \end{array}$$

The sampling distance a_w depends on the applied filter. Since we want to sparsify the signal we have to decide about an optimal position. We do this depending on the intrinsic dimension of the signal. For intrinsically 0D signals we choose the position of the corresponding Primitive as the sampling point:

$$\mathbf{x}^{(i,j)} = (s_1^{(i,j)}, s_2^{(i,j)})$$

For intrinsically 1D signals we choose the point within the image patch with highest energy on the line $L_{\theta T}$ orthogonal to the dominant orientation:

$$\mathbf{x}^{(i,j)} = \max_{(x_1,x_2) \in L_{\theta T}^{i,j}} \{m(x_1, x_2)\}.$$

For intrinsically 2D signals we select the amplitude maxima in each hexagonal patch:

$$\mathbf{x}^{(i,j)} = \max_{(x_1,x_2) \in A^{i,j}} \{m(x_1, x_2)\}.$$

There might occur two $\mathbf{x}^{(i,j)}, \mathbf{x}^{(i',j')}$ which have close distance, i.e., are caused by the same image event. To avoid two similar descriptors in our image representation we basically draw a circle around each $a_{(i,j)}$ and look whether there exist another $a_{(i',j')}$ close to $a_{(i,j)}$. If this is the case we delete the one with smaller energy.

3.3.4 Feature Extraction

Having defined a sparse sampling we can now extract the semantic attributes in the different modalities.

Orientation

The orientation of pixels is computed by averaging the orientations over the pixel positions with high amplitude:

$$\theta =< \theta >_{\{\theta(\mathbf{x})|m(\mathbf{x})>t\}}$$

The confidence c is set to the confidence for the intrinsic dimension 1 c_{D1}.

Coding of Contrast Transition using Phase Information

The phase can be used to interpret the kind of contrast transition at this maximum [41], e.g., a phase of $\frac{\pi}{2}$ corresponds to a dark–bright edge, while a phase of 0 corresponds to a bright line on dark background. The continuum of contrast transition at an intrinsic one-dimensional signal patch can be expressed by the continuum of phases (see figure 3.3).

We compute the phase φ by averaging the orientations over the pixel positions with high amplitude:

$$\varphi =< \varphi >_{\{\varphi(\mathbf{x})|m(\mathbf{x})>t\}} .$$

The confidence c_p is computed proportional to the variance of the phase at the pixel positions with high amplitude:

$$c_\varphi = \sigma^2(\varphi)_{\{\varphi(\mathbf{x})|m(\mathbf{x})>t\}}.$$

Coding of Color Information

We code color information at the left side, middle and right side of the edge $(\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r)$ by averaging over the different patches P^l, P^m, P^r. The radius r and the width of the middle patch are set according to properties of our filter.

$$\mathbf{c}_i =< c(\mathbf{x}) >_{\mathbf{x}\in P^i}$$

Again, the confidence for each patch is set according to the variance in each patch:

$$c_{c_i} = \sigma^2(\mathbf{c})_{P^i}.$$

Coding of Optic Flow Information

We extract optic flow information with the algorithm developed by Nagel [54] which turned out to be very good especially for edges. Since our edge detector deals with intrinsically one–dimensional signals every local optic flow algorithm faces the aperture problem which can only be disambiguated by context information, e.g., by information about the 3D motion (see chapter 6).

To achieve a more stable optic flow statement by averaging the optic flow vectors over the pixel positions with high amplitude.

$$\mathbf{o} =< \mathbf{o} >_{\{\mathbf{o}(\mathbf{x})|m(\mathbf{x})>t\}}$$

The confidence $c_\mathbf{o}$ is set according to the variance of the optic flow vectors:

$$c_\mathbf{o} =< \mathbf{o} >_{\{\mathbf{o}(\mathbf{x})|m(\mathbf{x})>t\}} .$$

We end up with the Primitive

$$E = (\mathbf{x}, f, \theta, \varphi, (\mathbf{c}_l, \mathbf{c}_m, \mathbf{c}_r), \mathbf{o}; c).$$

Chapter 4

Generating 3D Primitives

In this chapter we define a stereo similarity function that is used to define correspondences between image Primitives extracted from a left and right image of a stereo system. In figure 4.1 these correspondences are displayed. Once correspondences are defined we can, making use of the modalities position $\mathbf{x}^l, \mathbf{x}^r$ and orientation θ^l, θ^r compute a 3D position $\mathbf{X}$ and a 3D orientation $\mathbf{O}$ (see figure 4.2). It turns out that the use of all modalities supports stereo matching and that optimal results can be achieved through an optimal weighting across the modalities.

4.1 Remarks about Stereo Processing

In stereo processing with calibrated cameras we can reconstruct 3D points from two 2D points correspondences by computing the point of intersection of the two projective lines generated by the corresponding image points and the optical centers of the cameras (see figure 4.2 and, e.g., [14]). However, most meaningful image structure is intrinsically one-dimensional (see section 3.3.2), i.e., is dominated by edges or lines. Orientation at intrinsically one-dimensional image structures can be estimated robustly and precisely by various methods (see, e.g., [35]). Hence, it is sensible to use orientation information as well for the representation of 3D–information in visual scenes. From two corresponding 2D points with associated orientation. we can reconstruct a 3D point with associated 3D orientation, (see figure 4.2 and , e.g., [14, 65]): Each of the 2D-lines generated by the points with associated orientation together with the optical center of the camera span a 3D plane (figure 4.2). Then the intersection of both planes defines the 3D orientation of a 3D–line segment. Only in case that the two planes are identical reconstruction is not possible.

The problem at hand is to find correspondences between image structures in the left and right image. There is severe trouble connected to this problem: Since the scene is seen from different views *the image structure differs* in the left and right image, position of corresponding points are different and the local orientation at these points differ as well (see figure 4.3). In fact, these differences are the reason why reconstruction is possible: Dissimilarity in position (disparity) determines the depth while dissimilarity in orientation determines the 3D orientation (note that all geometric attributes of an oriented 3D point are covered by its 3D position and 3D orientation). The photometric information going beyond geometric information (in the following called *structural information* coded in φ) undergoes a complex transformation which has to take into account the transfer of pixel positions (depending on the 3D geometry of the projected object) as well as variation of reflection properties of surfaces according to view

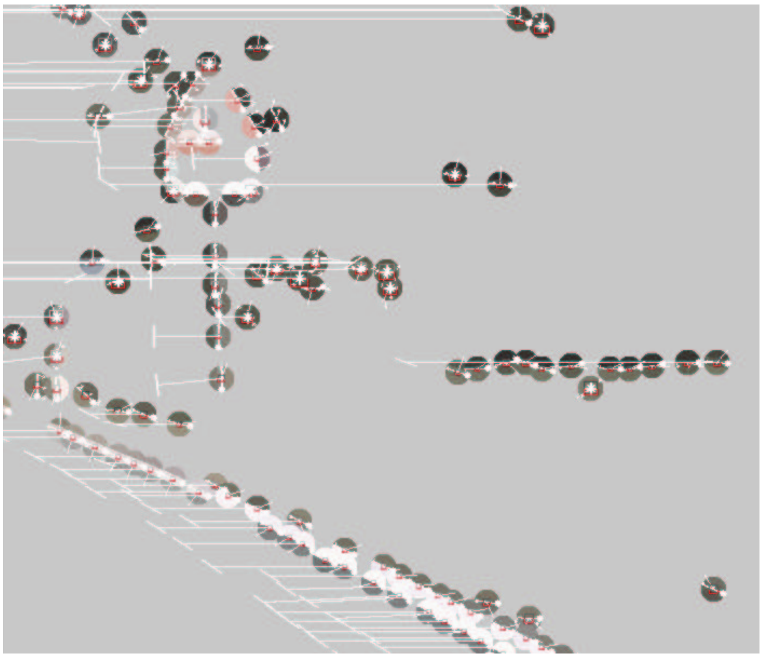


Figure 4.1: Features with associated Stereo correspondences.

variation and the occurrence of occlusion.

For 3D reconstruction we face the following *similarity–dissimilarity dilemma*: We want to find correspondences by *similarity* of the image patches but we want to reconstruct utilizing their *dissimilarity* in position and the geometric property orientation. While dissimilarity in position can be resolved by a transition of the image patch on the epipolar line the difference in orientation can only be resolved by more complex mechanisms (see, e.g.,[19]).

Some stereo similarity functions for intrinsically 1D information use geometric attributes (orientation, length) [2, 53]. However, ambiguity of geometric information leads to a large number of potential matches. Furthermore, significant variation of orientation in both images can occur for entities with small depth (see figure 4.3). In this paper we will show that we can improve stereo matching significantly by using structural information in addition to geometric information and we give measures for their relative importance.

Alternatively to methods that use geometric information only for feature matching, in classical stereo approaches often a kind of template matching is used to find correspondences in images: Local image patches are compared pixel-wise, see e.g., [66, 40]. These methods are also called 'area-based stereo' or 'intensity based correspondence analysis' [37]. The similarity function might be a (normalized) squared error (see [66]) or a (normalized) cross correlation (see [37]). In these approaches the above mentioned similarity–dissimilarity dilemma is not treated explicitly. Indeed, the underlying argument or 'hope' is that despite the deviation of orientation the template match is sufficiently close, a hope which is not necessarily justified (see figure 4.3). This 'hope' is with high likelihood fulfilled, when the depth in relation to the basis width of the stereo rig is sufficiently large. The performance of matching procedures will decline, when orientation difference is too large since they do not distinguish the two above-mentioned factors. Furthermore, they are not able to make use of 3D–orientation information since they do not represent 2D-orientations explicitly.

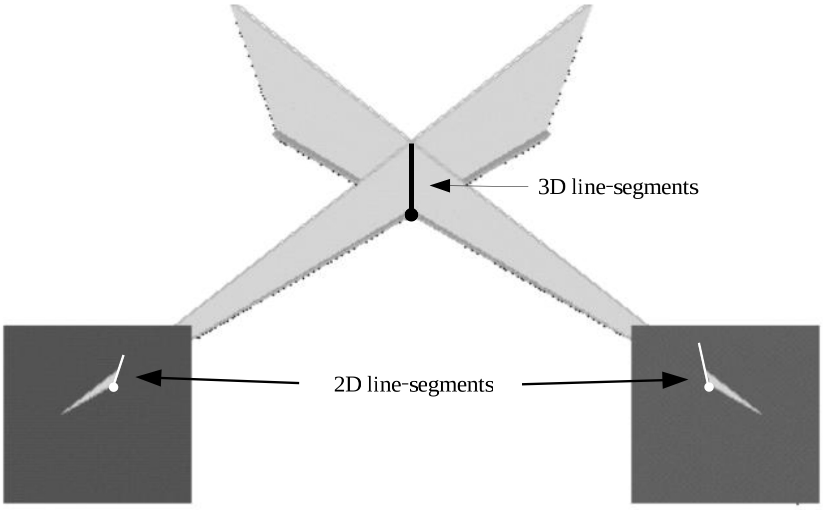


Figure 4.2: Reconstruction of 3D-point and 3D-orientation from two 2D-point and direction correspondences.

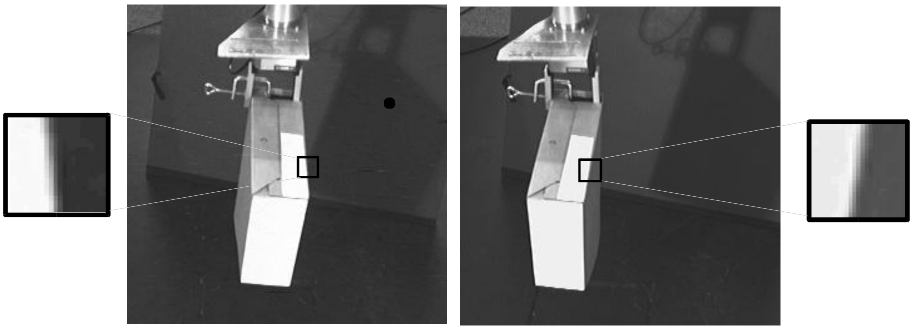


Figure 4.3: Dissimilarity of corresponding image patches in stereo images.

Some authors use both factors, orientation and structural information. In [19] variation of the local image patches are taken into account explicitly by applying an affine transformations of the grey values of the image patch. The parameters of this affine transformation have to be computed by finding a solution of an over-determined set of equations. Once these parameters are known, relative orientation difference of the image patches can be used for reconstruction. Of course, solving the set of equations can be a time demanding procedure. Taking assumptions about the 3D geometry into account (more specifically, assuming the edge being produced by the intersection of planes) the complexity of the affine transformation can be reduced [65] but still an optimization method has to be applied. Other problems concerned with this approach are that the assumption of plane surfaces is not necessarily full-filled. Furthermore, for edges caused by intersection of strictly homogeneous 3D–surfaces an optimal transformation can not be computed. Finally and most importantly, from the point of view of object representation a *more compact storage of structural information than the image patch itself is wanted.*

In this paper, we introduce a similarity function which uses geometric information (orientation) *and* structural information in a direct way, i.e., without the need of solving a set

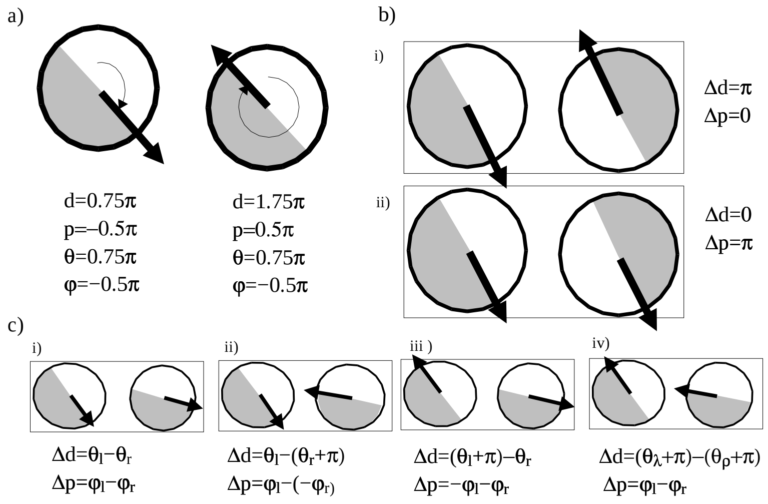


Figure 4.4: a) Given extracted orientation θ and phase φ the same image patch can be interpreted in terms of direction d and phase p as $(d = \theta + \pi, p = -\varphi)$ (left) or $(d = \theta, p = \varphi)$ (right). b) The similarity in structure and orientation of two image patches changes with the interpretation of direction. c) Theoretically possible interpretations of direction in the left and right image patch. Note that ii) and iii) are geometrically not possible according to the Stereo coherence constraint.

of equations. An intrinsically one-dimensional structure in a grey level image can be described by orientation (or geometric information) and information about its structure (e.g., it can be distinguished between being a dark/bright (bright/dark) edge or a bright (dark) line on dark (bright) background). Of course, there is a continuum between these different grey level structure. The local phase to take the grey level information into account (as one parameter in addition to orientation) in a very compact way (see, e.g., [26, 41, 17]).

As it was shown by e.g., [40, 36] color also is an important cue to improve stereo matching. We use color triplets to describe the left and right side of the edge in RGB space: Color at the left and right side of an edge is averaged to two color vectors indicating the mean color structure of two half sides (see figure 3.2).[1]

The paper is structured as following: In image processing we are able to compute a local *orientation*. However, taking more global interdependencies (such as consistency across different views) into account we can extend the concept of orientation to the concept of *direction*. In section 4.2 we therefore discuss the concept of direction in more detail. A similarity function is derived in section 4.3 that allows to steer explicitly the influence of the orientation deviation in contrast to the structural information and also the influence of phase versus color information. The relative importance of orientation, phase and color is investigated in section 4.4.

We would like to point out that it is not our aim to derive a perfect stereo system. Stereo is an ambiguous visual modality since the correspondence problem can become extremely awkward in complex scenes and mismatches lead to wrong 3D estimates. Integration of other visual modalities (see, e.g., [1, 48, 9]) and integration of ambiguous information over time (see, e.g. [13,

[1] More precisely we use for signal patches corresponding to lines (i.e., phase close to 0 or π) also a color value for the center line. For the sake of simplicity we neglect this in the following description.

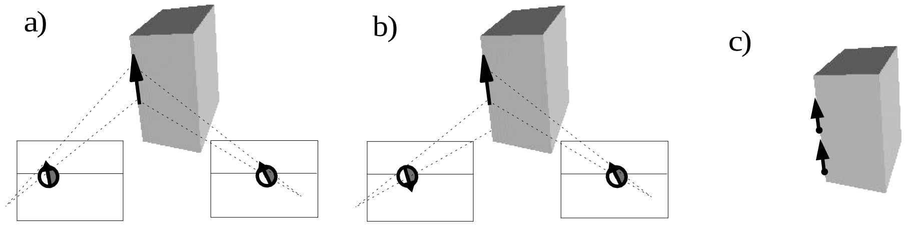


Figure 4.5: a) Geometrically possible interpretation of direction. b) Geometrically impossible interpretation of direction. c) Good Continuation.

