Technical Report about Time-Space Gestalts (Ecovision: Deliverable 3.2)

Norbert Krüger, Markus Lappe, Nicolas Pugeault and Florentin Wörgötter

November 29, 2003

Contents

Chapter 1

Introduction

In this report we describe the work performed within ECOVISION that addresses WP3. In the tenure of the project it has become clear that the subtasks formulated in WP3.2, WP3.3 and WP3.4 are closely intertwined. Therefore, although the deliverables addressing WP3.3 and WP3.4 are due in year three of the projects we describe aspects of WP3.3 concerning static inference and dynamic recursion that have been already addressed in this deliverable. We also describe preliminary work on grouping which focusses on aspects of WP3.4 that are used as a basis for the last year of the project. In this report we make use of journal, book and conference publications that have been evolved within the ECOVISION project in the last year, namely [77, 80, 72, 95, 96, 74]. Because of the high degree of linkage between the workpackages and also to allow a self-contend structure some aspects that have been already addresses in the last report are repeated in a modified form in chapter 2 and chapter 6.

In chapter 2 we describe the local feature processing in terms of multi-modal image Primitives (i.e., the receptive field structure) that has been motivated by hypercolumn-structures in V1. We discuss the analogy to hypercolumns and the role of the Primitives in a largely intertwined system. We have also added one important aspect to the Primitives that addresses the homogeneous-ness, edge-ness, or corner-ness of the image-patch. For that we found a continous measure in terms of the intrinsic dimension of the image patch that is described in chapter 3. In chapter 4 we extend the image Primitives to 3D-Primitives. This is done by finding correspondences between image Primitives in the left and right image. In the last year we have especially focussed on the interaction of stereo and optic flow that has been published in [95] which represents also an example of a static inference across visual modalities (for the application of dybamic recursions see chapter 5).

Rigid Body Motion is (beside statistical regularities addressed in chapter 7) an important regularity in visual data. Rigid-Body Motion (RBM) is directly applicable to 3D entities

(therefore as well to our 3D Primitives). It can be used to describe the change of visual entities over time. Its estimation from image correspondences has been an active research issue over the last decades. In the context of grouping it is essential that different kind of image correspondencs (such as point and line correspondences) can be mixed since groups usually consist of different kind of visual entities. The issue of grouping and RBM is at the core of the ECOVISION project. It requires the integration and utilization of statistical and determinsitic interdependencies. We therefore have devoted a review paper to this issue in which the RBM algorithm (that has been developped by Bodo Rosenhahn and Oliver Granert) that we use in our system is discussed in respect to grouping. We show that matematically not trivial problem to be capable to deal with different kind of correspondences makes this algorithm especially interesting. In chapter 5 we give a detailed description of the mathematical problems as well as the specific solution found by Bodo Rosenhahn et al.
The use of RBM for feature integration in dynamic recursions is addressed in chapter 6. This is an extension of the work described in [75]. In the last year we have made progress in different aspects. First, we have made use of the multi-modality of our Primitives for motion-integartion. Second, we have addressed the aspacts of metrics in different spaces (image-domain, stereo-domain and 3D space) and in this way have come to a better solution than in [75]. Thirdly, we have worked with out-door scenes recorded on co-operation with Hella. In [75] we hace worked with indoor-scenes in which the main object showed only limited variation in depth. However, in the out-door scens there occurs naturally large depth variation have. Both issues, multi-modality as well as the metric were essential to deal with the more complex data.
Finally, in chapter 7 we describe preparations for the representation of higher level segments. We use the linkage of collinear image Primitives for improvement of stereo processing. In the next year we will extend this work such that higher level segments in terms of groups of Primitives emerge that become established by multi-modal spatial-temporal recursions.

Chapter 2

Multi-modal Image Primitives

2.1 Introduction

In this chapter, we describe a new kind of image representation in terms of local multi–modal Primitives (see Figure 7.2). These Primitives are motivated by processing in the human visual system as well as by functional considerations. The work described here has been evolved from a project started in 1998 which has been focused on the integration of visual information [84]. The image representation described here is now a central pillar of the ongoing European project [24] that focuses on the functional modelling of early visual processes.

In the human visual system beside local orientation also other modalities such as colour and optic flow (that are also part of our multi–modal Primitives) are computed in the hyper-columns of V1 [53, 38]. *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 modelled analytically [54]. 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 [4, 61].

However, the human visual system acquires visual representations which allow for 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 [46] and by utilising spatial and temporal interdependencies* [92, 47]. This integration is manifested in the huge connectivity between brain areas in which the different visual modalities are processed as well as in the

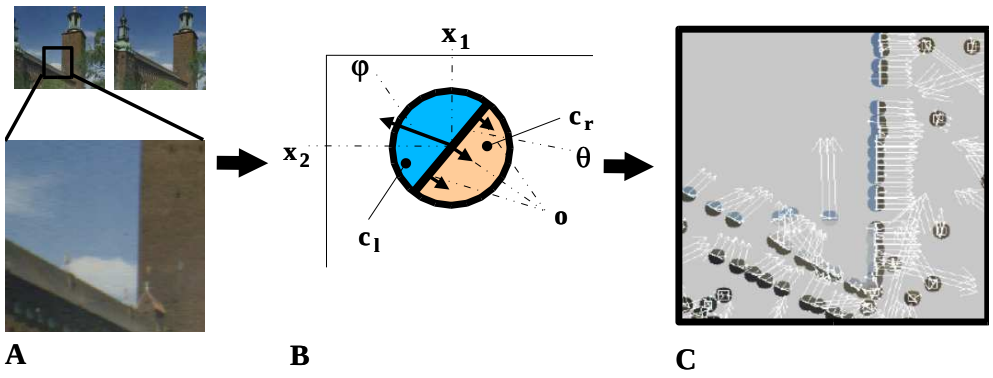


Figure 2.1: **left:** Image sequence and frame. **middle:** Schematic representation of the multi–modal Primitives. **right:** Extracted Primitives at position with high amplitude.

large number of feedback connections from higher to lower cortical areas [38]. The essential need for integrating visual information in addition to optimising single modalities to design efficient artificial visual systems has also been recognised in the computer vision community after a long period of work on improving single modalities [1].
However, integration of information makes it necessary that local feature extraction is subject to modification by contextual influences. As a consequence *adaptability* must be an essential property of the visual representation. Moreover, the exchange of information between visual events 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 the reason why we are after a *condensed* description of a local image patch, which however *preserves the relevant information*. Here relevance has to be understood not only in an information theoretical sense, but in a global sense (the system has to be subject to modifications by global interdependencies, in particular local entities have to be connectable to more complex entities) and action oriented sense (the transfered information has to be relevant for the actions the individual has to perform).
Taking the above mentioned considerations into account, the Primitives, which are the basic entities of our image representation, can be characterised by four properties:

Multi-modality: Different domains that describe different kinds of structures in visual data are well established in human vision and computer vision. For example, a local edge can be analysed by local feature attributes such as orientation or energy in certain frequency bands [81]. In addition, we can distinguish between line and step–edge like structures (contrast transition). Furthermore, colour can be associated

to the edge. This image patch also changes in time due to ego-motion or object motion. Therefore time specific features such as a 2D velocity vector (optic flow) are associated to our Primitives (see Figure 7.2).

Adaptability: Since the interpretation of local image patches in terms of the above mentioned attributes as well as classifications such as 'edge–ness' or 'junction–ness' are necessarily ambiguous when based on local processing [72], stable interpretations can only be achieved *through integration* by making use of contextual information [1]. Therefore, all attributes of our Primitives are equipped with a confidence that is essentially *adaptable according to contextual information* expressing the reliability of the attribute. Furthermore, feature attributes themselves are subject to correction mechanisms that use contextual information.

Condensation: Integration of information requires *communication between Primitives* expressing spatial [79, 73] and temporal dependencies [70]. This communication has necessarily to be paid for with a certain cost (as will be made explicit in section 2.3). This cost can be reduced by limiting the amount of information transferred from one place to the other, i.e., by reducing the bandwidth. Therefore we are after a *condensed* representation. Also for other tasks it is essential to store information in a *condensed way*, e.g., for the learning of objects to reduce memory requirements.

Meaningfulness: Communication and memorisation not only require a reduction of information. We want to reduce the amount of information within an image patch *while preserving perceptually relevant information*. This leads to *meaningful* descriptors such as our attributes position, orientation, contrast transition, colour and optic flow.

We will describe our feature processing in section 2.2 and will compare it to early human visual processing in Section 2.3.

2.2 Feature Processing and Application

In this section we describe the coding of modalities associated to our Primitives. In addition to the position $\mathbf{x}$, we compute the following semantic attributes and associate them to our Primitives (see also Figure 7.2).

Frequency: We describe the signal on different frequency levels f independently. Often the decision in which frequency band the relevant information does occur is difficult, therefore we leave this decision open to be decided at later stages of processing. It may be even that for the same position on different frequency levels there occur different kinds of

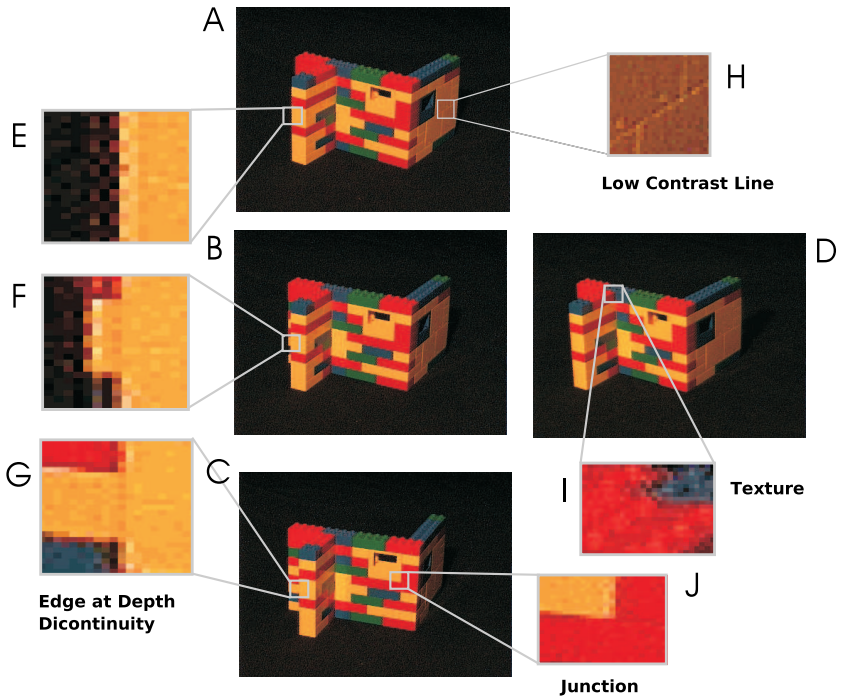


Figure 2.2: Examples of edge structures in an image sequence.

semantic information (for example, the top of the toy in Figure 2.2A on a high frequency level can be described as texture–like while on a lower frequency level it resembles an edge).

Orientation: The local orientation associated to the image patch is described by θ. The orientation θ is computed by interpolating across the orientation information of the whole image patch to achieve a more reliable estimate. This holds also true for the following feature attributes contrast transition, colour and optic flow.

Contrast transition: The contrast transition is coded in the phase φ of the applied filter [33]. The phase codes the local symmetry, for example a bright line on a dark background has phase 0 while a bright/dark edge has phase $-\pi/2$ (in Figure 2.3 the line that marks the border of the street is represented as a line or two edges depending on the distance from the camera). In case of boundaries of objects, the phase represents a description of the transition between object and background [67, 79].

Colour: Colour $(\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r)$ is processed by integrating over image patches in coincidence with their edge structure (i.e., integrating separately over the left $(\mathbf{c}^l)$ and right $(\mathbf{c}^r)$ side of the edge as well as a middle strip $(\mathbf{c}^m)$ in case of a line structure). In case of a boundary

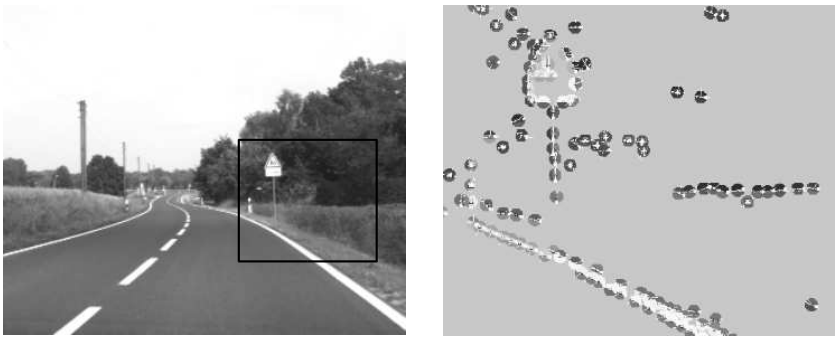


Figure 2.3: A: Original Image. B: Extracted Primitives with high amplitude.

edge of a moving object at least the colour at one side of the edge is expected to be stable (see Figure 2.2E–G) since it represents a description of the object.

Optic Flow: Local displacements $\mathbf{o}$ is computed by the well known optic flow technique [87].

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

$$E = (\mathbf{x}, f, \theta, \varphi, (\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r), \mathbf{o}; c).$$

In addition, 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 that is also subject to contextual adaptation.
We have applied our image representation to different contexts. First, an image patch also describes a certain region of the 3D space and therefore 3D attributes can be associated such as a 3D-position and a 3D-direction. In [73, 95], we have defined a stereo similarity function that makes use of multiple-modalities to enhance matching performance. Second, the Primitives can be subject to spatial contextual modification. We define groups of Primitives based on a purely statistical criterion in [79]. Once these groups are defined, we modulate the confidences of our Primitives: confidences are increased if the Primitives are part of a bigger group, otherwise the confidences are decreased. Thirdly, we have stabilised features according to the temporal context. In [70, 75], we make use of the motion of an object to predict feature occurrences and showed that we can stabilise stereo processing by modifying the confidences according to the temporal context.

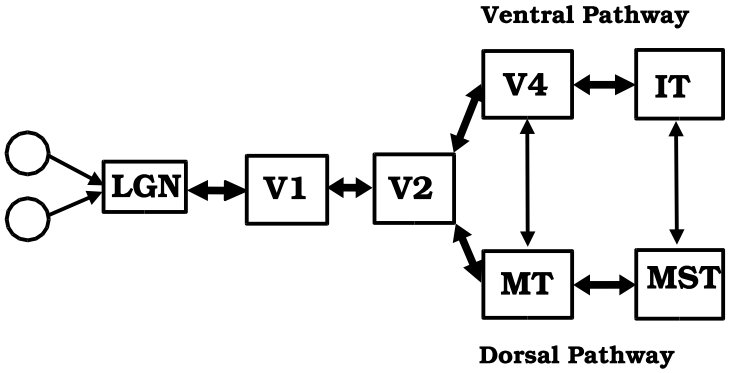


Figure 2.4: Flow of visual information in the human visual system (schematic).

2.3 Hyper-columns of Basic Processing Units in early Vision

In this section, we discuss aspects of the processing of visual information in the human visual system and draw analogies to our image representation.
The main stream of visual information in the human visual system goes from the two eyes to the LGN (Lateral Geniculate Nucleus) and then to area V1 in the cortex (see Figure 2.4 and [122]). There are two kinds of cell types involved (M (magnocellular) and P (parvocellular) cells) that have different response characteristics: M cells have a low spatial but high temporal resolution and are not colour sensitive. In contrast to M cells, P cells have a low temporal and high spatial resolution and are colour sensitive. Both kinds of cells project into two cortical pathways, the dorsal and ventral pathway (see Figure 2.4). The ventral pathway goes from the cortical area V1 to V2 to the Inferior Temporal Area (IT) and is believed to be mainly responsible for object recognition [112]. In the dorsal stream information is transferred from V1 to MT (Middle Temporal Area) to MST (Medial Superior Temporal Area) and is believed to be involved in the analysis of motion and spatial information.
V1 (or Visual Area 1) is the main input of both pathways. The structure of V1 has been investigated by Hubel and Wiesel in their ground-breaking work [52, 53]. V1 is organised in a retinotopic map that has a specific repetitively occurring pattern of substructures called hyper-columns. Hyper-columns themselves contain so called orientation columns and blobs (see Figure 2.5). The main input of V1 comes from the LGN and targets to layer 4 to which information of both eyes projects (see Figure 2.5Aiii).
The orientation columns are organised in an ordered way such that columns representing similar orientations tend to be adjacent to each other (see Figure 2.5Ai). However, it is