38, 65, 43]) has to be used to achieve robust information. However, the aim of this paper is to define and investigate an appropriate local similarity function which makes use of structural and geometric information and to derive statements about the relative importance of geometric versus structural information, and phase versus color information.

Our Primitives extracted from the left or right image respectively (see chapter 3) have the form

$$E^l = (\mathbf{x}^l, \theta^l, \varphi^l, (\mathbf{c}_l^l, \mathbf{c}_m^l, \mathbf{c}_r^l)) \text{ and } E^r = (\mathbf{x}^r, \theta^r, \varphi^r, (\mathbf{c}_l^r, \mathbf{c}_m^r, \mathbf{c}_r^r))$$

and here we define a stereo similarity function for these Primitives. Note that we neglect $\mathbf{c}_m$ here.

4.2 Direction and Orientation

In the feature processing described in chapter 3 we extract orientation information θ which takes values in $[0, \pi)$. However, when we add structural information at the local edge (see figure 4.4a) we can extend the concept of orientation to the concept of direction (see figure 4.4a and [26]), parameterized by $d \in [0, 2\pi)$ (Note that the local phase can change as well when we go from orientation to direction. We denote this corrected phase p instead of φ). Although for each single edge in figure 4.4a) or 4.4c) two interpretations for direction are possible we can overcome this ambiguity by taking global relations into account. For example, by assuming both edges as part of the same 3D Gestalt (see figure 4.5c) both assignments of direction have to be coherent.

In the stereo domain, the association of direction to two corresponding 2D line segments implies a 3D direction (see figure 4.5a). On the other hand, a 3D direction implies 2D directions of its projections (see figure 4.5a). We call this relation the *direction uniqueness constraint.*

Another constraint is *Stereo Direction Coherence:* Assuming parallel cameras. If for a line segment l^l in the left image holds $d < 1/2\pi$ or $d > 3/2\pi$ than for the corresponding line segment in the right image must hold the same. Figure 4.5a shows a valid interpretation of direction while figure 4.5b shows a non–valid interpretation of direction).[2]

Both constraints are used in the stereo similarity function described in section 4.3. The stereo direction constraint is used to reduce the number of possible correspondences of directed line segments. The direction uniqueness constraint states that reconstruction of directed 3D line segments is possible. It is one advantage of our similarity function that it associates a direction to line segments, i.e., can be used to overcome the ambiguity of local direction estimation.

[2] In case of non-parallel cameras the constraint reads: If for a line segment l^l in the left image holds that the vector of direction points above the epipolar line in the left image (constituted by the corresponding line segments in both images) than for the vector of direction of the corresponding line segment in the right image holds the same and vice versa.

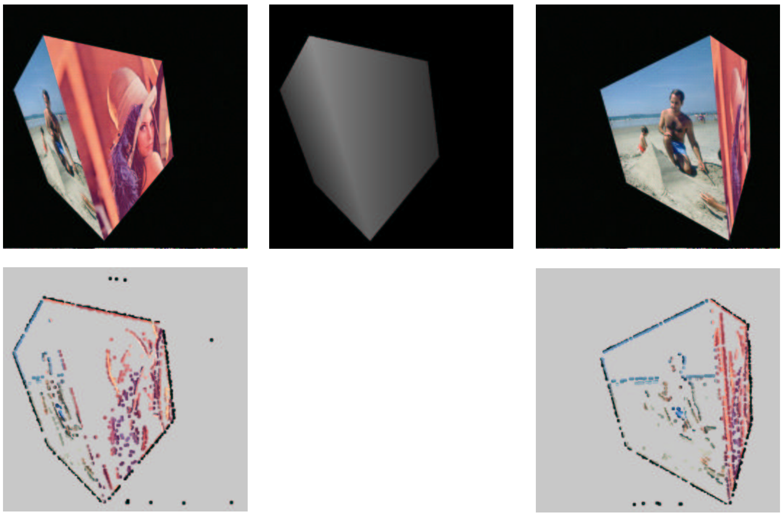


Figure 4.6: Top: Left Image (left), ground truth (middle), right image (right). Bottom: Extracted features in the left and right image.

4.3 A new Stereo Similarity Function

Our basic local features at position $\mathbf{x}$ can now be extended to $E = (\vec{x}, d, p, (\mathbf{c}_l, \mathbf{c}_r))$, $d \in [0, 2\pi]$ representing the direction, $\varphi \in [-\pi, \pi]$ representing the phase ($p = \varphi$ if $d = \theta$ and $p = -\varphi$ if $d = \theta + \pi$), $(\mathbf{c}_l, \mathbf{c}_r), \mathbf{c}_i \in [0,1] \times [0,1] \times [0,1]$ represent the color in RGB space. Since we want to neglect information about the magnitude we ensure $||\mathbf{c}|| = 1$.

A straightforward distance function between two line segments for stereo matching is the weighted sum of differences of the orientation and the structural information associated to the features extracted from the left and right image:

$$\begin{gathered} \mathcal{D}(e^l, e^r) = \\ \alpha(\Delta d) + (1 - \alpha)\left(\beta(\Delta\varphi) + (1 - \beta)(\Delta c)\right). \end{gathered} \tag{4.1}$$

$\alpha \in [0, 1]$ represents the weight for the geometric information compared to structural information, $\beta \in [0, 1]$ represents the weight for phase compared to color information. $\Delta d, \Delta p$ and Δc are all defined such that they take value in $[0, 1]$. Here we want to remark that all values $\Delta d, \Delta\varphi, \Delta c$ are normalized such that they have comparable mean and standard deviation according to [20].

The concept of direction is essential for the definition of structural information, since the structural part switches under a rotation of π (see figure 4.4a) and the two triplets $(\vec{c}_l, \vec{c}_r)$ switch as well). To define a stereo similarity function the concept of direction is essential as well. For instance, the image patches in figure 4.4bi) (if direction is interpreted as indicated) have same structure but opposite direction while the image patches in figure 4.4bii) have same direction but different structure.

However, in images only the orientation is locally measurable only while the structural part switches under the assumed underlying direction (see figure 4.4b and [26]). Although from a more global perspective consistent interpretation of direction could be associated to the line segments (see section 4.2) we can not decide locally which interpretation of direction is appropriate. Therefore, to compare two feature vectors $(d^l, p^l, (\mathbf{c}_l^l, \mathbf{c}_r^l)), (d^r, p^r, (\mathbf{c}_l^r, \mathbf{c}_r^r))$ in the stereo similarity function we have to look at all geometrically possible interpretations of direction for the left and right image patch. We have to deal with four cases:

1) The measured orientation in the left and right image equals the underlying direction: $d^l = \theta^l, d^r = \theta^r$ (see figure 4.4ci).

2) The measured orientation in the left image equals the underlying direction but the measured orientation in the right image is related to the direction by $d^r = \theta^r + \pi$. This also implies that the phase has opposite sign than the locally measured phase for the right image ($p^r = -\varphi^r$) and the color triplets switch as well (see figure 4.4cii).

3) The measured orientation in the left image equals is related to the direction by ($d^l = \theta^l + \pi$) which also implies that the phase has opposite sign than the locally measured phase $p^l = -\varphi^l$ (see figure 4.4ciii) and the color triplets switch as well. The underlying direction in the right image equals the measured orientation.

4) The measured orientation in the left image is associated to the direction $d^l = \theta^l + \pi$ and the measured orientation in the left image is associated to the direction $d^r = \theta^r + \pi$.

The exact definition of the stereo similarity function can be derived by treating and combining these four cases in an appropriate way (for details, see [44]).

4.4 Experiments

In this section we investigate the relative importance of geometrical versus structural information as well as the relative importance of phase versus color. That means we investigate the quality of stereo matching depending on the weights α and β in (4.1).

Measuring performance: To achieve statistically relevant statements about the importance of the different factors of visual information we need a large data base. Since a manual generation of ground truth from natural images is extremely tedious we use images created in a virtual environment by texture mapping with natural images: Natural images are mapped onto a randomly rotated cube (see figure 4.6). This ensures that we deal with data close to natural conditions as well as a known 3D–structure of the scene.

Our measure of performance is the number of 3D–line segments with associated disparity close to the ground truth (in our case we chose a deviation of 3 pixel, $t = 3$) divided by the number of extracted 3D–line segments.

In many stereo algorithms additional constraints are used to improve performance. The *epipolar line constraint* says that the corresponding point to a point in the left image must be on a epipolar line in the right image. This constraint is always valid (see, e.g., [14, 37]). The *uniqueness constraint* states that a 3D–point can not have two distinct projections in an image. This constraint is always valid as well. Other constraints such as *ordering* (i.e., for a point which is left to another point the corresponding points in the right image have to have the same order) and *restrictions on the absolute disparity* are only valid in most circumstances but there exist geometric exceptions.

In our simulations we only make use of the epipolar constraints and the uniqueness constraint but we do not use any kind of further restrictions to improve stereo since our aim is not come up with a system which optimal performance but to investigate the different factors of visual information according to their contribution to stereo matching. We use a set of 40 images. The chance level (i.e., when our similarities are defined randomly) of performance is 30.6%.

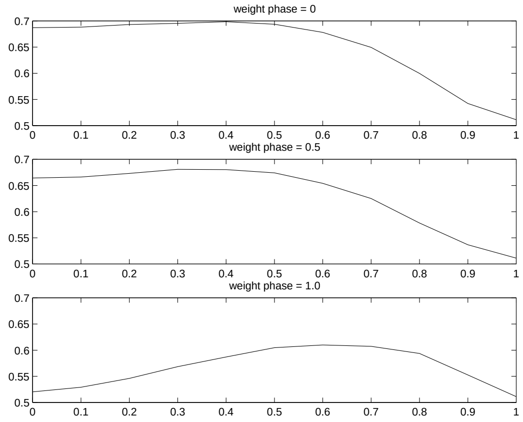


Figure 4.7: Slices of surface shown in figure 4.8. The x-axis represent α. Top: $\beta = 0$. Middle: $\beta = 0.5$, Bottom: $\beta = 1.0$.

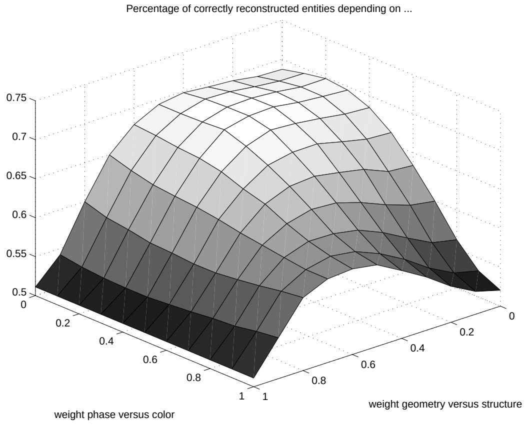


Figure 4.8: Performance for varying α and β. α represents the 'weight geometry versus structure' and β the 'weight phase versus color'.

The performance with a normalized cross correlation[3] on the grey level image is 67.7%, the performance with a normalized cross correlation (by adding the results of correlation in each sub–channel) on the color level image is 68.5% .

Contribution of structural versus geometric information in grey level images: Figure 4.7 (bottom) shows the variation of performance on the test set of 40 images for different α when we set $\beta = 1.0$, i.e., we use only phase and no color. We recognize a peak performance of 61% for $\alpha = 0.6$. We see that optimal performance is achieved *by combining* structural and geometric information (using direction only we get only 51.1% and using phase only we get 52%). Here, the importance of geometric information for stereo matching is slightly higher than for the structural information coded in the phase The performance is lower than for a normalized–cross correlation matching with 10 × 10 patches. However, we achieve a reasonable performance although we *reduce the signal to 2 parameters only (direction and phase) compared to 100 parameters for the normalized cross correlation!* This corresponds to a reduction by a factor of 50.

[3] The comparisons are made at the very same pixel positions for normalized cross correlation than for our new similarity function.

Contribution of structural versus geometric information using also color information: Figure 4.7 (top) shows the variation of performance performance on the test set different α when we set $\beta = 0.0$, i.e., we use color only as structural information. We see again that optimal performance is achieved by combining geometrical and structural information. We also recognize that the structural information coded in color leads to better results than by using phase only.

Figure 4.8 shows the performance when we vary α and β. The plots in figure 4.7 are slices of this figure. We achieve a top performance of 70.5% for $\alpha = 0.3, \beta = 0.3$. Once again we see that the *combination* of structural and geometric information gives optimal performance and we can also recognize that the *combination* of phase and color information gives the best results, i.e., both factors can be used complementary. Our parameterization of color leads to an increase of performance by 9.5% compared to the use of grey level information only. Structural information can now be weighted higher compared to grey level information (0.7 versus 0.4). However, since our similarity function distinguishes between left and right side of the edge some geometric information is coded as well in the color triplets.

For the optimal weights we achieve on our data set an even higher performance (70.5% versus 68.5%) than using normalized cross correlation for color images although we reduce a 100 dimensional image patch to 8 parameters only.

4.5 Conclusion

We have investigated a compact and explicit coding of local image information for intrinsically one–dimensional signals in terms of direction, phase and color attributes. We applied applying this coding within the stereo domain. By making use of the explicit separation of geometric and structural information we could compute the relative importance of the different sub–aspects for stereo processing. We could show that it is the combination of aspects that gives the best results. Moreover, we could show that we can achieve high matching performance although we reduce the image information by a factor of 50 for grey level images. For color images we can achieve an even higher matching performance than with a normalized cross correlation although we reduce the image information by a factor of more than 35. Therefore, this compact coding also promises efficient applicability for tasks that require low storage costs, such as, e.g., object coding and object recognition.

Chapter 5

Statistical Investigations of Primitives in the spatial Domain and its Applications

In this chapter we investigate the statistical interdependencies between the Primitives derived in chapter 3.

A large amount of research has been focused on the usage of Gestalt laws in computer vision systems (overviews are given in [64, 61]). The most often applied and also the most dominant Gestalt principle in natural images is collinearity [12, 42]. Collinearity can be exploited to achieve more robust feature extraction in different domains, such as, edge detection (see, e.g., [27, 31]) or stereo estimation [8, 61]. In most applications in artificial visual systems, the relation between features, i.e., the applied Gestalt principle, has been defined heuristically based on semantic characteristics such as orientation or curvature. Mostly, explicit models of feature interaction have been applied, connected with the introduction of parameters to be estimated beforehand, a problem recognized as extremely awkward in computer vision.

Gestalt principles are affected by multiple modalities. For example, figure 5.1 shows how collinearity can be intensified by the different modalities contrast transition, optic flow and color. This paper addresses statistics of natural images in these modalities. As a main result we found that statistical interdependencies corresponding to the Gestalt law "collinearity" in visual scenes become significantly stronger when multiple modalities are taken into account (see section 5.1). Furthermore, we discuss how these measured interdependencies can be used within artificial visual systems (see section 5.2).

For our statistics we use 9 image sequences with a total of 42 images of size 512x512 (18 images) and 384x288 (24 images). Our data (see figure 5.2 (left) for examples) contains varia-

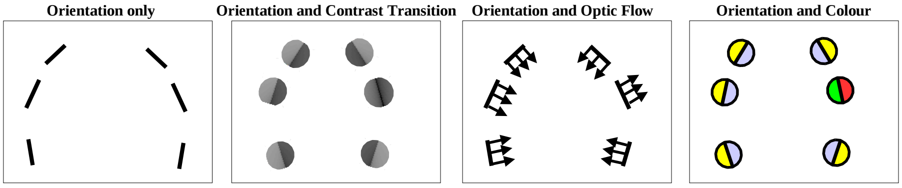

Orientation only
Orientation and Contrast Transition
Orientation and Optic Flow
Orientation and Colour

Figure 5.1: Grouping of visual entities becomes intensified (left triple) or weakened (right triple) by using additional modalities: Since the visual entities are not only collinear but show also similarity in an additional modality their grouping becomes more likely.

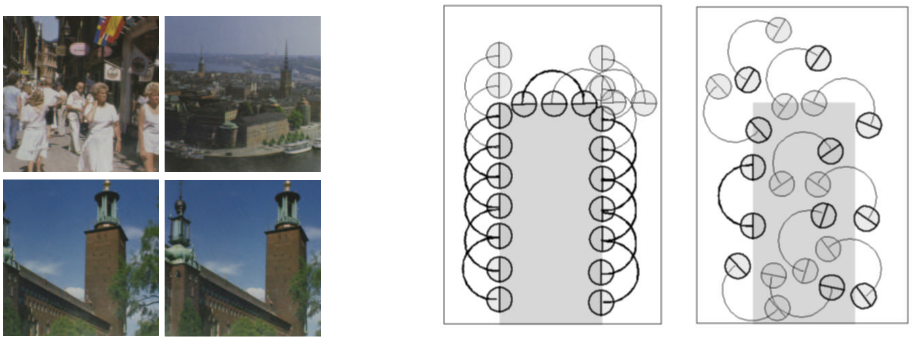


Figure 5.2: **Left:** Images of the data set (top) and 2 images of a sequence (bottom). **Right:** Explanation of the Gestalt coefficient $G(e^1|e^2)$: We define e^2 as the occurrence of a line segment with a certain orientation (anywhere in the image). Let the second order event e^1 be: "occurrence of collinear line segments two units away from an existing line segment e^2". Left: Computation of $P(e^1|e^2)$. All possible occurrences of events e^1 in the image are shown. Bold arcs represent real occurrences of the specific second order relations e^1 whereas arcs in general represent possible occurrences of e^1. In this image we have 17 possible occurrences of collinear line segments two units away from an existing line segment e^2 and 11 real occurrences. Therefore we have $P(e^1|e^2) = 11/17 = 0.64$. Right: Approximation of the probability $P(e^1)$ by a Monte Carlo method. Entities e^2 (bold) are placed randomly in the image and the presence of the event 'occurrence of collinear line segments two units apart of e^2' is evaluated. (In our measurements we used more than a 500000 samples for the estimation of $P(e^1)$). Only in 1 of 11 possible cases this event takes place (bold arc). Therefore we have $P(e^1) = 1/11 = 0.09$ and the Gestalt coefficient for the second order relation is $G(e^1|e^2) = 0.64/0.09 = 7.1$.

tions caused by object motion as well as camera motion. There is a total of 3900 feature vectors in the Data set (approximately 2600 from the outdoor images) and the statistic is based on 1555548 second order comparisons.

5.1 Statistical Interdependencies in Image Sequences

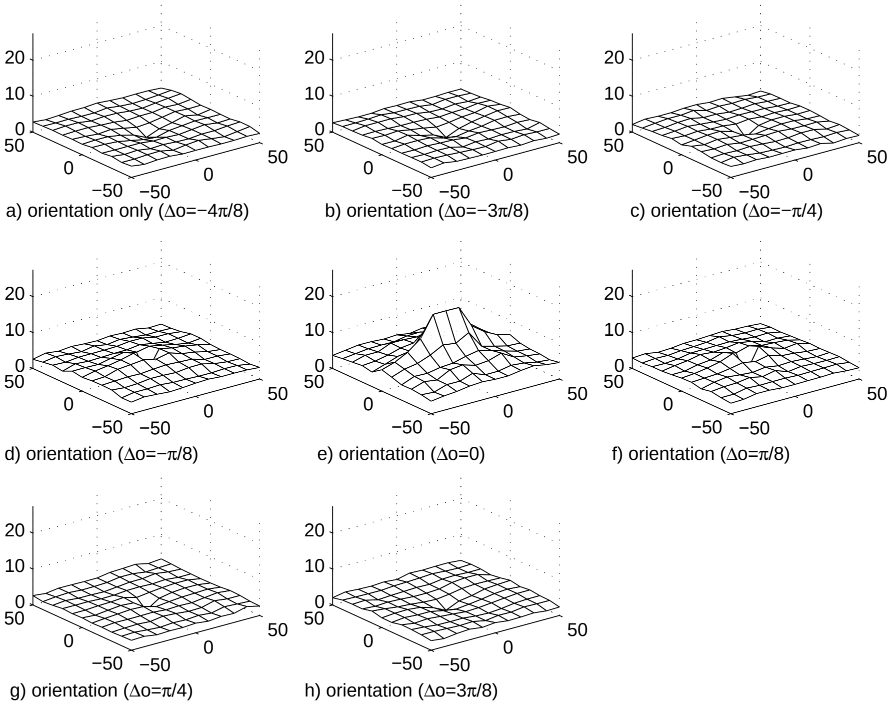
We measure statistical interdependencies between events by a mathematical term that we call 'Gestalt coefficient'. The Gestalt coefficient is defined by the ratio of the likelihood of an event e^1 given another event e^2 and the likelihood of the event e^1:

$$G(e^1, e^2) = \frac{P(e^1|e^2)}{P(e^1).} \tag{5.1}$$

For the modeling of feature interaction a high Gestalt coefficient is helpful since it indicates the modification of likelihood of the event e^1 depending on other events. A Gestalt coefficient of one says, that the event e^2 does not influence the likelihood of the occurrence of the event e^1. A value smaller than one indicates a negative dependency: the occurrence of the event e^2 reduces the likelihood that e^1 occurs. A value larger than one indicates a positive dependency: the occurrence of the event e^2 increases the likelihood that e^1 occurs. The Gestalt coefficient is illustrated in figure 5.2(right). Further details can be found in [48].

5.1.1 Second Order Relations Statistics of Natural Images

A large amount of work has addressed the question of efficient coding of visual information and its relation to the statistics of images. Excellent overviews are given in [74, 69]. While many



a) orientation only (Δo=−4π/8)

b) orientation (Δo=−3π/8)

c) orientation (Δo=−π/4)

d) orientation (Δo=−π/8)

e) orientation (Δo=0)

f) orientation (Δo=π/8)

g) orientation (Δo=π/4)

h) orientation (Δo=3π/8)

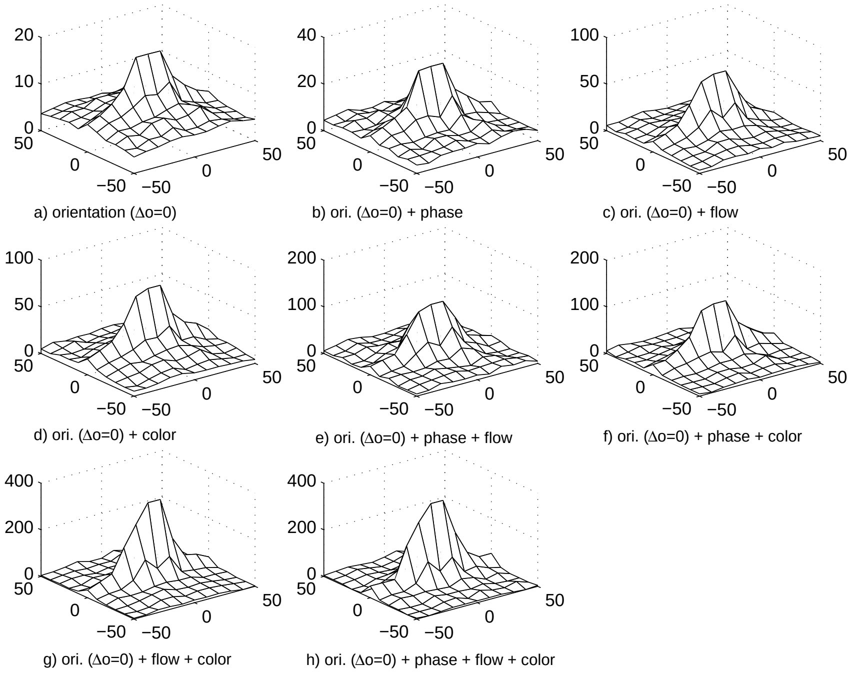
Figure 5.3: The Gestalt coefficient for differences in position from -50 to 50 pixel in x– and y– direction when orientation only is regarded. In a) the difference of orientation of the line segments is $\frac{\pi}{2}$ (the line segments are orthogonal) while in e) the difference of orientation is 0, i.e., the line segments have same orientation. The b), c), d) represent orientation difference between $\frac{\pi}{2}$ and 0. Note that the Gestalt coefficient for position (0,0) and $\Delta\theta = 0$ is set to the maximum of the surface for better display. The Gestalt coefficient is not interesting at this position, since e^1 and e^2 are identical.

publications were concerned with the statistics on the pixel level and the derivation of filters from natural images by coding principles (see, e.g. [55, 5]), recently statistical investigation in the feature space of local line segments have been performed (see, e.g., [42, 12, 22]) and have addressed the representation of Gestalt principles in visual data.

Here we go one step further by investigating the second order relations of events in our *multi–modal feature space* (see chapter 3)

$$E = (\vec{x}, \theta, \varphi, \mathbf{c}_l, \mathbf{c}_r, \mathbf{o}).$$

In our measurements we collect second order events in bins defined by small patches in the (x_1, x_2)–space and by regions in the modality–spaces defined by the metrics defined for each modality (for details see [48]). Figure 5.3 shows the Gestalt coefficient for equidistantly separated bins (one bin corresponds to a square of 10 × 10 pixels and an angle of $\frac{\pi}{8}$ rad). As already been shown in [42, 22] collinearity can be detected as significant second order relation as a ridge in the surface plot for $\Delta\theta = 0$ in figure 5.3e. Also parallelism is detectable as an offset of this surface. A Gestalt coefficient significantly above one can also be detected for small orientation differences (figure 5.3d,f, i.e., $\Delta\theta = -\frac{\pi}{8}$ and $\Delta\theta = \frac{\pi}{8}$) corresponding to the frequent occurrence of curved entities. The general shape of surfaces is similar in all following measurements concerned with additional modalities: *we find a ridge corresponding to collinearity and an offset corresponding to parallelism and a Gestalt coefficient close to one for*



a) orientation (Δo=0)
b) ori. (Δo=0) + phase
c) ori. (Δo=0) + flow
d) ori. (Δo=0) + color
e) ori. (Δo=0) + phase + flow
f) ori. (Δo=0) + phase + color
g) ori. (Δo=0) + flow + color
h) ori. (Δo=0) + phase + flow + color

Figure 5.4: The Gestalt coefficient for $\Delta o = 0$ and all possible combination of modalities.

all larger orientation differences. Therefore, in the following we will only look at the surface plots for equal orientation $\Delta o = 0$. These result shows that Gestalt laws are reflected in the statistics of natural images: *Collinearity and parallelism are significant second order events of visual low level filters* (see also [42]).

5.1.2 Pronounced Interdependencies by using additional Modalities

Now we can look at the Gestalt coefficient when we also take into account the modalities contrast transition, optic flow and color.

Orientation and Contrast Transition: We say two events $(\mathbf{x}, \theta)$ and $(\mathbf{x}', \theta')$ have similar contrast transition (i.e., 'similar phase') when $d(\varphi, \varphi') < t^{\varphi +}$. The metrics in the different modalities are precisely defined in [48]. $t^{\varphi}+$ is defined such that only 10% of the comparisons $d(\varphi, \varphi')$ in the data set are smaller than t^{φ}. Figure 5.4b shows the Gestalt coefficient for the events 'similar orientation and similar contrast transition'. In figure 5.5 the Gestalt coefficient along the x-axes in the surface plot of figure 5.4 is shown. The Gestalt coefficient on the x-axes correspond to the 'collinearity ridge'. The first column represents the Gestalt coefficient when we look at similar orientation only, while the second columns represent the Gestalt coefficient when we look at similar orientation and similar phase. We see a significant increase of the Gestalt coefficient compared to the case when we look at orientation only for collinearity.

This result shows that assuming a line segment with a certain contrast transition does exist in an image it is not only that the likelihood for the existence of a collinear line segment increases but that it also becomes more likely that it has similar contrast transition.

Orientation and Optic Flow: The corresponding surface plot is shown in figure 5.4c and the slice corresponding to collinearity is shown in the third column in figure 5.5. An even more pronounced increase of inferential power for collinearity can be detected.

Orientation and Color: Analogously, we define that two events have 'similar color structure'. The corresponding surface plot is shown in figure 5.4d and the slice corresponding to collinearity

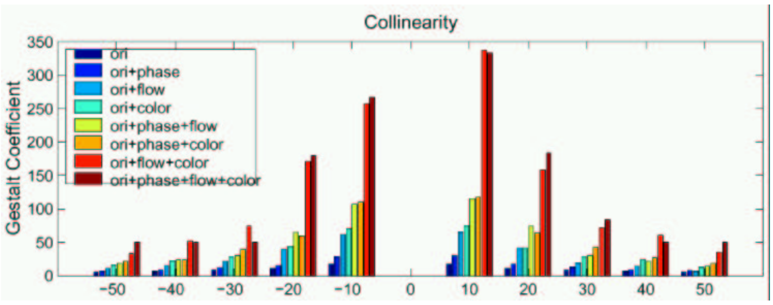


Figure 5.5: The Gestalt coefficient for collinear feature vectors for all combinations of modalities. For (0,0) the Gestalt coefficient is not shown, since e^1 and e^2 would be identical.

is shown in the fourth column in figure 5.5.

Multiple additional Modalities: Figure 5.4 shows the surface for similar orientation, phase and optic flow (figure 5.4e); similar orientation, phase and color (figure 5.4f) and similar orientation, optic flow and color (figure 5.4g). The slices corresponding to collinearity are shown in the fifth to seventh columns in figure 5.5. We can see that the the Gestalt coefficient for collinear line segments increases significantly. Most distinctly for the combination optic flow and color (seventh column). Finally, we can look at the Gestalt coefficient when we take all three modalities into account. Figure 5.4h and the eighth column in figure 5.5 shows the results. Again an increase of the Gestalt coefficient compared to the case when we look at only two additional modalities can be achieved.

5.2 Summary and Examples of Possible Applications

In this paper we have addressed the statistics of local oriented line segments derived from natural scenes by adding information to the line segment concerning the modalities contrast transition, color, and optic flow. We could show that statistical interdependencies in the orientation–position domain correspond to the Gestalt laws collinearity and parallelism and that they become significantly stronger when multiple modalities are taken into account. Essentially it seems that visual information bears a high degree of intrinsic redundancy. This redundancy can be used to reduce the ambiguity of local feature processing.

The results presented here provide further evidence for the assumption that despite the vagueness of low level processes stability can be achieved by *integration of information across modalities*. In addition, the attempt to model the application of Gestalt laws based on statistical measurements, as suggested recently by some researchers (see, [22, 12, 42, 67]) gets further support. Most importantly, the results derived in this paper suggest to formulate the application of Gestalt principles in a multi-modal way.

Illusionary contour processing (in which the Gestalt law 'collinearity' is tightly involved) occurs at a late stage (after approximately 6 months) during the development of the human visual system (see [6] and [47]). This late development of the above mentioned mechanisms makes it likely that those mechanisms depend on visual experience of the underlying structures in visual data. This also suggests a formalization of Gestalt laws in artificial systems depending on statistical measurements.

Motivated by the measurable reflectance of Gestalt principles in the statistics of natural images (as shown in this paper) and the late development of abilities in which these Gestalt principles are involved, it is our aim to replace heuristic definition of Gestalt rules by interaction

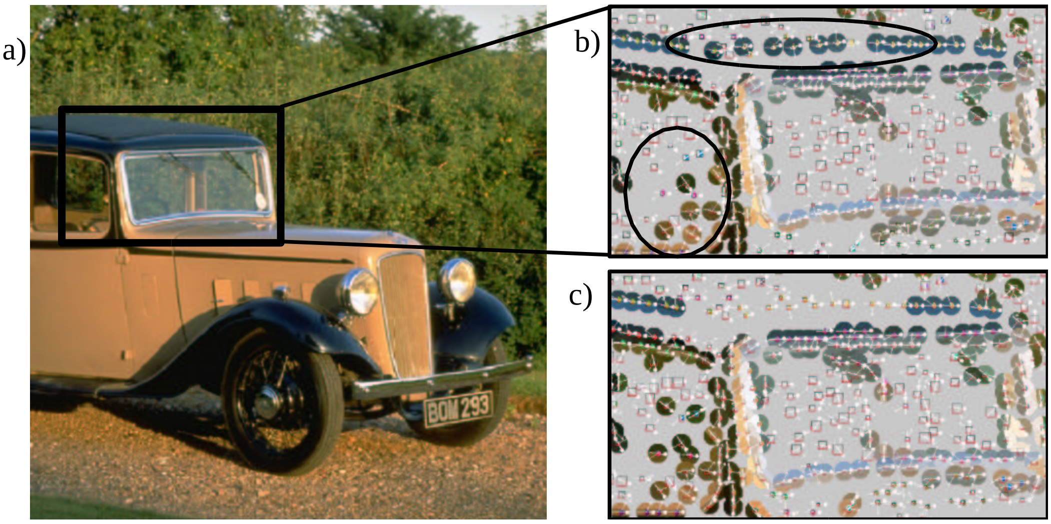


Figure 5.6: Left: Image of a car. Right top: Extraction of Primitives with grouping based on the Gestalt coefficient. Right bottom: Primitives extracted without grouping.

schemes based on statistical measurements. We want to describe two examples: A process of self–emergence of feature constellations and low–contrast edge detection. In both cases only a simple criterion based on the Gestalt coefficient is applied to realize the collinear relation.

Self–Emergence of Feature Constellations: The need of entities going beyond local oriented edges is widely accepted in computer vision systems across a wide range of different viewpoints. Their role is to extract from the complex distribution of pixels in an image patch (or an image patch sequence) a sparse and higher semantical representation which enables rich predictions across modalities, spatial distances and frames. Accordingly, they consist of groups of early visual features (such as local edges).

These higher feature constellations have been already applied in artificial systems but were needed to be defined heuristically. By using a link criterion based on the Gestalt coefficient (stating that there exist a link when the Gestalt coefficient is high) and the transitivity relation (if two pairs of entities are linked then all entities have to be linked) we are able to define a process in which groups of local entities do self emerge. Groups are coded by Primitives with very same color in their center (see figure 5.7.