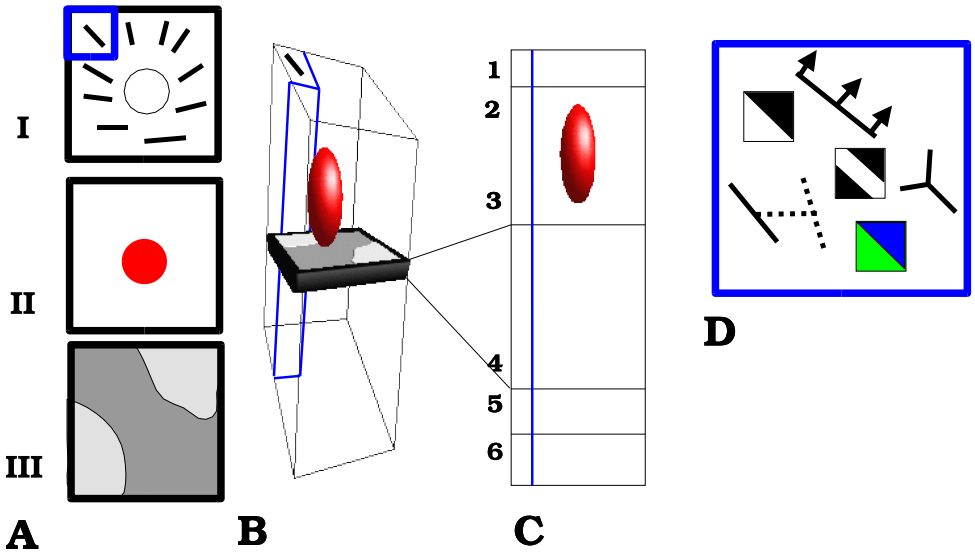


Figure 2.5: Hyper-columns in V1. A: There exist three physiological distinguishable substructure in a hyper-column: (i) in orientation columns information about oriented edge structure is represented in a topological way. (ii) Colour information is coded in so called 'blobs'. (iii) Information of both eyes are input to the fourth cortical layer (see also B). B: three–dimensional structure of a hyper-column. C: organisation in cortical layers. D: feature attributes that are coded in a hyper-column.

not only orientation that is processed in an orientation column but the cells are sensitive to additional attributes (see Figure 2.5D) such as disparity [6, 91], local motion [123], colour [53] and phase [57]. Also specific responses to junction–like structures could be measured [108]. Therefore, it is believed that in V1 basic local feature descriptions are processed similar to the feature attributes coded in our Primitives. However, since the processing is local,[1] the ambiguities of visual information is not resolved at this level. For example, response properties of neurons in V1 reflect the aperture problem [111]. This holds also for our Primitives since the flow is also computed by a local operation.

It is believed that mainly form is processed in the ventral pathway. Neurophysiological equivalents of illusionary contours can be detected in V2 but not in V1 [118]. This is not surprising since illusionary contours like in the Kanizsa triangle [59] presuppose an integration of information across a large spatial domain as well as across different feature types (e.g., edges and junctions) and can therefore only be processed at a later stage.

The different visual modalities are not computed independently but are combined. For

[1]There is a high connectivity within a hyper-column. There exist also connections across hyper-columns. However their distribution falls sharply with distance.

example in V1 the processing of motion is necessarily intertwined with the processing of orientation because of the aperture problem. In V4, colour and orientation is combined [123]. Accordingly, in our image representation the coding of colour is deeply intertwined with the coding of orientation. Colour is a feature that describes homogeneous surfaces. However, orientation describes discontinuities and can be used to separate the surfaces. In our image representation we therefore first compute orientation and then compute a left and a right colour according to this orientation.
In the dorsal pathway mainly motion is analysed. Like the occurrence of illusionary contours presuppose global interactions, the aperture problem can only be solved by taking the global context into account. This does not happen (and can not happen because of the local processing) in V1. However, in MT and MST many cell responses indicate a solution to aperture problem [89, 123]. Similar to the cells in V1, our Primitives also reflect the aperture problem. However, we can use the output of our Primitives to apply global mechanisms that disambiguate the local flow.
As in the ventral pathway, cells in the dorsal pathway show multi-modal response patterns. For example, a moving edge may not be visible as a luminance edge but can be constituted by colour or texture. MT cells respond to these kinds of structures although they are not sensitive to colour alone [114, 123].
Let us summarise. In V1 visual information is mainly locally processed. However, some semi–local interactions exist. The ambiguities of visual information can not be resolved at this stage of processing. A specialisation to form processing (along the ventral pathway V1–V2–V4–IT) and motion processing (along the dorsal pathway V1–V2–MT–MST) does occur.
As mentioned above, stable and reliable information can only be achieved by disambiguation through integration. However, this integration process makes the exchange of information within and across visual areas mandatory. As discussed before, intra–areal connections are very limited. However, inter–areal connection project to a much wider field of the next layer.
Regarding communication between visual areas we have to address two issues:

1) What is the bandwidth of information we want to transfer ("quantity")?

2) What kind of information do we want to communicate ("quality")?

The first question leads to a reflection about costs of communication. In any communication system transfer of information is associated to a cost which normally increases with the amount of information to be transferred and with the distance to be covered. This could concern the costs of "cables" but also the cost of the energy used for the transfer [3].
In the brain, the communication between two neurons is realized by an axon docking to the soma or the dendrites of other neurons. Accordingly, the complexity and, thus, the

"cost" of communication increases with the number of connections. This holds in a very general sense and may have been one driving force for the bandwidth reduction that is actually observed in neuronal visual processing. This bandwidth reduction most clearly manifests itself in mechanisms of visual attention and visual awareness. Focused attention is often taken as one central mechanisms used to reduce the bandwidth of computation as well as of information transfer in the brain to a manageable degree. Anatomically the bandwidth limitation requirement may be reflected by the density of fibres which connect different areas which is smaller than that which connects cells within a hyper-column.
A similar mechanism is also used in our image representation were we arrive at a significant reduction of information following the first processing stages. Compared to an average sized image patch of 15×15 pixels represented by a Primitive the output of a Primitive has less than 20 values, i.e., we have a compression rate of more than 96%. This rate becomes even higher when we compare the output of a Primitive to intermediate local stages of processing where feature attributes for all modalities are derived for each pixel.
The second question above concerns the quality of information which needs to be transferred between the different stages of visual processing. Here we refer back to what we have said above noting that pre-processed visual information is exceedingly ambiguous as the consequence of fundamental problems in image data acquisition as well as resulting from the intrinsic structure of the detectors (receptive fields). This leads to the situation that redundant information must be transferred because only through redundancy it can be assured that erroneous information can be disambiguated. For this it is required that a visual event which is represented by the firing of neuron A has a relevance for the event represented by B. Since event A is supposed to be used to correct event B both events need to be highly correlated. This can be quantified by the following measure of statistical interdependencies:

$$\frac{P(B|A)}{P(B)}. \tag{2.1}$$

If this term takes a high value then there is a high likelihood of the occurrence of event B when we know event A has occurred compared to the likelihood of the occurrence of the event B without prior knowledge. In this case, events A and B can be used to mutually correct each other because they are carrying shared (i.e., redundant) information. The expression (2.1) has been called 'Gestalt coefficient' in [68] where it was shown that applying binarised Gabor wavelets to natural images, a high Gestalt coefficient corresponds to the Gestalt laws Collinearity and Parallelism. As an extension of [68], it has been shown in [79] that by using our multi–modal Primitives we can increase the statistical interdependencies measured by (2.1) significantly compared to using orientation only [68]. That means that by using our Primitives we can increase interdependencies of visual events. In this way in our Primitives not only information is condensed but transferred to *more meaningful descriptors*.