Detection of low Contrast Edges Once the groups have self–emerged they can be used to detect low contrast edges and reduce falsely detected edges caused by structural noise *by modifying confidences of entities within the group an entity belongs to.* In figure 5.6b and c all Primitives above the very same threshold are displayed with a filled circle while features below this threshold are displayed without these circles. Note the detection of the low contrast edge (figure 5.6b, horizontal ellipse) when applying grouping based on the Gestalt coefficient and the reduction of non–meaningful features (vertical ellipse) without grouping.

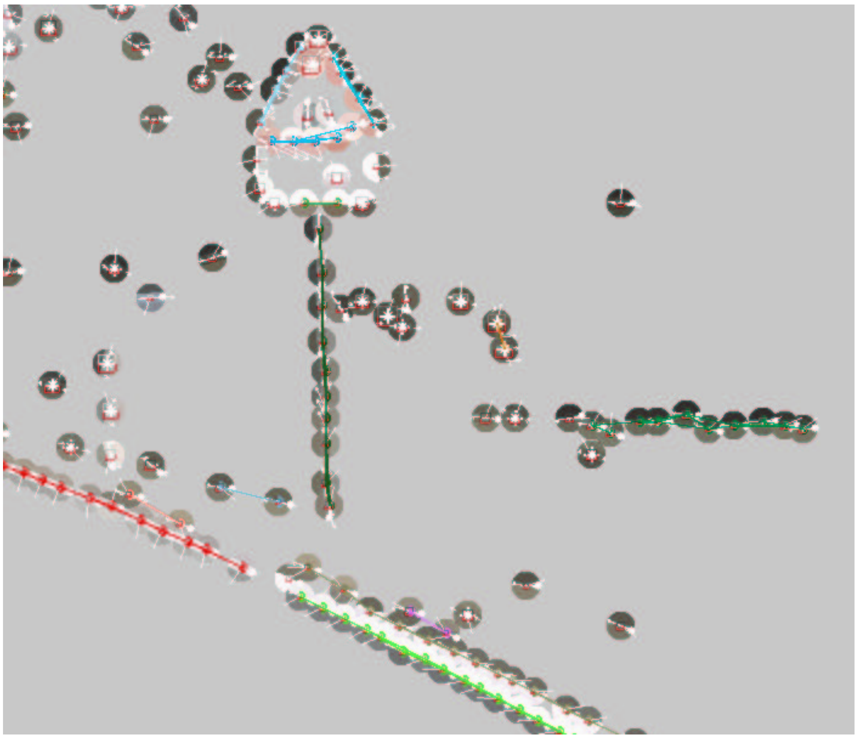


Figure 5.7: Grouping of Primitives.

Chapter 6

Utilizing Stereo and Motion to stabilize Primitives

In this chapter we use the temporal interdependency RBM to stabilize the 3D Primitives derived in chapter 3 and 4. It turns out that a 3D collinearity criterion is essential for a proper estimation of RBM. Therefore, we found a strong dependency between temporal and spatial Gestalts. In this chapter we already start to address aspects of the work packages 3.2 and 3.3.

Note that in this chapter we presuppose that image changes are only due to one motion. However, after segmenting the scene into different regions corresponding to different motions (see chapter 8) the method described here can also be applied to multiple motions.

6.1 Motion and Statistical Interdependencies

Two important regularities in visual data with distinct properties are *motion* (most importantly rigid body motion, RBM, see, e.g., [14]) and *statistical interdependencies* between features such as collinearity and symmetry (see, e.g., [64]).[1] RBM reflects a geometric dependency in the time-space continuum. If the 3D motion between two frames is known then feature prediction can be postulated since the change of the position and the semantic properties of features can be computed (see, e.g., [43]). This can be done by having physical control over the object (as in [43]) or by *computing the RBM* as done in this paper. A computation of the RBM makes the system more flexible since it allows for acquiring object knowledge by watching the object without grasping is. This is also one of the main contributions of this paper compared with [43]. However, having physical control over the object might also have advantageous in specific situation, e.g., when the RBM is controlled in such a way that especially cognitive interesting situation are created.

However, computation of RBM is a non trivial problem. A huge amount of literature is concerned with its estimation from different kinds of feature correspondences (see, e.g., [59, 63]), which are most commonly point and/or line correspondences. In our system, correspondences are established by optic flow. However, one fundamental problem of RBM-estimation is that methods are in general very sensitive to outliers. The pose estimation algorithm we do apply [63] computes the rigid body motion presupposing a 3D model of the object and a number of correspondences of 3D–entities between the object model and their projections in

[1] There exists evidence that abilities based on rigid body motion are to a much higher degree hard coded in the human visual system than abilities based on statistical interdependencies (for a detailed discussion see [47]).

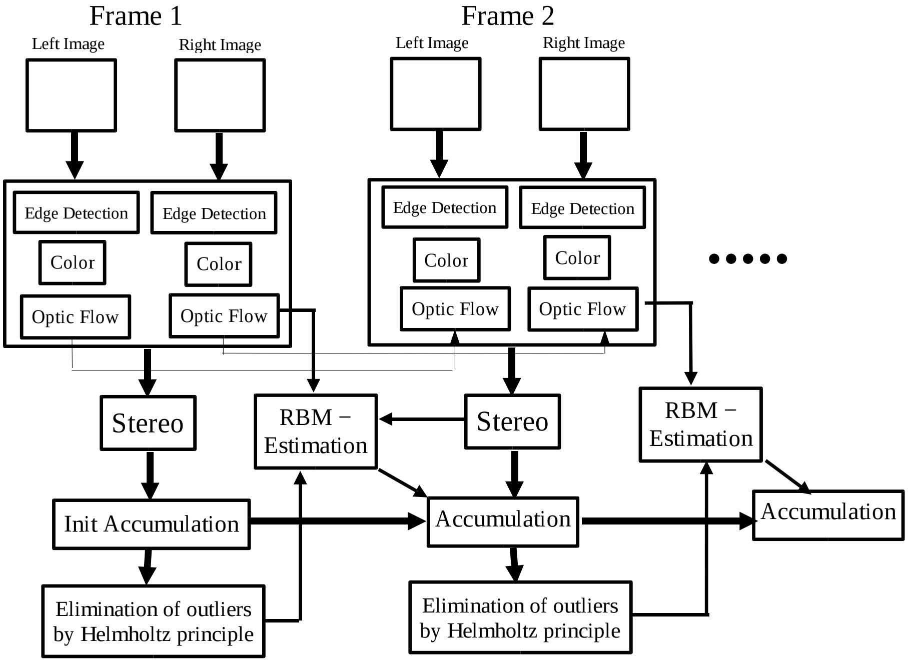


Figure 6.1: Scheme of Interaction of visual sub–modalities

the consecutive frame. In [63] a manually designed 3D model was used for pose estimation. Here, we want to replace this prior knowledge by substituting the manually created model by 3D information extracted from stereo. However, by using stereo we face the above mentioned problems of uncertainty and reliability of visual data as described above. Because of the sensitivity of pose estimation to outliers in the 3D–model we need to compensate these disturbances. We can sort out unreliable 3D–features by applying a grouping mechanism based on *statistical interdependencies* in visual data.

Once the RBM across frames is known (and for the computation of the RBM we need a quite sophisticated machinery) we can utilize and a scheme which uses the *deterministic regularity* RBM to disambiguate 3D entities over consecutive frames [43].

6.2 Visual Sub-modalities

Our system acquires stable representations from stereo image sequences. An overview of the system is given in figure 6.1.

At this point, we want to stress the difference between two different sources of disturbances:

- Outliers: 3D entities caused by wrong stereo correspondences. They have an irregular non–Gaussian distribution (see figure 6.2 (top row))

- Feature inaccuracy: Deviation of parameters of estimated 3D entities (e.g., 3D orientation and 3D position) caused by unreliable position and orientation estimates in images. This kind of disturbance can be expected to have Gaussian like distribution with its mean close to the true value.

Both kinds of disturbances have distinct distribution and the visual modules have a different sensitivity to both errors: for example, while outliers can lead to a completely wrong estimation of pose, feature inaccuracy would not distort the results of pose estimation that seriously.

We will deal with these two kinds of disturbances in distinct ways: *Outliers* are sorted out by a filtering algorithm utilizing the statistical interdependency "collinearity" in 3D and by a process of recurrent predictions based on rigid body motion estimation. Both processes *modify confidences* associated to Primitives. *Feature inaccuracy* becomes reduced by *merging* corresponding 3D line segments over consecutive frames. During the merging process semantic parameter (here 3D–position and 3D–orientation) are iteratively adapted.

In the following we briefly introduce the applied sub–modalities and their specific role within the whole system.

Stereo: In stereo processing with calibrated cameras we can reconstruct 3D points from two corresponding 2D points by computing the point of intersection of the two projective lines generated by the corresponding image points and the optical centers of the camera. However, most meaningful image structure is intrinsically one-dimensional [73], i.e., is dominated by edges or lines. Orientation at intrinsically one-dimensional image structures can be estimated robustly and precisely by various methods (see, e.g., [35]). Therefore, it makes sense to use orientation information also for the representation of visual scenes: from two corresponding 2D points with associated orientation we can reconstruct a 3D point with associated 3D orientation (in the following called '3D–line segment'). A more detailed description can be found in [43].

To find stereo correspondences in the left and right image we can use geometrical as well as structural information in form of phase and color. In [44] we can show that both factors are important for stereo–matching and that the optimal result is achieved by *combining* both kinds of information.

Note that out stereo algorithm does not make use of the ordering constraint (such as , e.g., in [60]) but that the system can also be used in case of depth continuities. The only restriction is the occurrence of local line segments in the scene.

The basic feature we extract from the stereo module is a reduction of our 3D Primitives to 3D line segment coded by its mid–point $\mathbf{X}$ and its 3D orientation coded by two parameter $\mathbf{O}$. Furthermore a confidence c is associated to the parametric description of the 3D entity. We therefore can formalize a 3D–line segment by

$$\mathbf{l} = (\mathbf{X}, \mathbf{O}; c). \tag{6.1}$$

All parameters are subject to modifications by contextual information (as described below) utilizing the Gestalt law Collinearity and the regularity RBM across frames.

Pose estimation: To be able to predict 3D–features in consecutive frames, we want to track an object in a stereo image sequence. More precisely, we want to find the rigid body motion from one frame to the consecutive frame. To compute the rigid body motion we apply the pose estimation algorithm [25] which requires a 3D model of the object as well as correspondences of image entities (e.g., 2D line segments) with 3D object entities (e.g., 3D line segments)[2]. A 2D–3D line correspondence defines a constraint on the set of possible rigid body motions that (using a linear approximation of a rigid body motion [25]) can be expressed by two linear equations.

[2] This pose estimation algorithm has the nice property that it can combine different kinds of correspondences (e.g., 3D point–3D point, 2D point–3D point, and 2D line–3D line correspondences) within one system of equations. This flexible use of correspondences makes it especially attractive for sophisticated vision systems which process multiple kinds of features such as 2D junctions, 2D line segments or 3D points.

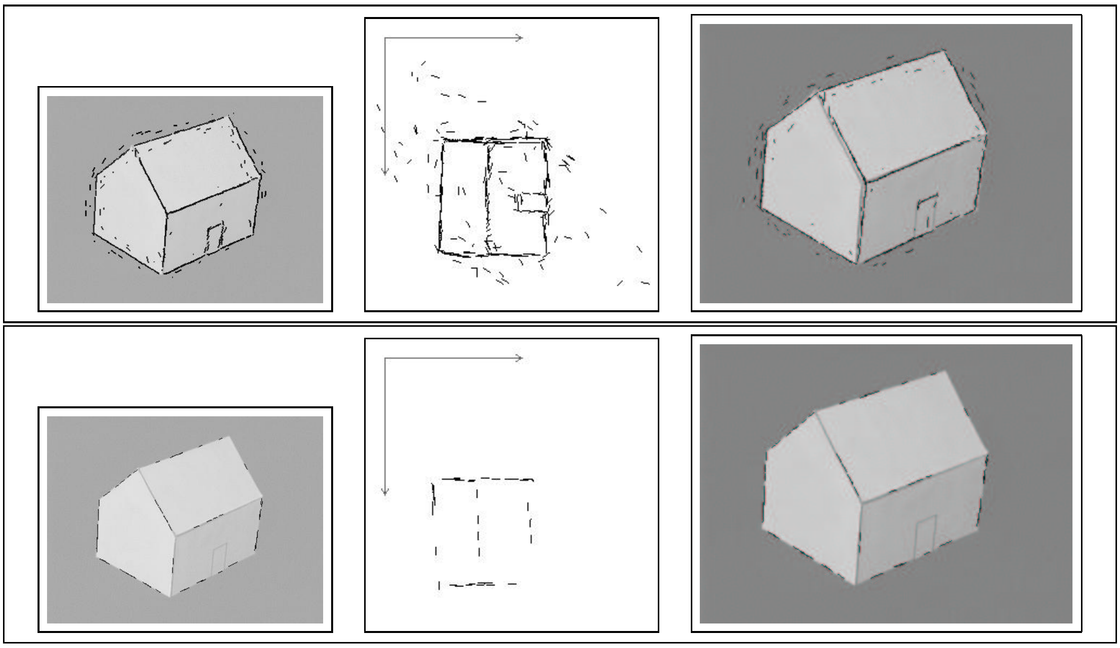


Figure 6.2: **Top:** Using the stereo module without ellimination procedure. Left: Projection onto the image. Middle: Projection onto the xz plane. Note the large number of outliers. Right: Pose estimation with this representation. Note the deviation of pose from the correct position. **Bottom:** The same after the elimination process. Note that all outliers could be eliminated by our collinearity criterion and that pose estimation does improve.

In combination with other constraints we get a set of linear equations for which a good solution can be found iteratively [25] using a standard least square optimization algorithm.

Optic flow: The 3D model of the object is extracted by our stereo algorithm. Correspondences between 3D entities (more precisely by their 2D projections respectively) and 2D line segments in the consecutive frame are found by the optic flow. After some tests with different optic flow algorithms (see [34]) we have chosen the algorithm developed by Nagel [54] which showed good results especially at intrinsically 1D structures. Correspondences are established by simply moving a local line segment according to its associated optic flow vector.

Using collinearity in 3D to eliminate outliers: The pose estimation algorithm is sensitive to outliers since these outliers can dominate the over–all error in the objective function associated with the equations established by the geometric constraints. We therefore have to ensure that no outliers are used for the pose estimation.

According to the *Helmholtz Principle*, every large deviation from a "uniform noise" image should be perceivable, provided this large deviation corresponds to an *a priori* fixed list of geometric structures (see [10]). The *a priori* geometric structure we do apply to eliminate wrong 3D–correspondences are *collinear structures in 3D*: We assume that (according to the Helmholtz principle) a local 3D line segment that has many neighbouring collinear 3D line segments is very unlikely to be an outliers and we only use those line segments for which we find at least a couple of collinear neighbours. More precisely, we lower the confidence c in (6.1) for all line segments that have only few collinear neighbours. Figure 6.2 (middle) shows the results of the elimination process for a certain stereo image). We can show that the elimination

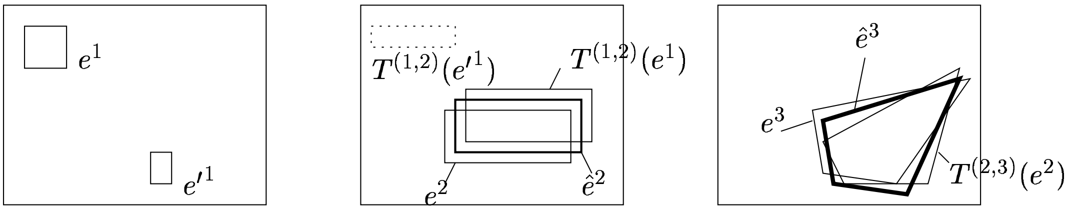


Figure 6.3: The accumulation scheme. The entity e^1 (here represented as a square) is transformed to $T^{(1,2)}(e^1)$. Note that without this transformation it is barely possible to find a correspondence between the entities e^1 and e^2 because the entities show significant differences in appearance and position. Here a correspondence between $T^{(1,2)}(e^1)$ and e^2 is found because a similar square can be found close to $T^{(1,2)}(e^1)$ and both entities are merged to the entity $\hat{e}^2$. The confidence assigned to $\hat{e}^2$ is set to a higher value than the confidence assigned to e^1 indicated by the width of the lines of the square. In contrast, the confidence assigned to e'^1 is decreased because no correspondence in the second frame is found. The same procedure is then applied for the next frame for which again a correspondence for e^1 has been found while no correspondence for e'^1 could be found. The confidence assigned to e^1 is increased once again while the confidence assigned to e'^1 is once again decreased (the entity has disappeared). By this scheme information can be accumulated to achieve robust representations.

process improves pose estimation (see figure 6.2 (right). For a more in depth discussion about applying Gestalt principles within our system see [48].

Acquisition of object representations across frames: Having extracted a 3D representation by the stereo module and having estimated the RBM between two frames we can apply an accumulation scheme (for details see [43]) which uses correspondences across frames to accumulate confidences for visual entities. Our accumulation scheme is of a rather general nature. Confidences associated to visual entities are increased when correspondences over consecutive frames are found and decreased if that is not the case. By this scheme, only entities which are validated over a larger number of frames (or for which predictions are often fulfilled) are considered as existent while outliers can be detected by low confidences (in Figure 6.3 a schematic representation of the algorithm for two iterations is shown). Since the change of features under an RBM can be computed explicitly (e.g., the transformation of the square to the rectangle from the first to the second frame), the rigid body motion can be used to predict the correspondences (see also [43]).

This accumulation scheme presupposes a metrical organisation of the feature space. If we want to compare visual entities derived from two frames even when we know the exact transformation corresponding to the rigid body motion, the corresponding entities cannot be expected to be exactly the same (the two squares in figure 6.3 are only similar not equal) because of factors such as noise during the image acquisition, changing illumination, non–Lambertian surfaces or discretization errors. Therefore it is advantageous to formalize a measure for the likelihood of correspondence by using a metric (for details see [43]). Once a correspondence is established we apply an update rule on the confidence c as well as the semantic parameters (x_1, x_2, x_3) and (θ, ϕ). for the confidence and semantic properties of the line segment (for details see [43] and [34]). That means that by the accumulation scheme our 3D line segments are embedded in the time domain, *they represent features in 3D-space and time.*

Integration of visual sub-modalities: The recurrent process based on the sub-modalities described above is organised as shown in figure 6.1. For each frame we perform feature extraction (edge detection, optic flow) in the left and right image. Then we apply the stereo algorithm and the elimination process based on the Helmholtz principle. Using the improved accumulated model (i.e., after eliminating outliers), we apply the pose estimation module which uses the stereo as well as the optic flow information. Once the correct pose is computed, i.e., the RBM between the frames is known we transform the 3D entities extracted from one frame to the consecutive frame based on the known RBM (for details see [43]).Then we are able to perform one further iteration of the accumulation scheme.

We have applied our system to different image sequences, one of them is shown in figure 3.1. Figure 6.4 (left) shows the results. At the top the extracted stereo representation at the first frame is shown while at the bottom the accumulated representation after 6 frame is shown. We see that the number of outliers can be reduced significantly. In figure 6.4 (right) the mean difference of the semantic parameters (3D–position and 3D–orientation) from a ground truth (manually measured beforehand) is shown. We see that the difference between the extracted representation (consisting of line segments with high confidence) compared to the ground truth for position and orientation decreases during accumulation. Further simulations can be found in [34].

6.3 Summary and Discussion

We have shown that through integration of different visual modalities we are able to extract reliable object representation from disturbed low level processes. Since we want to make use of the regularity RBM across frames we need to use a complex mechanism (which uses different sub–modules) that allows to compute the RBM. This mechanism also made use of statistical regularities to elliminate outliers for pose estimation. Our feature representation allows for a modification of features depending on contextual information. The confidence c codes the likelihood of the existence of the visual entity while semantic parameters describe properties of the entity. Both kind of descriptors are subject to modification by contextual information, i.e., by the statistical and deterministic regularities coded within the system.

Our system has some interesting properties compared to other systems. Firstly, differing to classical structure from motion approaches (see, e.g., [37, 14]) we do not intend to acquire 3D–information only but we are interested in attributes that are relevant for perceptive tasks. Here, our representations consist of line segments. We use these representation for the task of tracking of objects. Representations based on line segments have also been used for object recognition (see, e.g., [46]). We have applied our accumulation scheme to geometric 3D entities. However, this scheme is generic and we intend to apply our accumulation scheme to other visual domains (such as color, texture or other appearance based information) to extract richer and more powerful object representations.

Secondly, in our representations semantic properties of features and their reliability are explicitly coded. Both, semantic properties and the reliability are subject to contextual influences. The integration of contextual information and its modelling by recurrent processes that modify reliabilities is the central aim of our current project and our method differs in that respect to classical structure from motion algorithms (see, e.g., [28]). Here, we have used the reliability information to improve the RBM estimation by picking out certain 'good' feature constellations from the big feature pool. This way to handle outliers works complementary to other well established methods such as RANSAC ([7]). Coding information with its reliability allows to keep hypotheses that are (looking at them locally) unlikely but may become likely taking the

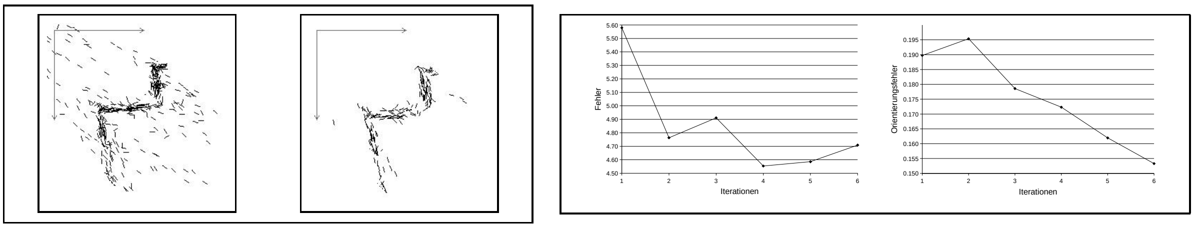


Figure 6.4: **Left:** Projection of representations (extracted from the image sequence shown in figure 3.1) on xz-plane at the beginning of accumulation (left) and after 6 steps of accumulation (right). **Right:** Deviation of 3D position (left) and 3D orientation (right) from the ground truth during accumulation. Estimation of both semantic parameters improves during accumulation.

context into account.

Part B

Kalman-based context sensitive filters based on deterministic spatial Gestalts

Chapter 7

Basic ideas

7.1 Objectives

To develop models of adaptive systems for 3D dynamic scene understanding. Such systems will rely upon two perceptual channels: motion, and stereo.

They should have the following features:

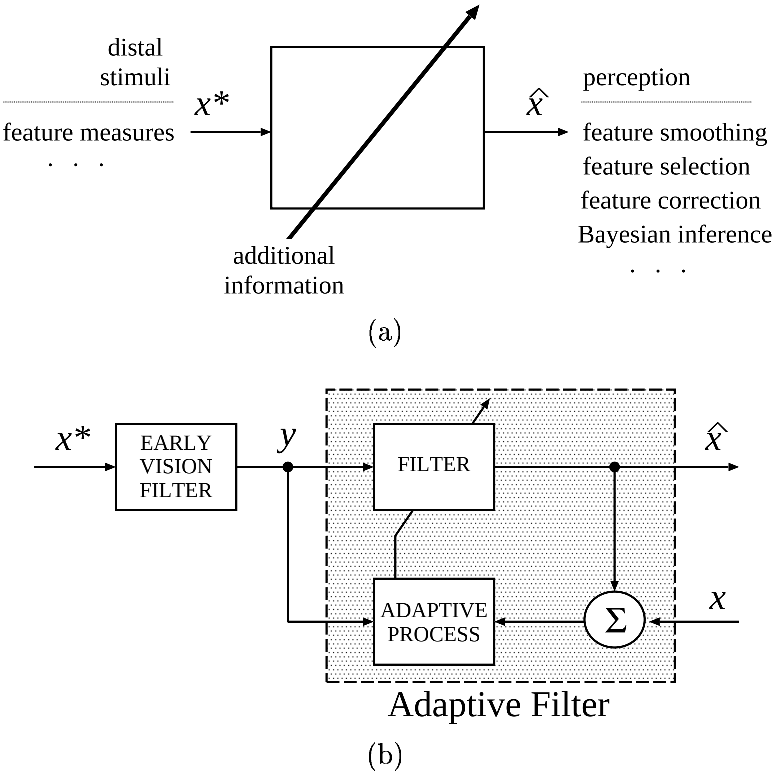
- to be adaptive with respect to the visual context. This would imply that they change their response in relation with the statistical/structural properties of the visual signal;
- to embed such context-sensitive characteristic in recurrent interconnection architectures. In this perspective "context" = "responses of nearby cells".

7.2 Approach

7.2.1 Adaptive filtering

Perception can be viewed as an inference process to gather properties of real-world, or *distal*, stimuli (e.g., an object in space) given the observations of *proximal* stimuli (e.g., the object's retinal image). The distinction between proximal stimulus and distal stimulus touches on something fundamental to sensory processes and perception. The proximal stimulus, not the distal stimulus, actually sets the receptors' responses in motion. Considering the ill posedness of such (inverse) problem, one should include (*a priori*) constraints to reduce the dimension of the allowable solutions, or, in other terms, to reduce the uncertainty on visual measures. These considerations apply both if one tackles the problem of interpretation (understanding) as a whole, and if one considers the confidence on single feature measurements (see Fig. 7.1a). Each measure of an observable property of the stimulus is indeed affected by an uncertainty (not only due to the additive neural noise, but also to the fact that the visual properties are themselves random processes) that can be removed, or, better, reduced by making use of additional information (context information, *a priori* knowledge, etc.). *Early cognitive vision* can be related to that segment of perceptual vision that takes care of reducing the uncertainty on visual measures by capturing coherent properties over large (overlapping) retinal locations (Gestalts), a step that precedes the true understanding of the scene.

In this perspective, early cognitive vision can be casted as an *adaptive filter* in which some kind of early-cognitive algorithm plays the role of the *adaptive process.* A general adaptive filtering system is shown in Fig. 7.1b, where $\boldsymbol{x}^*[k]$ is the *unknown* stimulus (the *state*) at time



(a)

(b)

Figure 7.1: Schematic representation of adaptive early vision filters.

step k, $\boldsymbol{y}[k] = A\boldsymbol{x}^*[k]$ is the obervation of the stimulus (the *measure*), $\hat{\boldsymbol{x}}[k]$ is the estimated stimulus, and $\boldsymbol{x}[k]$ is the reference signal (i.e., what we know about $\boldsymbol{x}^*[k]$). The purpose of a general adaptive system is to filter the input signal $\boldsymbol{y}[k]$ (measure) to invert (in some sense) the measure operator and gain an estimation of the solution of $\boldsymbol{x}^*[k] = A^{-1}\boldsymbol{y}[k]$ by making use of the knowledge $\boldsymbol{x}[k]$.

Remarks:

- The filter evolves in time. It takes some time (convergence time) to the filter to "learn" to act as the inverse operator A^{-1}, by embedding information about the unknown stimulus in its structure (cf. the learning phase of classical neural networks). After convergence, if the reference signal is sufficiently representative of the unknown signal the filter could work in an open-loop configuration (without further adaptation) on the basis of what it has already known of the model. In this condition, new measures are not further used to refine the estimation that will be considered as an *a priori* estimate. The closed-loop configuration, on the contrary, guarantees a continuous adaptation leading to *a posteriori* state estimates. In the following, we will see that this distinction between *a priori* and *a posteriori* estimates will be more evident in the formulation of the Kalman filter due to its intrincic recurrent nature by which the *a priori* estimation (based on the previous experience) is corrected by actual measure to give rise to the *a posteriori* estimate.

- By observing the adaptive filter block we observe that it is characterized by two inputs: the measures $\boldsymbol{y}$ and the reference signal $\boldsymbol{x}$ (the model). Depending on the application of interest or on the modeling abstraction level, different signals can be considered as outputs: apart from the adaptive filter output $\hat{\boldsymbol{x}}$, several hidden signals can be tapped from the adaptive process block. Their physical interpretation will depend on the specific architecture of the adaptive filter implemented.

- We have stated that the reference signal represents what we know about the signal x^*; in general such a knowledge can be provided by:

  1. the *context*
     - properties of the spatial neighborhood (e.g., responses outside the classical receptive field) [spatial context]
     - (spatially local-) temporal properties (cf. the contraints posed by rigid body motion) [(spatio-)temporal context]
     - punctual/local stimulus properties with respect to another modality [multimodal context]
  2. the *state of the perceptual agent* (e.g., alert state, task dependency, expectation, etc.)
  3. *a priori* information (e.g., cognitive models such as shading, familiarity, perspective, etc.)

  In the framework of WP3 we focus only on data-driven (bottom-up) (exogenous) information provided by the context, disregarding the other two model-driven (top-down) (endogenous) components.

7.2.2 Kalman filter

Basic concepts

The Kalman Filter (KF) is an optimal recursive linear estimator, in the sense that it can iteratively process new measures as they arrive, on the basis of the knowledge about the system accrued by previous measurements. Accordingly, a recursive process equation is required to describe the reference signal (the model). Due to its recurrent formalization it appears particularly promising to design *context-sensitive filters* (CSFs) based on recurrent cortical-like interconnection architectures.

Formally, the two inputs to the filter are:

the *process equation*

$$\boldsymbol{x}[k] = \boldsymbol{\Phi}[k, k-1]\, \boldsymbol{x}[k-1] + \boldsymbol{S}[k-1]\, \boldsymbol{s}[k-1] + \boldsymbol{n}_1[k-1] \tag{7.1}$$

and

the *measurement equation*

$$\boldsymbol{y}[k] = \boldsymbol{C}[k]\, \boldsymbol{x}[k] + \boldsymbol{n}_2[k] \tag{7.2}$$

The matrix $\boldsymbol{\Phi}[k, k-1]$ is a known state transition matrix that relates the state at the previous time step $k-1$ to the state at the current step k. The matrix $\boldsymbol{S}[k]$ takes into the account an optional control input to the state. The matrix $\boldsymbol{C}[k]$ is a known measurement matrix. The process and measurement uncertainty are represented by $\boldsymbol{n}_1[k] = N(0, \boldsymbol{\Lambda}_1[k])$ and $\boldsymbol{n}_2[k] = N(0, \boldsymbol{\Lambda}_2[k])$ The space spanned by the observations $\boldsymbol{y}[1], \boldsymbol{y}[2], \cdots, \boldsymbol{y}[k-1]$ is denoted by $\boldsymbol{\mathcal{Y}}_{k-1}$.

Casting

Let us interpret the meaning of the input/output signal of the KF in relation with our perceptual problem. Concerning the measurement equation, still hold the same observations made above. It is instead worthwhile to specifically comment on the process equation and the filter's output.

Process equation - Assuming $\boldsymbol{x}$ a vector containing the values of a bunch of visual features over a fixed spatial region, Eq. 7.1 models the temporal evolution of the relationships among such features, according to specific rules embedded in the transition matrix $\boldsymbol{\Phi}$. By example, if we consider just one feature (e.g., motion velocity), $\boldsymbol{x}[k]$ will represent the "model" optic flow values at time step k, for all the (discrete) locations of the considered spatial regions (the velocity state). If $\boldsymbol{\Phi}$ has a diagonal structure, the process equation will describe the "model" temporal evolution of punctual velocities, independently of the spatial neighborhood values (temporal context). On the other hand, if $\boldsymbol{\Phi}$ shows a non-diagonal structure, the process equation models a "model" temporal evolution of the state that takes into account *also* spatial relationships (spatio-temporal context). More generally, if we build a state vector that collects more multiple features (e.g., motion, stereo, etc.), by proper specification of the transition matrix $\boldsymbol{\Phi}$, the process equation can potentially model any type of *multimodal* spatio-temporal relationships (multimodal context).

Filter output - Apart from the KF output $\hat{\boldsymbol{x}}$, we could be interested in making the measurements more confident. Accordingly, the output will be $\hat{\boldsymbol{y}}[k|\boldsymbol{\mathcal{Y}}_k]$, to be compared with $\boldsymbol{y}[k]$. The additional (contextual) information will be provided by Kalman innovation $\boldsymbol{\alpha}[k]$. We expect that, if the model is correct, the uncertainty associated to the *a posteriori* estimate of the actual measure $\hat{\boldsymbol{y}}[k|\boldsymbol{\mathcal{Y}}_k]$ is inferior to the uncertainty associated to the actual measure itself $\boldsymbol{y}[k]$. In this perspective, starting from the general scheme of the KF, and taking into account that it is possible to derive a recursive expression for $\boldsymbol{\alpha}[k]$, the resulting architectural scheme of the KF is shown in Fig. 7.2.

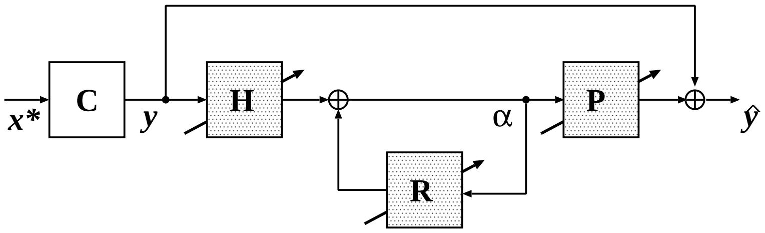


Figure 7.2: Schematic illustration of the KF. Grayed blocks represent adaptive linear operators, depending on the state, the measures, and on the statistics of noise.