Chapter 3

A continuous formulation of intrinsic Dimension

3.1 Introduction

Natural images are dominated by specific local sub–structures, such as edges, junctions, or texture. Sub–domains of Computer Vision have analyzed these sub–structures by making use of certain conepts (such as, e.g., orientation, position, or texture gradient). These concepts were then utilized for a variety of tasks, such as, edge detection (see, e.g., [17]), junction classification (see, e.g., [100]), and texture interpretation (see, e.g., [97]). However, before interpreting image patches by such concepts we want know whether and how these apply. For example, the idea of orientation does make sense for edges or lines but not for a junction or most textures. As another example, the concept of position is different for a junction compared to an edge or an homogeneous image patch. For a junction the position can be unambiguously defined by the point of intersection of lines, for edges the aperture problem leads to a definition of the position as a one-dimensional manifold and for an homogeneous image patch it is impossible to define a position in terms of local signal attributes. Hence, before we apply concepts like orientation or position, we want to classify image patches according to their junction–ness, edge–ness or homogeneous–ness.
The intrinsic dimension (see, e.g., [125, 30]) has proven to be a suitable descriptor in this context. Homogeneous image patches have an intrinsic dimension of zero (i0D), edge–like structures are intrinsically 1–dimensional (i1D) while junctions and most textures have an intrinsic dimension of two (i2D). There exists also related classifications such as the rank of a image patch [41], the rank taking discrete values zero, one, or two. Another related formulation is the distinction between constant, simple and isotropic signals [55]. The association of intrinsic dimension to a local image structure has mostly be done by

a discrete classification [125, 30, 55]. To our knowledge, so far there exists no continuous definition of intrinsic dimensionality that covers all three possible cases (i0D, i1D, and i2D). However, there exist attempts to find a continuous formulation between i1D and i2D signals [41].
In contrast to, e.g, curvature estimators (see, e.g., [7, 90]), the intrinsic dimensionality does not make any assumption about specific structural attributes of the signal but is is based a purely statistical criterion: The concept of curvature does make sense for curved lines but not for junctions or most complex textures. However, the intrinsic dimension is a sensible descriptor also for these kind of signals (see also [64]).
In section 3.2.1, we will show that the intrinsic dimension is a local descriptor that is spanned by two axes: one axis represents the variance of the spectral energy and one represents the a weighted variance in orientation. In this paper, we will review diverse definitions of intrinsic dimension. In section 3.2.2, we will show that they can be subsumed within the above mentioned scheme. Since the intrinisc dimension is a two–dimensional structure, no continuous one–dimensional definition is sensible. Moreover, we will show in section 3.2.1 *that the topological structure of instrinsic dimension essentially has the form of a triangle.* We will then give one possible concrete definition of intrinsic dimension that realizes its triangular structure in section 3.3.1.
A classification of edge–ness or corner–ness based on a local image patch without taking the context into account always faces the problem of the high degree of ambiguity of visual information (see, e.g., [1]). Taking into account this ambiguity we do not want to come to a final decision about the junction–ness of edge–ness of an image patch but we want to associate confidences to such classifications. Assigning confidences instead of binary decisions at low level stages of processing has been proven useful since it allows for stabilizing such local classifications according to the context (see, e.g., [1, 69]). By making use of barycentric coordinates (see, e.g., [20]), we will utilize the triangular structure of intrinsic dimension to express confidences for the different possible interpretation in section 3.3.2. This leads to *continuous definition of intrinsic dimensionality that covers i0D, i1D and i2D signals.* Finally, in section 3.4 we show examples of our continuous classification of image patches of different intrinsic dimension.
To our knowledge, this paper is the first work that makes the triangular structure of intrinsic dimensionality explicit and which gives a continuous definition that covers all three possible cases of intrinsic dimension.

3.2 The Concept of intrinsic Dimensionality

The *intrinsic dimensionality* in image processing is a formalization of what is commonly called "edgeness" vs. "junction–ness". The term intrinsic dimensionality itself is much more general. In [11], p. 314, it says that "a data set in d dimensions is said to have

an *intrinsic dimensionality* equal to d' if the data lies entirely within a d'-dimensional subspace", but indeed, the concept of intrinsic dimensionality is much older [116].
In image processing, the intrinsic dimensionality was introduced by [125] to define heuristically a discrete distinction between edge–like and corner–like structures. However, here we want to adopt the more general definition in [11] to image processing. For this, we have to consider the *spectrum* of an image patch (see figure 3.1):

- if the spectrum is concentrated in a point[1], the image patch has an intrinsic dimensionality of null (i0D),
- if the spectrum is concentrated in a line[2], the image patch has an intrinsic dimensionality of one (i1D), and
- otherwise the image patch has an intrinsic dimensionality of two (i2D).

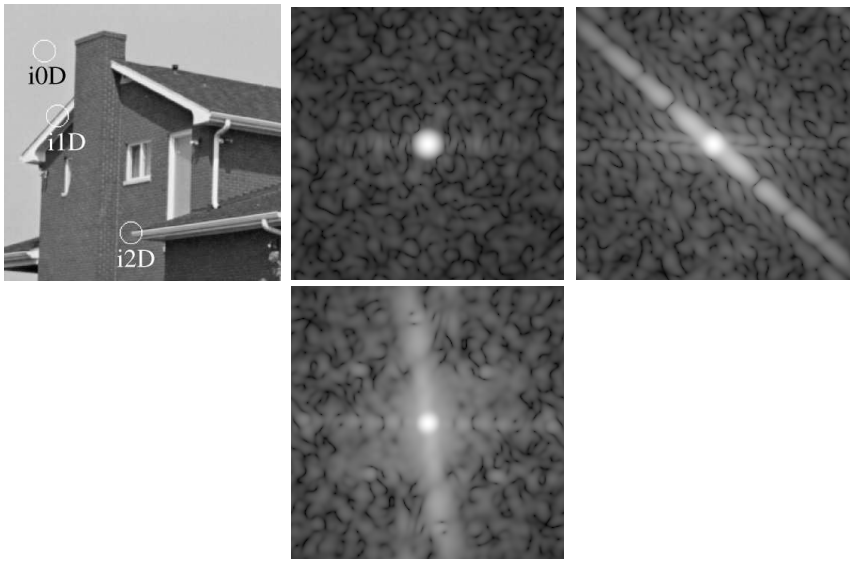


Figure 3.1: Illustration intrinsic dimensionality. In the image on the left, three neighborhoods with different intrinsic dimensionalities are indicated. The other three images show the local spectra of these neighborhoods, from left to right: i0D, i1D, and i2D.

Each of these three cases can be characterized more vividly. Constant image patches correspond to i0D patches. Edges, lines, and sinusoid-like textures obtained by projecting

[1]Note that due to the Hermitian spectrum of a (real valued) image, this point can only be the origin, i.e., the DC component.
[2]With the same argument as in footnote 1, this line goes through the origin.

1D functions (simple signals [41]) correspond to i1D patches. All other structures like corners, junctions, complex textures, and noise correspond to i2D patches.
Taking a closer look at the concept of intrinsic dimensionality, two fundamental problems pop up:

1. The intrinsic dimensionality as it is defined above is a discrete feature in {i0D, i1D, i2D}. However, every real signal consists of a combination of intrinsic dimensionalities – there are hardly any totally constant or ideal i1D image patches in real images. Hence, we would like to have a *continuous* definition of intrinsic dimensionality.

2. The topology of the iD-space is yet undefined. In case of a discrete space, the relations between the different intrinsic dimensionalities is obvious, all dimensionalities are mutually adjacent. The topology of the continuous iD-space is considered in the subsequent section.

In the following section we discuss a new model for representing the intrinsic dimensionality in a continuous, topologically appropriate way. The subsequent section gives an overview of known methods for estimating the intrinsic dimensionality and relates them to our new model.

3.2.1 The Intrinsic Dimensionality has a 2D Triangular Structure

For the estimation of the intrinsic dimensionality of an image patch, we need to apply a measure for the spread of the spectral data, either to a point or to a line. The classical approach from statistics for such a measure is the *variance* of the data. Since a change of the coordinate system results in new stochastic variables, the computation of the variance depends on the coordinate system, for instance in Cartesian coordinates vs. polar coordinates. Different coordinate systems lead to further diversification of practical approaches.
To be more concrete, the variance of the spectral data with respect to the origin in a cartesian coordinate system is defined by

$$\sigma_{\mathrm{O}}^2 = \frac{1}{N} \int_{\Omega} |\mathbf{u}|^2 |F(\mathbf{u})|^2 \, d\mathbf{u} \ , \tag{3.1}$$

where $\mathbf{u}$ is the frequency vector, Ω is the region of integration in the Fourier domain[3] and

$$N = \int_{\Omega} |F(\mathbf{u})|^2 \, d\mathbf{u} \tag{3.2}$$

[3] In practice, this is mostly a Gaussian window, i.e., we consider the windowed Fourier transform (2D version of the short-time Fourier transform).