Chapter 8

Progress results

The implementation of the Kalman-based context-sensitive filter will be articulated by considering (a) two different channels (motion and stereo) and (b) three increasing level of "context" complexity: (1) spatial, (2) spatio-temporal, and (3) multimodal.

In this Chapter we will report the results obtained for the **motion channel** with **spatial context**.

8.1 Problem formulation

Given motion information represented by an optic flow field, we want to recognize if a group of velocity vectors belong to a specific pattern, on the basis of their relationships in a spatial neighborhood. Casting the problem as a Kalman filter, the detection occurs through a spatial recurrent filter that checks the consistency between the spatial structural properties of the input flow field pattern and a structural rule expressed by the process equation of the KF.

8.2 Spatial "motion Gestalts"

8.2.1 First-order differential structures

Local spatial features around a given location of a flow field, can be of two types: (1) the average flow velocity at that location, and (2) the structure of the local variation in a the neighborhood of that locality [39]. The former relates to the *smoothness constraint* or *structural uniformity.* The latter relates to *linearity constraint* or *structural gradients* (linear deformations). Velocity gradients provide important cues about the 3-D layout of the visual scene. On a local scale, velocity gradients caused by the motion of objects provide perception of their 3-D structure (structure from motion and motion segmentation), whereas, on a global scale, they specify the observer's position in the world, and his/her heading.

Formally, linear (first-order) deformations can be described by a deformation tensor

$$\mathbf{T} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} \partial v_x/\partial x & \partial v_x/\partial y \\ \partial v_y/\partial x & \partial v_y/\partial y \end{bmatrix} . \tag{8.1}$$

Thence, if $\boldsymbol{x} = (x, y)$ is a point in a spatial image domain, the linear properties of a motion field $\boldsymbol{v}(x, y) = (v_x, v_y)$ around the point $\boldsymbol{x}_0 = (x_0, y_0)$ can be characterized by a Taylor expansion, truncated at the first order:

$$\boldsymbol{v} = \bar{\boldsymbol{v}} + \bar{\mathbf{T}}\boldsymbol{x} \tag{8.2}$$

or

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = \begin{bmatrix} \bar{v}_x \\ \bar{v}_y \end{bmatrix} + \bar{\mathbf{T}} \begin{bmatrix} x \\ y \end{bmatrix} \tag{8.3}$$

where $\bar{\boldsymbol{v}} = (\bar{v}_x, \bar{v}_y) = \boldsymbol{v}(x_0, y_0)$ and $\bar{\mathbf{T}}$ is the tensor operator evaluated in the point $\boldsymbol{x}_0$. Considering that Eq. 8.3 can be rewritten as

$$\begin{aligned} \begin{bmatrix} v_x \\ v_y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \bar{v}_x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \bar{v}_y \; &+ \; \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \left.\frac{\partial v_x}{\partial x}\right|_{\substack{x=x_0 \\ y=y_0}} + \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \left.\frac{\partial v_x}{\partial y}\right|_{\substack{x=x_0 \\ y=y_0}} \\ &+ \; \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \left.\frac{\partial v_y}{\partial x}\right|_{\substack{x=x_0 \\ y=y_0}} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \left.\frac{\partial v_y}{\partial y}\right|_{\substack{x=x_0 \\ y=y_0}} \end{aligned}$$

flow field can be locally described through two-dimensional maps ($\boldsymbol{f} : \Re^2 \mapsto \Re^2$) representing *cardinal* flow components:

$$\boldsymbol{v} = \boldsymbol{\alpha}^x \bar{v}_x + \boldsymbol{\alpha}^y \bar{v}_y + \boldsymbol{d}^x_x \left.\frac{\partial v_x}{\partial x}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}^x_y \left.\frac{\partial v_x}{\partial y}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}^y_x \left.\frac{\partial v_y}{\partial x}\right|_{\boldsymbol{x}_0} + \boldsymbol{d}^y_y \left.\frac{\partial v_y}{\partial y}\right|_{\boldsymbol{x}_0} \tag{8.4}$$

where $\boldsymbol{\alpha}^i$ are pure translations:

$$\boldsymbol{\alpha}^x : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \boldsymbol{\alpha}^y : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and $\boldsymbol{d}^i_j$ represent cardinal deformations, basis of the linear deformation space[1]:

$$\boldsymbol{d}^x_x : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} x \\ 0 \end{bmatrix} \quad \boldsymbol{d}^x_y : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} y \\ 0 \end{bmatrix} \quad \boldsymbol{d}^y_x : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} 0 \\ x \end{bmatrix} \quad \boldsymbol{d}^y_y : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} 0 \\ y \end{bmatrix} .$$

The set of partial derivatives evaluated in $\boldsymbol{x}_0$ are scalar quantities that relate to the "gradient rate". It is worthy to note that the components of pure translations could be incorporated in the corresponding deformation components, thus obtaining generalized deformation components with shifted motion boundaries, and, to the limit, to deformation components without motion boundaries. Although this does not affect the significance of the Taylor expansion (see Eq. 8.4), the so modified elementary components, present very different structural properties. Since we assume a template-based approach (in the sense that single neurons cannot extract single components but only perform pattern matching operations), the linear decomposition of the optic flow has significance only for the specification of a proper representation space. Specific templates would be designed to optimally sample that representation space. In this work, without loss of generality, we consider two different classes of deformation templates (opponent and non-opponent), each characterized by two gradient types (stretching and shearing), see Fig. 8.1. Shearing patterns are characterized by a right angle between the direction of motion and the direction of the steepest ascent of the gradient, whereas in stretching patterns both those vectors align. Due to their ability to detect the presence and the orientation of velocity gradients and kinetic boundaries, such cardinal EFCs and proper combinations of them resemble the characteristics of the cells in the Middle Temporal visual area (MT) [52] [71] [62]. It is straightforward to derive that these MT-like components are well suited to provide the building blocks for the more complex receptive field properties encountered in the Medial Superior

[1] This basis defines a four-dimensional linear subspace of the first order properties of the velocity field.

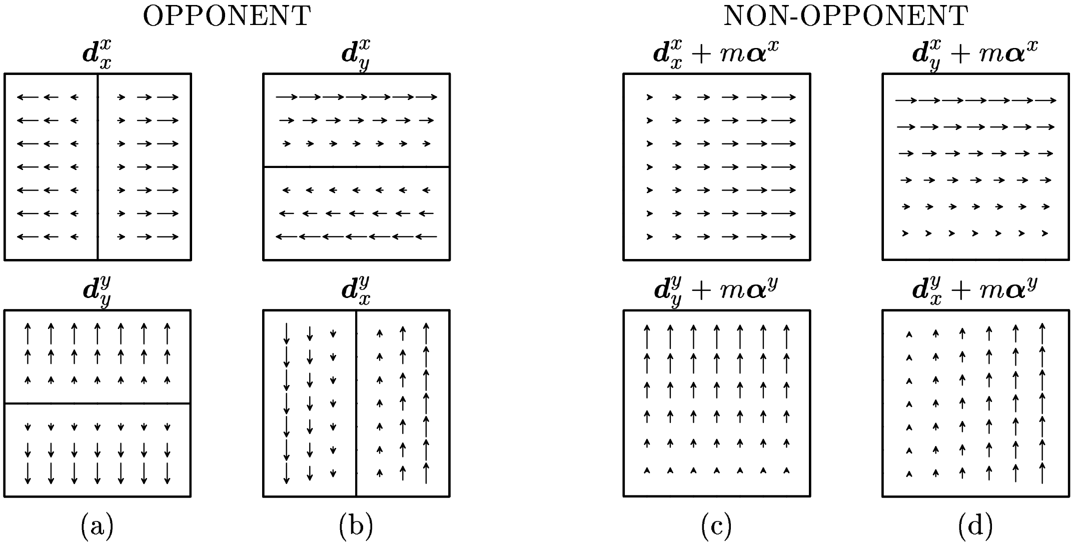


Figure 8.1: Basic gradient type Gestalts considered. In stretching-type components (a,c) velocity varies *along* the direction of motion; in shearing-type components (b,d) velocity gradient is oriented *perpendicularly* to the direction of motion. Non-opponent patterns are obtained from the opponent ones by a linear combination of pure tranlations and cardinal deformations: $\boldsymbol{d}_j^i + m\boldsymbol{\alpha}^i$, where m is a proper positive scalar constant.

Temporal visual area (MST) [11] [50] [56]. It is, indeed, well known that the tensor $\bar{\mathbf{T}}$ can be decomposed as:

$$\bar{\mathbf{T}} = \begin{bmatrix} \bar{T}_{11} & \bar{T}_{12} \\ \bar{T}_{21} & \bar{T}_{22} \end{bmatrix} = E \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \omega \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + S_1 \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + S_2 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

where

$$E = \frac{\bar{T}_{11} + \bar{T}_{22}}{2} \qquad \omega = \frac{\bar{T}_{12} - \bar{T}_{21}}{2} \qquad S_1 = \frac{\bar{T}_{11} - \bar{T}_{22}}{2} \qquad S_2 = \frac{\bar{T}_{12} + \bar{T}_{21}}{2}$$

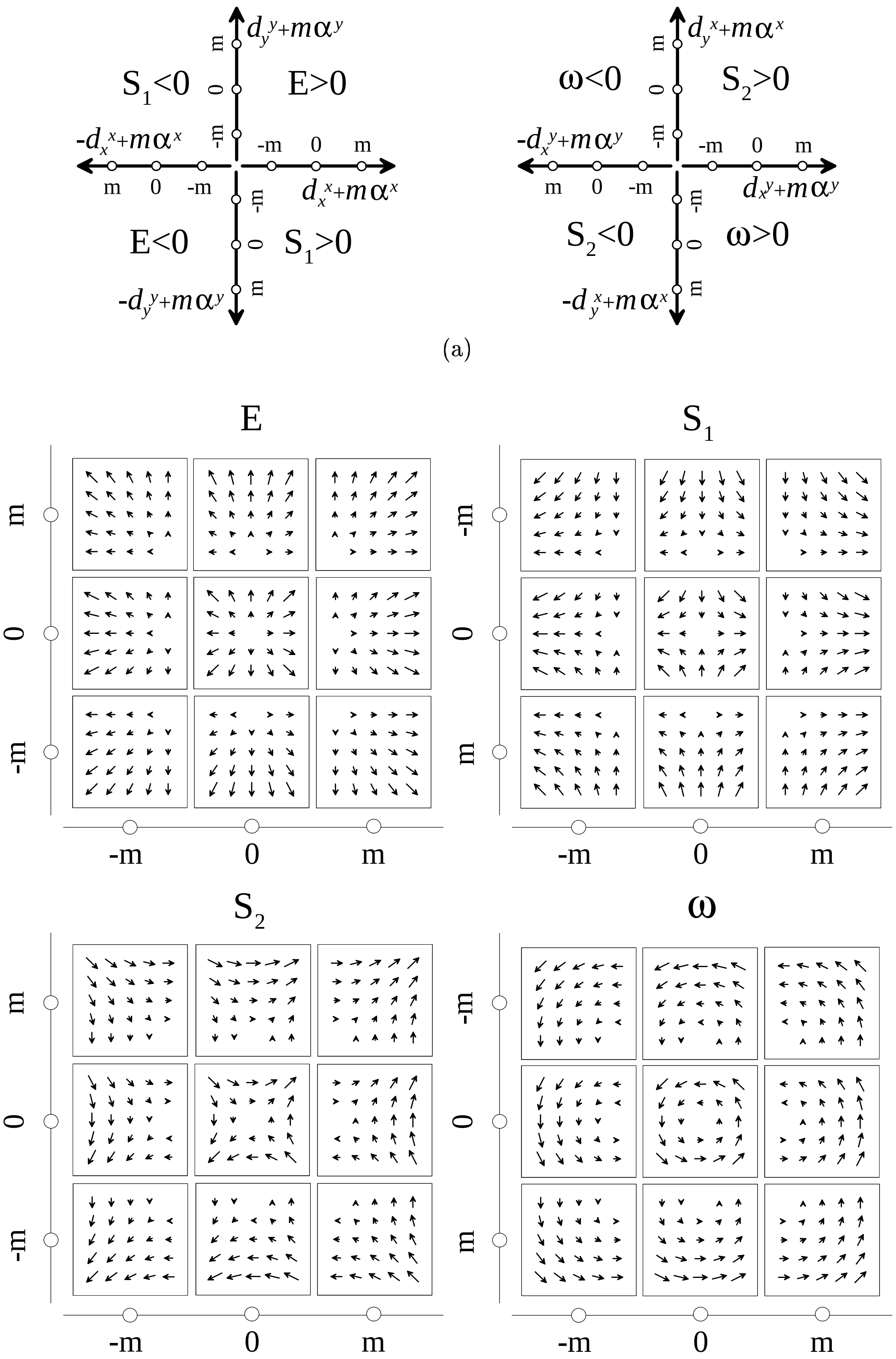
are the divergence, the curl and the two components of shear deformation of the vector field, respectively. Accordingly, the optic flow can be decomposed in an isotropic expansion, a rigid rotation and two components of shear (cf. [39]):

$$\boldsymbol{v} = \boldsymbol{\alpha}^x \bar{v}_x + \boldsymbol{\alpha}^y \bar{v}_y + \frac{1}{2}(\boldsymbol{d}_x^x + \boldsymbol{d}_y^y)E + \frac{1}{2}(\boldsymbol{d}_x^x - \boldsymbol{d}_x^y)\omega + \frac{1}{2}(\boldsymbol{d}_x^x - \boldsymbol{d}_y^y)S_1 + \frac{1}{2}(\boldsymbol{d}_y^x + \boldsymbol{d}_x^y)S_2$$

or

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} v_x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} v_y + \frac{1}{2} \begin{bmatrix} x \\ y \end{bmatrix} E + \frac{1}{2} \begin{bmatrix} -y \\ x \end{bmatrix} \omega + \frac{1}{2} \begin{bmatrix} x \\ -y \end{bmatrix} S_1 + \frac{1}{2} \begin{bmatrix} y \\ x \end{bmatrix} S_2$$

These mixed EFCs constitute, together with the pure translations, an equivalent representation basis for the linear properties of the velocity field. Also in this case, the considerations on the template-matching character of computation and on the necessity of building a proper set of templates that optimally span the representation space still hold. Accordingly, templates with shifted center of motion would be introduced (cf. [11]), see Fig. 8.2. Anyway, mixed EFCs are *derived* quantities that define a higher level representation space. Such EFCs are indeed



(a)

(b)

Figure 8.2: (a) Two deformation subspaces obtained by the set of cardinal EFCs with different values of the parameter m. The quadrants of each subspace characterize an elementary deformation, as evidenced in (b) for expansion ($E > 0$), horizontal positive shear ($S_1 > 0$), oblique positive shear (S_2), and counterclockwise rotation ($\omega > 0$).

rather complex since not only the speed, but also the direction of feature motion varies as a function of spatial position. Rigid body motion often generates simpler flow fields characterized by unidirectional patterns, as the cardinal EFCs considered in this study.

8.2.2 The Common-cause paradigm

To achieve effective motion segmentation in real sequences, distinguishing object and ego motions we can rely on the different causes that originate the optic flow patterns. We can use the common cause idea to define motion Gestalts from motion templates at two levels. At local level, the coherent motion of objects in the field leads to local coherent motion patterns, either if the motion of the object is due to its own motion or due to motion of the observer. These local coherent motion patterns must not be just uniform motion, but depending on the 6D motion properties (of either the object or the observer) can consist of various smooth motion patterns. However, they are always resulting from the same cause and will thus form a single, predictable Gestalt. At the local level, the Gestalt is therefore defined by the similarities in the the local motions of a single coherent pattern. At global level, ego-motion induces visual motion of the entire visual field in a common-cause way. As this common cause is directly coming from the ego-motion parameters of the observer it appears sensible to define high level motion Gestalts by virtue of the ego-motion parameters. Thus, a template for the estimation of a particular ego motion (or a particular observer heading) comprise a motion Gestalt at the global level. At the global level, the Gestalt is therefore defined by the motion parameters of the observer or object motion. This global/local approach has a direct link to problems in real driving scenes, in which multiple motions can occur at the same time (own vehicle, other vehicles).

The relationships between local and global properties in the optic flow can be modeled by considering the statistical interdependences among local patterns of the optic flow (e.g. a study of the joint probability density function for EFCs in ecological flows).

8.3 The context-sensitive filter

The problem of evidencing the presence of a certain complex feature in the optic flow on the basis of both local and contextual information, is posed as an adaptive filtering problem (estimation), where local information act as the input *measurements* and context acts as the *reference signal*, e.g., representing a specific motion Gestalt. In the following, we propose a solution in the form of a generalized Kalman filter theory [30] [23].

8.3.1 Measurement equation

For the sake of simplicity, we assume to have direct noisy measurements $\tilde{\boldsymbol{v}}(i,j)$ of the actual velocity field $\boldsymbol{v}(i,j)$. The difference between these two variables can be represented as a constant noise term $\boldsymbol{\varepsilon}(i,j)$:

$$\tilde{\boldsymbol{v}} = \boldsymbol{v} + \boldsymbol{\varepsilon} \; . \tag{8.5}$$

Due to the intrinsic noise of the nervous system, the neural representation of the optic flow $\mathbf{v}(i,j)[k]$ can be expressed by a *measurement equation*:

$$\mathbf{v}[k] = \tilde{\boldsymbol{v}} + \boldsymbol{n}_1[k] = \boldsymbol{v} + \boldsymbol{\varepsilon} + \boldsymbol{n}_1[k] \tag{8.6}$$

where $\boldsymbol{n}_1$ represents the uncertainty associated with a neuron's response. In this case the measurement matrix $\boldsymbol{C} = \boldsymbol{I}$. The approach can be straightforwardly generalized to consider indirect motion information, e.g., by the gradient equation:

$$-I_t[k] = \boldsymbol{\nabla}^T I[k]\tilde{\boldsymbol{v}}[k] + \boldsymbol{n}_1[k]$$

where $\boldsymbol{\nabla}^T I$ and I_t are the spatial image gradient and temporal derivative, respectively, of the image at a given spatial location and time. It is worthy to note that here the linear operator relating the quantity to be estimated to the measurement I_t is *also* a measurement [68].

8.3.2 Process equation

In the present case, the reference signal should reflect structural regularities in space of the input optic flow. These structural regularities can be described statistically and/or geometrically. In any case, they can be defined by a process equation that models spatial relationships by the transition matrix $\boldsymbol{\Phi}$:

$$\boldsymbol{v}[k] = \boldsymbol{\Phi}\boldsymbol{v}[k-1] + \boldsymbol{n}_2[k-1] + \boldsymbol{s} \tag{8.7}$$

with

$$\lim_{k\to\infty} \boldsymbol{v}[k] = \boldsymbol{v} \qquad \text{if} \quad \boldsymbol{n}_2 = 0$$

where the state transition matrix is *de facto* a spatial interconnection matrix that implements a specific Gestalt rule (i.e., a specific EFC); $\boldsymbol{s}$ is a constant driving input; $\boldsymbol{n}_2$ represents the process uncertainty. The space spanned by the observations (of the neural response) $\mathbf{v}[1]$, $\mathbf{v}[2]$,. . . , $\mathbf{v}[k-1]$ is denoted by $\boldsymbol{\mathcal{V}}_{k-1}$ and represents the internal noisy representation of the optic flow. We assume that both $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are independent, zero-mean and normally distributed $\boldsymbol{n}_1[k] = N(0, \boldsymbol{\Lambda}_1)$ and $\boldsymbol{n}_2[k] = N(0, \boldsymbol{\Lambda}_2)$. Considering that we are tackling spatial context only, the meaning of the temporal step k requires a clarification: in this circumstance k does not model physical time, but only the *iteration time* necessary for the convergence of the KF. Indeed, it takes some time to the context information to propagate, and produce an effect on the filter output. The time step k will have full physical meaning when we will generalize the approach to model spatio-*temporal* context. The matrix $\boldsymbol{\Phi}$ models space-invariant nearest-neighbor interactions within a finite region Ω in the (i, j) plane that is bounded by a piece-wise smooth contour. Interactions occur, separately for each component of the velocity vectors (v_x, v_y), through anisotropic interconnection schemes:

$$\begin{aligned} v_{x/y}(i,j)[k] &= w_N^{x/y} v_{x/y}(i,j-1)[k-1] + w_S^{x/y} v_{x/y}(i,j+1)[k-1] + s_{x/y}(i,j) \\ &+ \ w_W^{x/y} v_{x/y}(i-1,j)[k-1] + w_E^{x/y} v_{x/y}(i+1,j)[k-1] + n_1^{x/y}(i,j)[k-1] \end{aligned}$$

where (s_x, s_y) is a steady additional control input, which models the boundary conditions. The process equation has a *structuring effect* constrained by the boundary conditions that yields to structural equilibrium configurations, characterized by specific first-order EFCs. The resulting pattern depends on the anisotropy of the interaction scheme and on the boundary conditions. By example, considering, for the sake of simplicity, a rectangular domain $\Omega = [-L, L] \times [-L, L]$, the cardinal EFC $\boldsymbol{d}_x^x$ can be obtained through:

$$\begin{array}{llll} w_N^x = w_S^x = 0 & w_N^y = w_S^y = 0 & s_x(i,j) = \left\{ \begin{array}{cl} -\lambda & \text{if } i = -L \\ \lambda & \text{if } i = L \\ 0 & \text{otherwise} \end{array} \right. & s_y(i,j) = 0 \\ w_W^x = w_E^x = 0.5 & w_W^y = w_E^y = 0 & & \end{array}$$

where the boundary value λ controls the gradient slope. In a similar waywe can obtain the other components.

Given Eqs. (8.6) and (8.7), we may write the optimal filter for optic flow Gestalts. The filter allows to detect, in noisy flows, intrinsic correlations, as those related to EFCs, by checking, through spatial recurrent interactions, that the spatial context of the observed velocities conforms to the Gestalt rules, embedded in $\boldsymbol{\Phi}$. To understand how the CSF works, we define the *a priori* state estimate at step k given knowledge of the process at step $k-1$, $\hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}]$, and the *a posteriori* state estimate at step k given the measurement at the step k, $\hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_k]$. The aim of the CSF is to compute an *a posteriori* estimate by using an *a priori* estimate and a weighted difference between the current and the predicted measurement:

$$\hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_k] = \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}] + \boldsymbol{G}[k]\ \ (\mathbf{v}[k] - \hat{\mathbf{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}]) \tag{8.8}$$

The difference term in Eq. (8.8) is the *innovation* $\boldsymbol{\alpha}[k]$ that takes into account the discrepancy between the current measurement $\mathbf{v}[k]$ and the predicted measurement $\hat{\mathbf{v}}[k|\boldsymbol{\mathcal{V}}_{k-1}]$. The matrix $\boldsymbol{G}[k]$ is the Kalman gain that minimizes the *a posteriori* error covariance:

$$\boldsymbol{K}[k] = E\left\{(\boldsymbol{v}[k] - \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_k])(\boldsymbol{v}[k] - \hat{\boldsymbol{v}}[k|\boldsymbol{\mathcal{V}}_k])^T\right\}\ . \tag{8.9}$$

Eqs. 8.8 and 8.9 represent the mean and covariance expressions of the CSF output. The covariance matrix $\boldsymbol{K}[k]$ provides us only information about the properties of convergence of the KF and not whether it converges to the correct values. Hence, we have to check the consistency between the innovation and the model (i.e., between observed and predicted values) in statistical terms. A measure of the reliability of the KF output is the Normalized Innovation Squared (NIS):

$$NIS_k = \boldsymbol{\alpha}^T[k]\ \boldsymbol{\Sigma}^{-1}[k]\ \boldsymbol{\alpha}[k] \tag{8.10}$$

where $\boldsymbol{\Sigma}$ is the covariance of the innovation. It is possible to exploit Eq. (8.10) to detect if the current observations are an instance of the model embedded in the KF [3].

8.4 Results

Fig. 8.3 shows the responses of the CSF in the deformation subspaces for two different input flows. Twentyfour EFC models have been used to span the deformation subspaces shown in Fig. 8.2a. The grey level in the CSF output maps represents the probability of a given Gestalt according to the NIS criterium: lightest grey indicates the most problable Gestalt. Besides Gestalt detection, context information reduces the uncertainty on the measured velocities, as evidenced, for the circled vectors, by the Gaussian densities, plotted over the space of image velocity.

To assess the performances of the KF we applied the CSFs, also to real world optic flows. In particular, we considered the driving sequences provided by Hella, which show road scenes taken by a rear-view mirror of a moving car and present different situations of an overtaking car. A “classical” algorithm [51] has been used to extract the optic flow (see Fig. 8.4a). To achieve motion segmentation in these real sequences we analyzes overlapping local regions of the optic flow with the twenty-four EFC models. In this way, we can compute a dense distribution of the local Gestalt probabilities for the overall optic flow. Thence, we obtain, according to the NIS criterium, the most reliable local velocity patterns, that is the patterns of local Gestalts that characterize the sequence (see Fig. 8.4b).

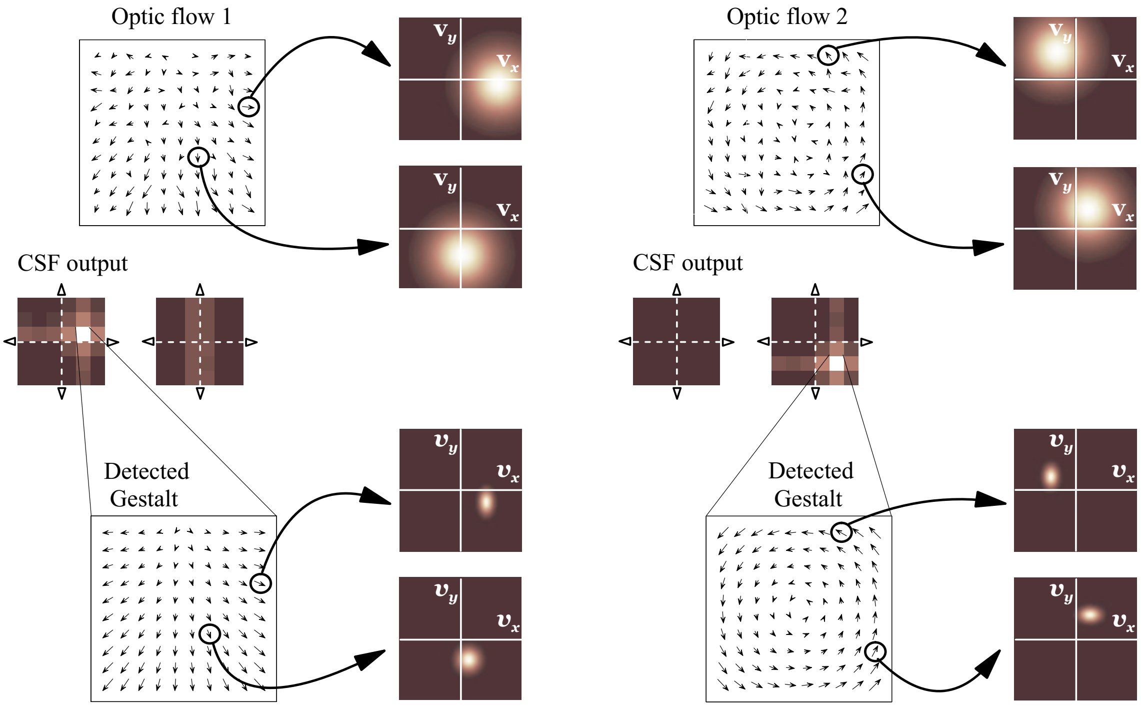


Figure 8.3: Example of Gestalt detection in noisy flows.

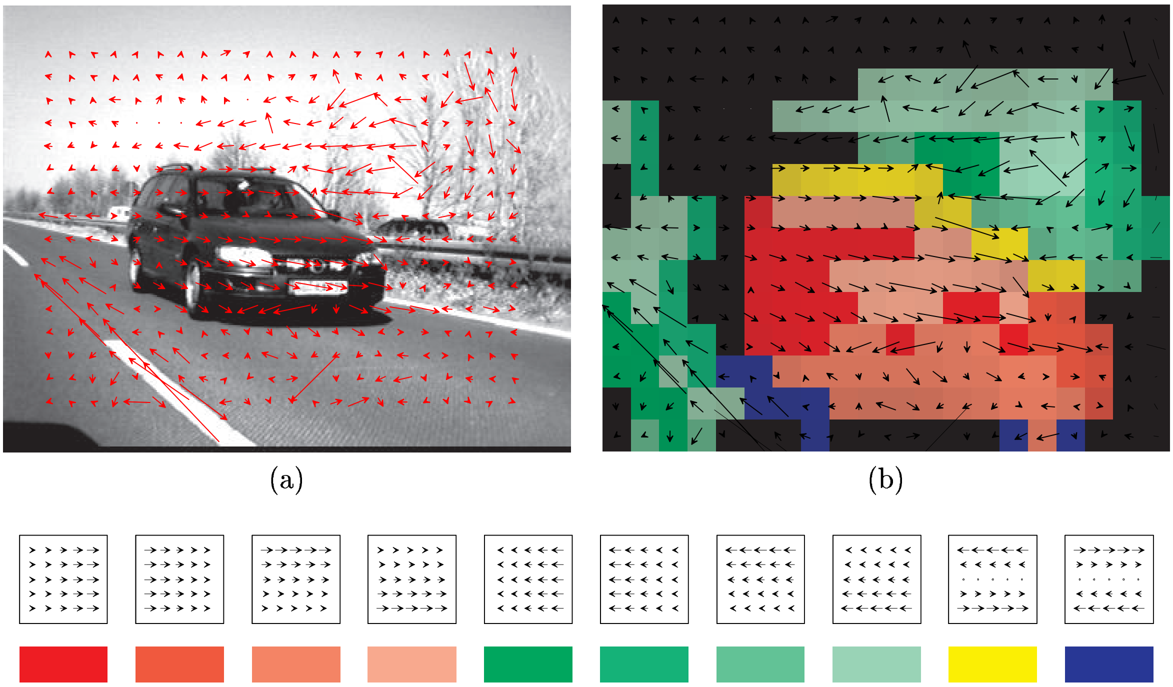
8.5 Discussion and conclusion

Measured optic flow fields are always somewhat erroneous and/or ambiguous. First, we cannot compute the actual spatial or temporal derivatives, but only their estimates, which are corrupted by image noise. Second, optic flow is intrinsically an image-based measurement of the relative motion between the observer and the environment, but we are interested in estimating the actual motion field. However, real-world motion field patterns contain intrinsic properties that allow to define Gestalts as groups of pixels sharing the same motion property. By checking the presence of such Gestalts in optic flow fields we can make their interpretation more confident. We propose an optimal recurrent filter capable of evidencing motion Gestalts corresponding to 1st-order spatial derivatives or elementary flow components (EFCs). A Gestalt emerges from a noisy flow as a solution of an iterative process of spatially interacting nodes that correlates the properties of the visual context with that of a structural model of the Gestalt.

The CSF behaves as a template model. Yet, its specificity lies in the fact that the template character is not built by highly specific feed-forward connections, but emerges by stereotyped recurrent interactions (cf. the process equation). Furthermore, the approach can be straightforwardly extended to consider adaptive cross-modal templates (e.g, motion and stereo). By proper specification of the matrix $\mathbf{\Phi}$, the process equation can, indeed, potentially model any type of multimodal spatio-temporal relationships (i.e., multimodal spatio-temporal context).

8.5.1 Related scenarios

In general, KF represents a recursive solution to an inverse problem of determing the distal stimulus based on the proximal stimulus, in case



(a)

(b)

(c)

Figure 8.4: a) The computed optic flow superimposed on a frame of the sequence provided by Hella. b) The resulting motion segmentation superimposed to the optic flow evidenced in a). c) The color code of the most probable EFCs ecountered. We use the HSV color space: the *hue* identifies the direction of motion, the *saturation* discriminates the kind of EFC for a given hue and the *value* represents the probability of the given EFC. The two opponent EFCs are identified only by the hue, while the saturation and the value are fixed. The color black indicates that, for the considered region, the reliability of the segmentation is below a given threshold. It is worthy to note that the CSFs detected the motion edges (yellow and blue colors) between the areas of coherent motions (green and red regions for leftward and rightward motions, respectively).

1. we adopts a stochastic version of the regularization theory involving Bayes' rule
2. we assume Markovianity
3. we consider linear Gaussian models (linearity and Gaussian normal densities).

- The first condition can be motivated by the fact that the a priori contraints necessary to regularize the solution can be described in probabilistic terms. Bayes' rule allows the computation of the *a posteriory* probability $p(\boldsymbol{x}|\boldsymbol{y})$ as follows:

  $$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}$$

  where $p(\boldsymbol{x})$ is the *a priori* probability densities for the distal stimulus and represents *a priori* knowledge about the visual scene; $p(\boldsymbol{y}|\boldsymbol{x})$ is the likelihood function for $\boldsymbol{x}$. This function represents the transformation from the distal to proximal stimulus and includes information about noise in the proximal stimulus. Finally, $p(\boldsymbol{y})$ is the probability of obtaining the proximal stimulus. The inverse problem of determining the distal stimulus can be solved by finding $\hat{\boldsymbol{x}}$ that maximizes the *a posteriori* probability, $p(\boldsymbol{x}|\boldsymbol{y})$. Such $\hat{\boldsymbol{x}}$ is called a maximum a posteriori (MAP) estimator. Although the Bayesian framework is more general than the standard regularization, there exist a relationship between the deterministic and stochastic methods of solving inverse problems. Under some assumptions about the probability densities, maximizing the *a posteriory* probability $p(\boldsymbol{x}|\boldsymbol{y})$ is, indeed, equivalent to minimizing the Tikhonov functional[2].