is a normalization constant. The variance with respect to a line is given by

$$\sigma_{\mathrm{L}}^2 = *min_{\mathbf{n}} \frac{1}{N} \int_{\Omega} |\mathbf{n}^T \mathbf{u}|^2 |F(\mathbf{u})|^2 \, d\mathbf{u} \ , \tag{3.3}$$

where $\mathbf{n}$ is obtained to be parallel to i1D signals, i.e., it represents the orientation. The variance σ_{O}^2 defines some kind of measure of the local grey level variation whereas the the variances σ_{L}^2 reflects the dynamic perpendicular to the main orientation.
If we change to polar coordinates $\mathbf{u} \mapsto (q, \theta)$, we get two new variances, the radial variance

$$\sigma_{\mathrm{R}}^2 = \frac{1}{N'} \int_0^Q q^2 \int_0^{2\pi} |F(q\cos\theta, q\sin\theta)|^2 \, d\theta \, dq \ , \tag{3.4}$$

where Q is the radius of Ω, and the angular variance

$$\sigma_{\mathrm{A}}^2 = *min_{\theta_0} \frac{1}{N'} \int_{\theta_0-\pi}^{\theta_0+\pi} (\theta - \theta_0)^2 \int_0^Q |F(q\cos(\theta-\theta_0), q\sin(\theta-\theta_0))|^2 \, dq \, d\theta \ , \tag{3.5}$$

where the normalization constant N' is given similar to eq:norm, performing the integration in polar coordinates. The angle θ_0 represents the local orientation.
The two characterizations $(\sigma_{\mathrm{O}}^2, \sigma_{\mathrm{L}}^2)$ and $(\sigma_{\mathrm{R}}^2, \sigma_{\mathrm{A}}^2)$ are different in detail, but related. The most important difference between the two variances σ_{O}^2 and σ_{R}^2 is the different weighting of the frequency components due to the missing Jacobian of the coordinate transform. The two variances σ_{L}^2 and σ_{A}^2 differ more essentially, since σ_{A}^2 becomes undefined for $\sigma_{\mathrm{R}}^2 = 0$. The orientation variance σ_{A}^2 corresponds to formulations of *intensity invariant measures of the 'i1D-ness'*. This intensity invariance prevents a probabilistic, triangular formulation of intrinsic dimensionality since the i0D case is neglected. Examples for such traditional, intensity invariant measures are the coherence [55] (see also next section) and the isotropy factor [30].
For the idealized cases of purely i0D, i1D, and i2D signals, the variances obviously behave as given in table 3.1. By a proper normalization ($\tilde{\sigma}_{\mathrm{O}}^2 = k\sigma_{\mathrm{O}}^2$, etc.), the entries "large"

intrinsic dimensionality	i0D	i1D	i2D
σ_{O}^2	0	large	large
σ_{L}^2	0	0	large
σ_{R}^2	0	large	large
σ_{A}^2	undefined	0	large

Table 3.1: Intrinsic dimensionality and variances. A zero variance means a ideal concentration of the spectral data.

in this table can be replaced by "1", yielding an overall range of $[0,1] \times [0,1]$, i.e., the

iD-space spanned by $\tilde{\sigma}_R^2$ and $\tilde{\sigma}_A^2$ corresponds to a *2D square*. The entry "undefined", however, cannot be simply replaced by a value between zero and one, all values coexist with the same right. In other words, one edge of the square is singular. To solve for this singular edge, a straightforward idea is to multiply $\tilde{\sigma}_A^2$ with $\tilde{\sigma}_R^2$ (see section 3.3), which can be considered as a replacement for the Jacobian, i.e., we consider a space similar to $(\tilde{\sigma}_O^2, \tilde{\sigma}_L^2)$ instead. Since the $(\tilde{\sigma}_R^2, \tilde{\sigma}_A^2)$ space is a 2D square, we obtain a *2D triangle* for $(\tilde{\sigma}_O^2, \tilde{\sigma}_L^2)$, see figure 3.2.

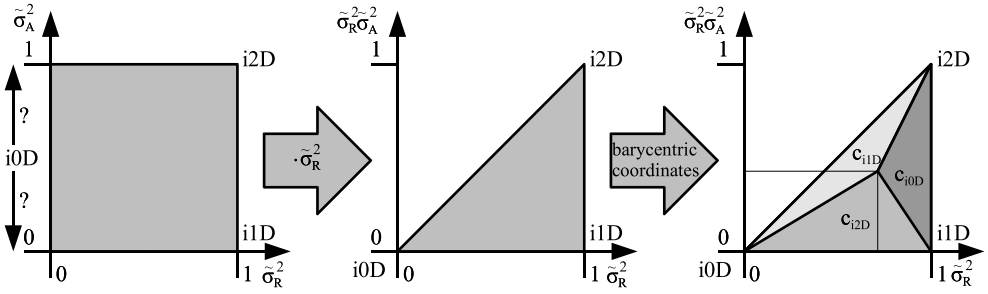


Figure 3.2: About the topology of iD-space. Left: traditional iD-space (square), center: our iD-space (triangle), right: parametrization of the iD-triangle by barycentric coordinates.

Each of the corners of the triangle corresponds to a certain intrinsic dimensionality. The topology of the triangle allows to vary the intrinsic dimensionality continuously from any case to any other case. This observation is very important in practice since, as stated further above, every real signal consists of a combination of intrinsic dimensionalities. The parameterization of the iD-triangle is described in detail in section 3.3.

3.2.2 Approaches for Estimating the Intrinsic Dimensionality

The various approaches which occurred in the literature so far mainly differ with respect to two aspects: (1) the computation of the variances and (2) the coordinate system. Nearly all systematic approaches to measure the intrinsic dimensionality are known as or are equivalent to the *structure tensor* [10, 37].
Basically, the variances can either be computed by outer products of first order derivatives or by combinations of quadrature filter responses, see [41, 55] for an overview. There are other but still related methods, e.g., polynomial expansions [27] and higher order spherical harmonics [32]. Most approaches make use of Cartesian coordinates to compute and to represent the variances, but an evaluation in polar coordinates is at least a plausible alternative, see section 3.3 and [22].

The first approach to what is nowadays called structure tensor is based on averaging the outer product of derivatives. This method was independently invented by Bigün and Granlund [10] and Förstner and Gülch [37]. In [35] a deeper analysis of the structure tensor from a statistical point of view is developed. The idea is to approximate the auto-covariance function by a truncated Taylor series expansion in the origin. The term which is obtained by this expansion is given by

$$\mathbf{J} = \int_{\Omega} \mathbf{u}\mathbf{u}^T |F(\mathbf{u})|^2 \, d\mathbf{u} \ . \tag{3.6}$$

Applying the power theorem [14] and the derivative theorem, we end up with

$$\mathbf{J} = \int_{\omega} (\nabla f(\mathbf{x}))(\nabla f(\mathbf{x}))^T \, d\mathbf{x} \ , \tag{3.7}$$

where ω is a local region of integration in the spatial domain (a Gaussian window in case of a windowed Fourier transform). This tensor $\mathbf{J}$ can be visualized by an ellipse, where the length of the two main axes correspond to the two eigenvalues λ_1 and λ_2 of the tensor. The mean of the two eigenvalues (the trace of the tensor) corresponds to the variance with respect to the origin σ_{O}^2, and the smaller eigenvalue λ_2 corresponds to the line-variance σ_{L}^2. Therefore, the two axes of the iD-triangle are given by $(\lambda_1 + \lambda_2)/2$ and λ_2 and an appropriate normalization.
The tensor feature which is typically used in context of estimating the intrinsic dimensionality is the *coherence*[4] [10]:

$$c = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2} \ . \tag{3.8}$$

The coherence c is related to the variances σ_{O}^2 and σ_{L}^2 (and to $\tilde{\sigma}_{\mathrm{A}}^2$) by $c = 1 - 2\sigma_{\mathrm{L}}^2/\sigma_{\mathrm{O}}^2 \approx 1 - \tilde{\sigma}_{\mathrm{A}}^2$. A common method for distinguishing i1D and i2D structures is to threshold the coherence. This is also the theoretic background of the Harris-Stephens corner detector [43]. A drawback of all coherence-based methods is *that an additional energy threshold has to be applied in order to single out constant (i1D) regions.* Our new triangle model (figure 3.2, section 3.3) allows us to postprocess the iD information without applying threshold at this early step.
A method which is related to the structure tensor but which is different in detail is based on generalized 2D quadrature filters [32]. The idea of this approach is to compute responses of steerable quadrature filters which are adapted to the main orientation of the signal and to the perpendicular orientation. The filters are polar separable and window out the information in the respectively perpendicular orientation. The effective amplitude response of the filter set is isotropic.

[4] In [37] the coherence is squared, which is unnecessary if the eigenvalues are ordered.