- The second concept, the Markovianity, captures the step-by-step local nature of the interactions in a cooperative system, and makes possible Kalman recursion, by allowing to express *global* properties of the state in terms of its *local* properties. Under these hypotheses the conditional probability that the system is in a particular state at any time is determined by the distribution of states at its immediately preceding time. That is, the conditional distribution of that states of a system given the present and past distributions depends only upon the present. Specifically, considering the visual signal as a Random Field, the Markovianity hypothesis implies that the joint probability distribution of that random field has associated positive-definite, translational invariant conditional probabilities that are spatially Markovian (Markov Random Fields)[3].

- The third assumption represents the necessary conditions to achieve the exact, analytical solution of the KF.

[2] The regularization method of solving ill-posed inverse problems was formulated by Tikhonov in the early 60s [70]. In this method, the solution is obtained by finding $\hat{\boldsymbol{x}}$, which minimizes a functional:

$$E = ||A\boldsymbol{x} - \boldsymbol{y}||^2 + \lambda||P\boldsymbol{x}||^2$$

where λ is a regularization parameter. The first norm evaluates how close the distal stimulus is to the proximal stimulus, and the second norm evaluates how well the *a priori* constraints are satisfied. If the proximal stimulus is reliable, λ should be small, otherwise λ should be large. In Tikhonov's theory, A is assumed to be a linear operator, $P\boldsymbol{x}$ a linear combination of the first p derivatives of the distal stimulus $\boldsymbol{x}$, and the norm are quadratic.

[3] A system is temporally Markovian if its state at a particular time depends on its state at the immediate preceding time but not on any of its states at earlier times. Similarly, a system is spatially Markovian if the states of its constituent elements depend on those of their neighbors, but not on the states of units that are spatially remote. These local temporal and spatial properties can be described mathematically using the probalistic language of Markov chains and processes, and Markov Random Fields (MRFs). Formally, by considering a random field $F = \{F_1, \ldots, F_m\}$ as a family of random variables defined on the set $\mathcal{S}$, in which each random variable F_i takes a value f_i in $\mathcal{L}$, F is said to be a MRF on $\mathcal{S}$ with respect to a neighborhood system $\mathcal{N}$ iff the

Relationships with MRFs. MRF theory is a branch of probality theory for analyzing the spatial or contextual dependencies of physical phenomena. It is used in visual/image processing to model context dependent entities such as image pixels and correlated features probalistic distributions of interacting labels, i.e., to describe in statistical terms the structure and correlations present in natural images. Formally, MRF theory tells us how to model the *a priori* probability of contextual dependent patterns. Contextual constraints may be expressed locally in terms of conditional probabilities $p(f_i|\{f'_i\})$, where $\{f'_i\}$ denotes the set of values at the other sites $i' \neq i$, or globally as the joint probability $p(f)$. Because local information is more directly observed, it is normal that a global inference is made on local properties. How to make a global inference using local information becomes a non-trivial task. MRF theory provides a mathematical foundation for solving this problem, by relating global and local properties of a cooperative system. Information on the nearest neighborhood is used to calculate conditional probabilities. In their pioneering work, Geman and Geman [24] expressed the statistical properties of the natural images in terms of cooperative interactions among pixel elements. The images are endowed with an artificial equilibrium dynamics that evolves the lattice system through a series of configurations to a near-optimal low energy state. Depending on the task being addressed, the optimal states obtained by MRF image processing methods, are those for which noise, blur, and other artifacts have been removed (image reconstruction) and/or where pixels belonging to the same entity have been identified (segregation and segmentation).

Remark 1: MRF methods for image processing usually assume to have the direct accessibility to the "system", whereas in Kalman filter theory only system's measures are observable. More generally we can refer to dynamic (discrete time) *state space models* [29] [72] [49] (cf. also Hidden MRF) given by

$$\boldsymbol{x}[k] \sim p(\boldsymbol{x}[k] \mid \boldsymbol{x}[0], \ldots, \boldsymbol{x}[k-1]) \ , \qquad \text{system}$$

$$\boldsymbol{y}[k] \sim p(\boldsymbol{y}[k] \mid \boldsymbol{x}[k]) \ , \qquad \text{observations}$$

where $\boldsymbol{y}[k]$ contain the observations at time step k, while $\{\boldsymbol{x}[k]\}$ is an underlying stochastic process which in some cases may have a physical meaning while in other cases it is merely included in order to describe the distribution of the observation process properly. Typically, some prior distribution is placed on $\boldsymbol{x}[0]$. An important task when analyzing data by state space models is estimation of the underlying state process, based on measurements from the observation process. The interest might be on $\boldsymbol{x}[k]$ itself, or merely is a tool for making prediction on $\boldsymbol{y}[k]$. In this perspective, the process (state) equation can be a MRF. The presence of the measurement equation (observations) makes more evident the distinction between the feed-forward and feed-back components of the filter.

Remark 2: Although it is straightforward to derive, in the case of dynamic state-space models (MRF models in time series) for linear Gaussian models, the KF, as an efficient and exact algorithm for computing inference, *spatial* MRFs should be reformulated to be mathematically identical to dynamic models and make the KF work.

following two conditions are satisfied:

$$p(f) > 0, \forall f$$
$$p(f_i|f_{S-\{i\}}) = p(f_i|f_{\mathcal{N}_i})$$

where $f_{\mathcal{N}_i} = \{f_{i'}|i' \in \mathcal{N}_i\}$ at the site neighboring i.

Remark 3: The process equation, thought as a state space model describes the statistical properties of the system (visual signal). In this sense, it can be used to model statistical Gestalt rules (good continuation, common fate, etc.) with typical constraint priors, such as "smoothness", "continuity", etc. Yet, optic flow patterns generated by ego- or rigid-body motion, show specific features that cannot be described only in statistical terms, since the velocity vectors in different spatial locations are subject to topological and geometric constraints. It is worthy to note that the process equation adopted in our study, models a structural property of the state space. In that sense, it is possible to describe *specific* vector configurations over (large) spatial regions (i.e., "that radial pattern outflowing from P" *vs* "radiality"). Accordingly, the filter behaves as a template-matching model. To look for Gestalts on the basis of statistical properties a different approach should be followed, requiring the definition of a process equation on a statistical basis.

Acknowledgments We wish to thank Martin Mühlenberg of HELLA KG Hueck & Co. for having provided the overtaking car sequences.

Bibliography

[1] J. Aloimonos and D. Shulman. *Integration of Visual Modules — An extension of the Marr Paradigm.* Academic Press, London, 1989.

[2] N. Ayache. *Stereovision and Sensor Fusion.* MIT Press, 1990.

[3] Y. Bar-Shalom and X.R. Li. *Estimation and Tracking, Principles, Techniques, and Software.* Artech House, 1993.

[4] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1971.

[5] A.J. Bell and T. Sejnowski. Edges are the 'independent components' of natural scenes. *Advances in Neural Information Processing Systems*, 9:831–837, 1996.

[6] B.I. Bertenthal, J.J. Campos, and M.M. Haith. Development of visual organisation: The perception of subjective contours. *Child Development*, 51(4):1072–1080, 1980.

[7] Fischlerand Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 1981.

[8] R.C.K. Chung and R. Nevatia. Use of monucular groupings and occlusion analysis in a hierarchical stereo system. *Computer Vision and Image Understanding*, 62(3):245–268, 1995.

[9] A. Cozzi and F. Wörgötter. Comvis: A communication framework for computer vision. *International Journal of Computer Vision*, 41:183–194, 2001.

[10] A. Desolneux, L. Moisan, and J.M. Morel. Edge detection by the Helmholtz principle. *JMIV*, 14(3):271–284, 2001.

[11] C.J. Duffy and R.H. Wurtz. Response of monkey MST neurons to optic flow stimuli with shifted centers of motion. *J. Neuroscience*, 15:5192–5208, 1995.

[12] H. Elder and R.M. Goldberg. Inferential reliability of contour grouping cues in natural images. *Perception Supplement*, 27, 1998.

[13] O. Faugeras and L. Robert. What can two images tell us about the third one? *International Journal of Computer Vision*, 18(1), 1996.

[14] O.D. Faugeras. *Three–Dimensional Computer Vision.* MIT Press, 1993.

[15] M. Felsberg. PhD Thesis, 2002.

[16] M. Felsberg and G. Sommer. A new extension of linear signal processing for estimating local properties and detecting features. *Proceedings of the DAGM 2000*, pages 195–202, 2000.

[17] M. Felsberg and G. Sommer. The monogenic signal. *IEEE Transactions on Signal Processing*, 41(12), 2001.

[18] D. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America*, 4(12):2379–2394, 1987.

[19] W. Förstner. Image matching. In R.M. Haralick and L.G. Shapiro, editors, *Computer and Robot Vision.* Addison Wesley, 1993.

[20] K. Fukunaga, editor. *Introduction to statistical pattern recognition (2nd ed).* Academic Press, 1990.

[21] M.S. Gazzaniga. *The Cognitive Neuroscience.* MIT Press, 1995.

[22] W.S. Geisler, J.S. Perry, B.J. Super, and D.P. Gallogly. Edge co–occurrence in natural images predicts contour grouping performance. *Vision Research*, 41:711–724, 2001.

[23] A. Gelb. *Applied Optimal Estimation.* Artech House, MIT Press, 1974.

[24] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741, 1984.

[25] O. Granert. Posenschaetzung kinematischer ketten. *Diploma Thesis, Universität Kiel*, 2002.

[26] G. H. Granlund and H. Knutsson. *Signal Processing for Computer Vision.* Kluwer Academic Publishers, Dordrecht, 1995.

[27] G. Guy and G. Medioni. Inferring global perceptual contours from local features. *International Journal of Computer Vision*, 20:113–133, 1996.

[28] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2000.

[29] A.C. Harvey. *Forecasting, structural time series models and the Kalman filter.* Cambridge University Press, Cambridge, 1989.

[30] S. Haykin. *Adaptive Filter Theory.* Prentice-Hall International Editions, 1991.

[31] F. Heitger, R. von der Heydt, E. Peterhans, L. Rosenthaler, and O. Kübler. Simulation of neural contour meachnisms: representing anomalous contours. *Image and Vision Computing*, 16:407–421, 1998.

[32] D.D. Hoffman, editor. *Visual Intelligence: How we create what we see.* W.W. Norton and Company, 1980.

[33] K. Ikeuchi and B.K.P. Horn. Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, 17:141–184, 1981.

[34] Thomas Jäger. Interaktion verschiedener visueller Modalitäten zur stabilen Extraktion von objektrepräsentationen. *Diploma thesis (University of Kiel)*, 2002.

[35] B. Jähne. *Digital Image Processing – Concepts, Algorithms, and Scientific Applications.* Springer, 1997.

[36] J.R. Jordan and A.C. Bovik. Using chromatic information in edge based stereo correspondence. *Computer Vision, Graphics and Image Processing: Image Understanding*, 54:98–118, 1991.

[37] R. Klette, K. Schlüns, and A. Koschan. *Computer Vision - Three-Dimensional Data from Images.* Springer, 1998.

[38] R. Koch. Model-based 3-D scene analysis from stereoscopic image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 49(5):23–30, 1994.

[39] J.J. Koenderink. Optic flow. *Vision Res.*, 26(1):161–179, 1986.

[40] A. Koschan. How to utilize color information in dense stereo matching and in edge based stereo matching? *Proceedings of ICARCV*, pages 419–423, 1994.

[41] P. Kovesi. Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3):1–26, 1999.

[42] N. Krüger. Collinearity and parallelism are statistically significant second order relations of complex cell responses. *Neural Processing Letters*, 8(2):117–129, 1998.

[43] N. Krüger, M. Ackermann, and G. Sommer. Accumulation of object representations utilizing interaction of robot action and perception. *Knowledge Based Systems*, 15:111–118, 2002.

[44] N. Krüger, M. Felsberg, C. Gebken, and M. Pörksen. An explicit and compact coding of geometric and structural information applied to stereo processing. *Proceedings of the workshop 'Vision, Modeling and VISUALIZATION 2002'*, 2002.

[45] N. Krüger, T. Jäger, and Ch. Perwass. Extraction of object representations from stereo imagesequences utilizing statistical and deterministic regularities in visual data. *DAGM Workshop on Cognitive Vision*, 2002.

[46] N. Krüger and G. Peters. Orassyll: Object recognition with autonomously learned and sparse symbolic representations based on metrically organized local line detectors. *Computer Vision and Image Understanding*, 77:49–77, 2000.

[47] N. Krüger and F. Wörgötter. Different degree of genetical prestructuring in the ontogenesis of visual abilities based on deterministic and statistical regularities. *Proceedings of the Ẃorkshop On Growing up Artifacts that LiveŚAB 2002*, 2002.

[48] N. Krüger and F. Wörgötter. Multi modal estimation of collinearity and parallelism in natural image sequences. *to appear in Network: Computation in Neural Systems*, 2002.

[49] H.R. Künsch. State space and hidden Markov models. In *Complex Stochastic Systems, no. 87 in Monographs on Statistics and Applied Probability*, pages 109–173. Chapman and Hall, London, 2001.

[50] M. Lappe, F. Bremmer, M. Pekel, A. Thiele, and K.P. Hoffmann. Optic flow processing in monkey STS: A theoretical and experimental approach. *J. Neuroscience*, 16:6265–6285, 1996.

[51] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *Proc. DARPA Image Understanding Workshop*, pages 121–130, 1981.

[52] V.L. Marcar, D.K. Xiao, S.E. Raiguel, H. Maes, and G.A. Orban. Processing of kinetically defined boundaries in the cortical motion area MT of the macaque monkey. *J. Neurophysiol.*, 74(3):1258–1270, 1995.

[53] G. Medioni and R. Nevatia. Segment-based stereo matching. *Computer Vision, Graphics and Image Processing*, 31, 1985.

[54] H.H. Nagel. On the estimation of optical flow: Relation between different approaches and some new results. 33:299–324, 1987.

[55] B.A. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[56] G.A. Orban. The analysis of motion signal and the nature of processing in the primate system. In *Artificial and Biological Vision System*, pages 24–56. ESPRIT Basic Research Series, 1992.

[57] G. Palm. On associative memory. *Biological Cybernetics*, 36:19–31, 1980.

[58] W.A. Phillips and W. Singer. In search of common foundations for cortical processing. *Behavioral and Brain Sciences*, 20(4):657–682, 1997.

[59] T.Q. Phong, R. Horaud, A. Yassine, and P.T. Tao. Object pose from 2-d to 3-d point and line correspondences. *International Journal of Computer Vision*, 15:225–243, 1995.

[60] M. Pollefeys, R. Koch, and L. van Gool. Automated reconstruction of 3d scenes from sequences of images. *Isprs Journal Of Photogrammetry And Remote Sensing*, 55(4):251–267, 2000.

[61] S. Posch. *Perzeptives Gruppieren und Bildanalyse.* Habilitationsschrift, Universität Bielefeld, Deutscher Universitäts Verlag, 1997.

[62] S. Raiguel, M.M. Van Hulle, D.K. Xiao, V.L. Marcar, and G.A. Orban. Shape and spatial distribution of receptive fields and antagonist motion surrounds in the middle temporal area (v5) of the macaque. *Eur. J. of Neurosci.*, 7:2064–2082, 1995.

[63] B. Rosenhahn, N. Krüger, T. Rabsch, and G. Sommer. Automatic tracking with a novel pose estimation algorithm. *Robot Vision 2001*, 2001.

[64] S. Sarkar and K.L. Boyer. *Computing Perceptual Organization in Computer Vision.* World Scientific, 1994.

[65] C. Schmid and A. Zisserman. Automatic line matching across views. *Proc. IEEE Conference on Computer Vision and Pattern Recognition,*, pages 666–671, 1997.

[66] Y. Shiray. *Three-dimensional Computer Vision.* Springer (Berlin), 1987.

[67] M. Sigman, G.A. Cecchi, C.D. Gilbert, and M.O. Magnasco. On a common circle: Natural scenes aand gestalt rules. *PNAS*, 98(4):1935–1949, 2001.

[68] E.P. Simoncelli. Bayesian multi-scale differential optical flow. In *Handbook of Computer Vision and Applications*, pages 297–322. Academic Press, 1999.

[69] E.P. Simoncelli and B.A. Ohlshausen. Natural image statistics and neural representations. *Anual Reviews of Neuroscience*, 24:1193–1216, 2001.

[70] A.N. Tikhnov and V.Y. Arsenin. *Solutions of ill-posed problems.* Winston, Washington, DC, 1977.

[71] S. Treue and R.A. Andersen. Neural responses to velocity gradients in macaque cortical area MT. *Visual Neuroscience*, 13:797–804, 1996.

[72] M. West and J. Harrison. *Bayesian forecasting and dynamic models.* Springer-Verlag, New York, 1997.

[73] C. Zetzsche and E. Barth. Fundamental limits of linear filters in the visual processing of two dimensional signals. *Vision Research*, 30, 1990.

[74] C. Zetzsche and G. Krieger. Nonlinear mechanisms and higher–order statistics in biologial vision and electronic image processing: review and perspectives. *Journal of Electronic Imaging*, 10(1):56–99, 2001.