The resulting feature vector consists of five features among which we find the local orientation, the local amplitude with respect to the local orientation A_{M}, and the local amplitude perpendicular to the local orientation A_{m}. Local amplitudes computed by a quadrature filter are related to variances in the Fourier domain. Assuming that a quadrature filter has a sufficiently small bandwidth, the filter output approximates the Fourier components at the center frequency [41], page 171. Hence, the local amplitude increases with increasing variance of the spectrum.
The squareroot of the ratio of the two amplitudes $c_{\mathrm{I}} = \sqrt{A_{\mathrm{m}}/A_{\mathrm{M}}}$ is called *isotropy factor* and it corresponds to $\sqrt{(1-c)/(1+c)}$ [30]. Hence, the orientation variance $\tilde{\sigma}_{\mathrm{A}}^2$ is given by

$$\tilde{\sigma}_{\mathrm{A}}^2 \approx 2\frac{\sigma_{\mathrm{L}}^2}{\sigma_{\mathrm{O}}^2} = 1 - c = \frac{2c_{\mathrm{I}}^2}{1+c_{\mathrm{I}}^2} = \frac{2A_{\mathrm{m}}}{A_{\mathrm{M}}+A_{\mathrm{m}}} \quad .$$

Due to the isotropy of the filter set, the mean of the two amplitudes corresponds to a total local amplitude of the signal and hence to the variance σ_{O}^2. Therefore, after normalization the amplitudes can be used for parameterizing the iD-triangle: $(A_{\mathrm{M}} + A_{\mathrm{m}})/2$ as the first coordinate and A_{m} as the second coordinate. Indeed, evaluating A_{m} and applying a threshold has been used for corner detection in [32].

3.3 Triangular Definition of intrinsic Dimension

Having shown in section 3.2.1 that the topological structure of intrinsic dimensionality is essentially a triangle, we now derive a realization of intrinsic dimensionality that makes use of its triangular structure. Instead of a binary classification (as done, e.g., in [125, 30, 55]), we compute 3 values $c_{0D}, c_{1D}, c_{2D}, c_i \in [0,1]$ that code confidences for the intrinsic 0–dimensionality, intrinsic 1–dimensionality and intrinsic 2–dimensionality of the signal.
In section 3.3.1 we will concretize the origin–variance and line–variance introduced in section 3.2.1 and use these measures to span a triangle whose corners represent the extremes of purely i0D, i1D and i2D signals. In section 3.3.2 we then use barycentric coordinates to assign the confidences.

3.3.1 Local Amplitude and Orientation Variance as two axes spanning the Triangle

Our image processing starts with a filter operation which is based on generalized quadrature filters [33]. These filters perform a *split of identity*, i.e., the signal becomes orthogonally divided into its amplitude m (indicating the likelihood of the presence of a structure), its geometric information (orientation) θ and its phase φ.
We express our realization of the intrinsic dimenionality triangle in polar coordinates. To compute the origin–variance we first apply a normalization function N that transfers

the amplitude m that has values in $[0, \infty]$ to the interval $[0, 1]$ by performing a smooth thresholding using a sigmoidal function. The shape of the sigmoid function does depend on the local and global contrast. In this way even at low contrast image patches image structures can be detected.[5] Assuming a sufficiently small bandwidth of our filters, our measure for the origin–variance at a pixel position $\mathbf{x}_0$ is simply given by the normalized amplitude (see section 3.2.2):

$$\hat{\sigma}_R = N(m(\mathbf{x}_0)).$$

To compute our measure for line–variance at pixel position $\mathbf{x}_0$, we compute a weighted variance measure of the local orientation. First, we define a set $A(\mathbf{x}_0)$ representing the local neighbourhood of $\mathbf{x}_0$ and we compute the mean orientation $E_A[\theta]$ on A. Weighting is performed according to the normalized magnitude. Our measure for line variance then becomes

$$\hat{\sigma}_L = \tilde{\sigma}_{\mathrm{A}}^2 \cdot \tilde{\sigma}_{\mathrm{R}}^2 = \sum_{\mathbf{x} \in A} \left(N\left(m(\mathbf{x})\right) \cdot d\left(\theta(\mathbf{x}), E_A[\theta]\right)\right).$$

Note that $\sum_{\mathbf{x} \in A} d\left(\theta(\mathbf{x}), E_A[\theta]\right)$ basically represent $\tilde{\sigma}_{\mathrm{A}}^2$ and the mutiplication with $N\left(m(\mathbf{x})\right)$ corresponds to the mutiplication with $\tilde{\sigma}_{\mathrm{R}}^2$.
The metric d takes the singularity of the orientation at 0 and π into account and performs a normalisation that ensures that $\hat{\sigma}_L$ takes values in $[0, 1]$. The measure $\hat{\sigma}_L$ defines the second axis of our triangle.
As a final step we apply the squashing function $f(x) = x^c$ to steer the distribution of values in [0,1]. Origin–variance and line variance are finally defined by

$$\sigma_O = \hat{\sigma}_0^{c_1}$$
$$\sigma_L = \hat{\sigma}_L^{c_2}$$

where the parameters $c_1 = \frac{1}{6}$ and $c_2 = \frac{1}{2}$ have to be proven useful. σ_0 and σ_L span the triangle (see figure 3.2). Note that by definition it holds $\sigma_L < \sigma_O$.
Since we have defined the axes of our triangle we can now associate the different intrinsic dimensions to its corners:
An intrinsically zero dimensional (i0D) image patch is characterized by a low origin variance ($\sigma_O \approx 0$). Then it also holds $\sigma_L \approx 0$ since $\sigma_L < \sigma_O$ by definition. In the triangle shown in figure 3.2 (right) intrinsically zero dimensional (i0D) image patches correspond to the coordinate $(0, 0)$. Although $m \approx 0$, the local image patch can also be a projection of a 3D–edge (that usually corresponds to i1D signals) or a junction (that usually corresponds to i2D signals). The low contrast may be caused by e.g., accidental background–object constellation or an accidental surface/illumination constellation. To account for these

[5] This normalization has been be proven to be useful in the object recognition system [**?**] where it is discussed in detail.

ambiguities we will (based on the representation introduced here) define confidences that express the likelihood of the signal being i0D, i1D or i2D.
An intrinsically one dimensional image patch is characterized by a high origin variance and a low line variance within the image patch. In the triangle in figure 3.2 (right) this corresponds to the coordinate (1,0). Note that orientation can only be meaningfully associated to an intrinsically one–dimensional signal patch. In contrast, for a homogenous image patch (i0D) or a junction (i2D) the concept of orientation does not make any sense. 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.
An intrinsically two dimensional image patch is characterized by high origin variance and high line variance. This corresponds to the coordinate $(1, 1)$ in the triangle shown in figure 3.2 (right). 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.

3.3.2 Coding intrinsic dimensionality by barycentric coordinates:

Having defined a triangle with its corners representing the extremes in intrinsic dimensionality, we can now code confidences associated to the different intrinsic dimensions (c_{0D}, c_{1D}, c_{2D}) by using barycentric coordinates (see, e.g., [20]). Given a point inside a triangle, the Barycentric coordinates describe twice the area of the triangle opposite to the corners of triangle (see figure 3.2).
A measurement of $\sigma 0$ and σ_{L}^2 defines a point inside the triangle $(0, 0), (0, 1), (1, 1)$:

$$\mathbf{p} = (p_x, p_y) = (\sigma_0, \sigma_L).$$

Our confidences are the barycentric coordinates of this point:

$$\begin{aligned} c_{0D} &= 1 - p_x \\ c_{1D} &= p_x - p_y \\ c_{2D} &= p_y \end{aligned}$$

Note that since $0 \leq p_y \leq p_x \leq 1$ and $p_x \in [0, 1]$ it holds $0 \leq c_i \leq 1$. The three confidences add up to one since

$$c_{0D} + c_{1D} + c_{2D} = (1 - p_x) + (p_x - p_y) + p_y = 1.$$

3.4 Simulations

We have applied our definition of intrinsic dimension within a new kind of image representation which is based on multi–modal Primitives (see, e.g. [77]). These Primitives carry information about orientation, colour, optic flow, depth in a condensed way and are used for scene analysis in the European project ECOVISION [24]. To all attributes in the different modalities confidences are associated that are subject to contextual modification. Our continous definition of intrinsic dimension is used as an additional descriptor that codes information about the edge–ness or junction–ness of the Primitive. This allows for, e.g., a use of orientation information for i1D structures only. Figure 3.3 shows the extracted Primitives from an image and for some of them the position in the triangular representation of the intrinsic dimensionality.
The continuous formulation of intrinsic dimension has a number of potential applications domains. For example, in optic flow analysis it can be used to distinguish between normal flow (a i1D signal patches) and potentially correct flow (at i2D image patches). The continuous formulation could allow for an appropriate weighting of flow vectors for global optic flow interpretation. Another example is the accumulation of ambiguous information over time (see, e.g., [70]). The continuous formulation would allow for the postponing of a final decision about edgen–ess or junction-ness to a rather later stage of processing that can make use of a number of time frames.
An extension of this work, in which the triangle formulation is extended to a cone representation allowing for a probabilsitic of image patches is described in [31].

Acknowledgment: We would like to thank Florentin Wörgötter for fruitful discussions. This workd has been funded by the European Project ECOVISION.

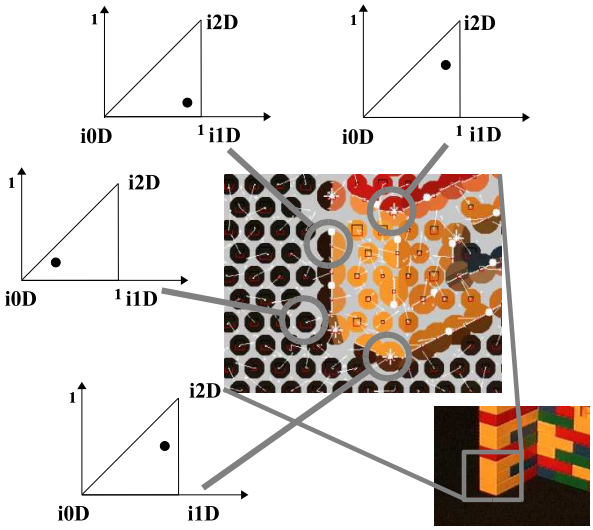


Figure 3.3: Primitives of different intrinsic dimensionality (i2D signals are indicated by a star and i1D signals by a line at its centers, i0D signals have no special indicator but have smaller radius. For some Primitives the triangular representation is shown.

Chapter 4

From 2D-Primitives to 3D-Primitives

4.1 Introduction

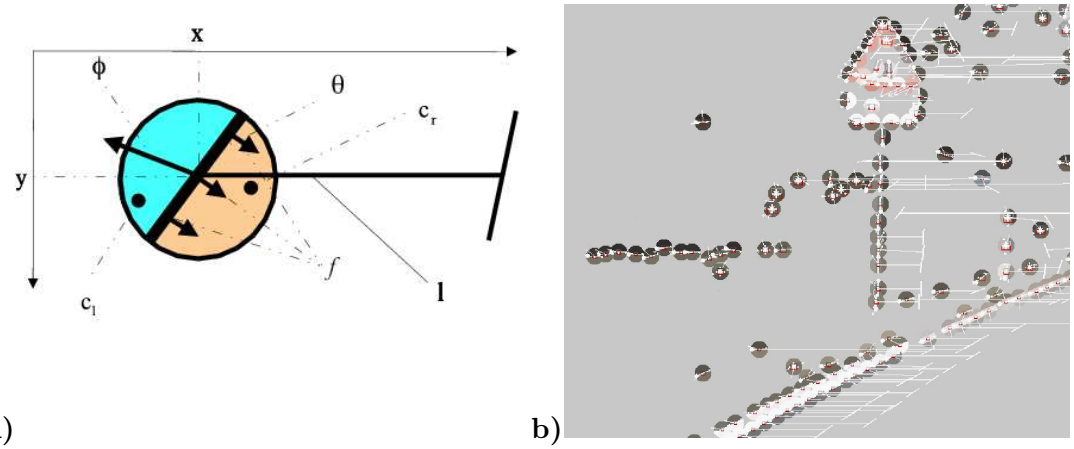
In stereo processing with calibrated cameras we can reconstruct a 3D point from two 2D point correspondences or from two corresponding 2D points with associated orientation, we can reconstruct a 3D point with associated 3D orientation, (e.g. [29, 107]). The problem at hand is to find correspondences between image structures in the left and right image.

To find correspondences, stereo similarity functions between image patches or features in the left and right image need to be defined. Some similarity functions use geometric attributes (such as, orientation or length) [4, 85]. 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. Alternatively to methods that use geometric information only for feature matching, some authors use both factors, orientation and structural information. For example, in [36] variations of the local image patches are taken into account explicitly by applying an affine transformation of the image patch grey values. 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. Making 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 [107] 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 homoge-

neous 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 that makes use of geometric and structural information in a direct way, *i.e.* without the need of solving a set of equations. To improve stereo matching we also use colour and temporal information. In [58, 65] it has been shown that the use of colour can improve stereo matching significantly. Our work confirms this result. Going beyond [58, 65], we are able to give a statement about the relative importance of colour compared to other visual modalities. We make further use of temporal information in terms of the optic flow.
Our similarity function is based on multi–modal image descriptors (see figure 4.1 and [76]) that covers geometric information (orientation), structural information (phase), colour and temporal information (optic flow). We will show that the use of multiple modalities improves stereo matching performance. Since our similarity function explicitly steers the influence of the different visual modalities, we are able to give concrete weights for their relative importance. We can also show that optimal weights are reasonably robust over different scenes.
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, 79, 21]) and integration over time (see, e.g. [28, 63, 107, 70]) 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 multiple aspects in visual scenes. We derive statements about the relative importance of the different visual aspects. Finally and most importantly, we show (by comparison to a normalized cross–correlation comparison) that our image representation leads to a condensation of information (up to a factor of 96.6%) while preserving the relevant information.
The paper is structured as following: In section 4.2, we briefly describe our feature processing. A distance function for optic flow vectors is described in section 4.3. Using this, we integrate the optic flow in a similarity function that also covers orientation, phase, and colour. This similarity functions allows us to steer explicitly the influence of the different visual attributes. The relative importance of orientation, phase, colour and optic flow is investigated in section 4.4.

4.2 Feature Processing

In this section we describe the processing of information (orientation, phase, colour and optic flow) used in our stereo algorithm. Note that in [79] the same kind of features are used to determine their statistical relationship in natural images.

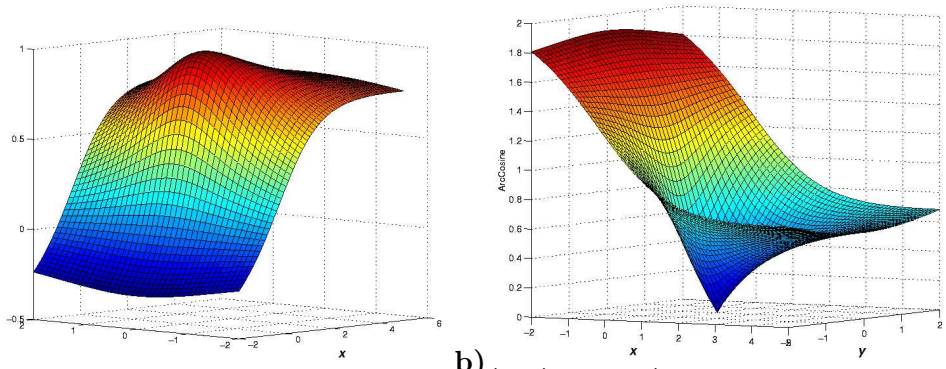


a) **b)**

Figure 4.1: **a)** Schematic representation of a basic feature vector. Position is coded by (x, y), orientation by θ, phase by ϕ, and colour by (c_l, c_r), the colour on both sides of the edge. l is the disparity between the Primitiveand its match in the other image. . **b)** Here the previously described Primitivesare extracted from an image. The white lines represent the disparities l for all the Primitivesand point to the position of the matching primitive in the other image.

We will use a systematic mathematical description of geometric and structural information of grey level images based on the monogenic signal [33]. The monogenic signal performs a *split of identity*, *i.e.* it orthogonally divides the signal into energetic information (indicating the likelihood of the presence of a structure), its orientation θ and its structure (expressed in the phase ϕ). Features are extracted in local image patches which position is parameterized by $X = (x, y)$ (see figure 4.1**a)**). In our simulations we only use features for which the variance of orientation within a small patch is below and the magnitude is above certain thresholds, *i.e.* features that correspond to image patches of intrinsic dimension close to one, since orientation and phase are only defined for intrinsicly one-dimensional signals (*c.f.* 3). The phase ϕ can be used to interpret the kind of contrast transition at this maximum [67], e.g., a phase of $\frac{\pi}{2}$ corresponds to a dark–bright edge, while a phase of zero 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. The local phase as additional feature allows us to code structural grey level information into account (as one parameter in addition to orientation) in a very compact way (see, e.g., [41, 67, 33]).

As it was shown by e.g. [58, 66], colour is also an important cue to improve stereo matching. The pixel data of the image contains the three components red, green and



a) **b)**

Figure 4.2: **a)** Graph of the dot product $\hat{f}_1 \cdot \hat{f}_2$, with $\hat{f}_1$ being the normalized 3D vector equivalent to the 2D vector $f_1 = (x, y)$, $x \in [-2, +5]$ and $y \in [-2, +2]$, and $\hat{f}_2$ being the normalized 3D vector equivalent to the 2D vector $f_2 = (1, 0)$: $\hat{f}_2 \simeq (0.7071, 0, 0.7071)$. **b)** Graph of the distance function $d(f_1, f_2)$.

blue. As we are already using the intensity information through the phase, we want a colour vector excluding this information. We decide to use the YUV colour space (cf. [105]), Y containing the intensity information, and U and V coding the colour. This allows us to reduce the colour information from 3 to 2 dimensions with a simple linear transformation.

The colour information of a Primitiveis defined by the colour on both sides of an edge, and, in the case of a line structure (if $\phi \simeq 0$ or $\phi \simeq \pi$), the colour of the line itself. The colour information vector is then $C = (c_l, c_m, c_r)$. The three component vectors $c_l = (c_U^l, c_V^l)$, $c_m = (c_U^m, c_V^m)$ and $c_r = (c_U^r, c_V^r)$ with $c_j^i \in [0, 1]$ hold the U and V values of the left side, the center and the right side of the edge. Consequently the colour information we are using is *6–dimensional*.

To this feature description we add the optic flow local measurement, using the well known *Nagel* algorithm (cf. [87]).

As a result we got a multimodal visual Primitivethat gives a rich but condensed description of a local image patch. For more details concerning this kind of image representation we refer to [**?**].

The resulting Primitivesare represented by the following vector:

$$E = (X, \theta, \phi, (c_l, c_m, c_r), f) \tag{4.1}$$

With $X = (x, y)$ being the position of the Primitivein the image, $\theta \in [0, 2\pi]$ the orientation and $\phi \in [-\pi, \pi]$ the phase. Finally $f = (u, v)$ is the optic flow vector at this location.

4.3 A Multi-Modal Similarity Function

To address the problem of stereo correspondances, we need to define a metric to estimate the quality of a match between two local Primitives$E$ and E' (being Primitivesas defined in equation (4.1)). A similarity function involving measures of the distances in orientation $d_\theta(E, E')$, phase $d_\phi(E, E')$ and colour $d_c(E, E')$ of the Primitiveshas already been proposed in [**?**]. Here we extend this similarity function, including our optic flow distance $d_f(E, E')$.

For the optic flow information to be integrated in the stereo correspondances discrimination, a distance metric between any pair of optic flow vectors (f_1, f_2) has to be defined. The vectors may be dissimilar in length or orientation. We want a similarity function so that the vectors have a low similarity if their orientation is widely different. If the orientation is close, then the vector would have a higher similarity if their lengths are close. The dot product of the normalized two vectors is proposed as distance for vectors by [8]. If for a vector $f = (x, y)$ we consider the equivalent homogenous 3D vector $f_{3D} = (x, y, 1)$, then the normalized homogeneous vector is: $\hat{f} = (\hat{x}, \hat{y}, \hat{z}) = \frac{f_{3D}}{||f_{3D}||}$, so that $||\hat{f}|| = 1$. Then the dot product of the normalized 3D equivalent of two vectors gives a possible value for those two vectors similarity:

$$sim(f_1, f_2) = \hat{f}_1 \cdot \hat{f}_2 \tag{4.2}$$

This formula allows comparison for length as well as orientation: high difference in orientation (more than 45 degrees) yields a very low similarity whatever the length of the vectors, which is consistent with our perception of optic flow similarity. The use of normalized 3D vectors assure a consistent behaviour while comparing vectors of any size range.

The graph 4.2**a)** shows this function 4.2 for vectors of coordinates $f_1 = (x, y)$ with the vector $f_2 = (1, 0)$. This curve is effectively a representation of the similarity of two vectors. Similar vectors have a high value (up to one for identity), also vectors sharing a close orientation keep a higher similarity while the function value reduces sharply for vectors of widely divergent orientation

In order to get a distance funtion between the optic flow vector, and to improve the steepness of the curve close to the identity we apply the ArcCosine function to 4.2. Our distance becomes:

$$d(f_1, f_2) = ArcCos(\hat{f}_1 \cdot \hat{f}_2) \tag{4.3}$$

The high steepness of this function (cf. figure 4.2**b)**) allows us to identify the best match in a set of closely related vectors.

The resulting similarity function can be written as follows:

$$D_w(E, E') = w_\theta d_\theta(E, E') + w_\phi d_\phi(E, E') + w_c d_c(E, E') + w_f d_f(E, E') \tag{4.4}$$

with $w = (w_\theta, w_\phi, w_c, w_f)$ the weighting of the modalities distances between the two Primitivesso that $w_\theta, w_\phi, w_c, w_f \in [0, 1]$ and $w_\theta + w_\phi + w_c + w_f = 1$

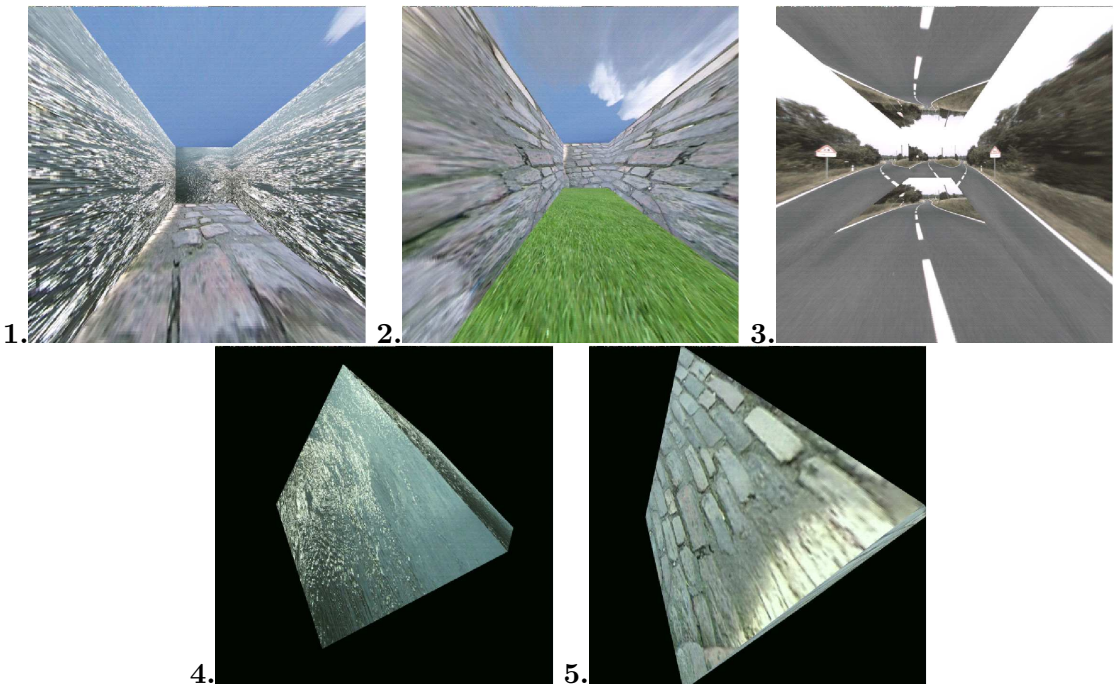


Figure 4.3: The five scenes used for the test.

All the modalities measured for those local Primitiveshave very different nature and distribution. As we want to combine them we need to normalize them somehow beforehand. We applied a normalization function proposed in [113].

4.4 Results

In this section, we investigate the relative importance of the modalities defining a Primitive(as in equation 4.1) for the task of stereo correspondances identification.
Concretely it means the quality of the stereo matching obtained using the similarity function defined in section 4.3, depending on the weights of each modality.

4.4.1 Data

We tested the quality of our stereo matching correspondances using artificial 3D scenes with natural textures (figure 4.3). The scenes feature a camera motion along a textured corridor, or rotating cubes, with varying textures.
On one hand, those scenes provide us an accurate ground truth for the scene depth (knowing the exact scene layout, camera projection matrices and motion), and so an

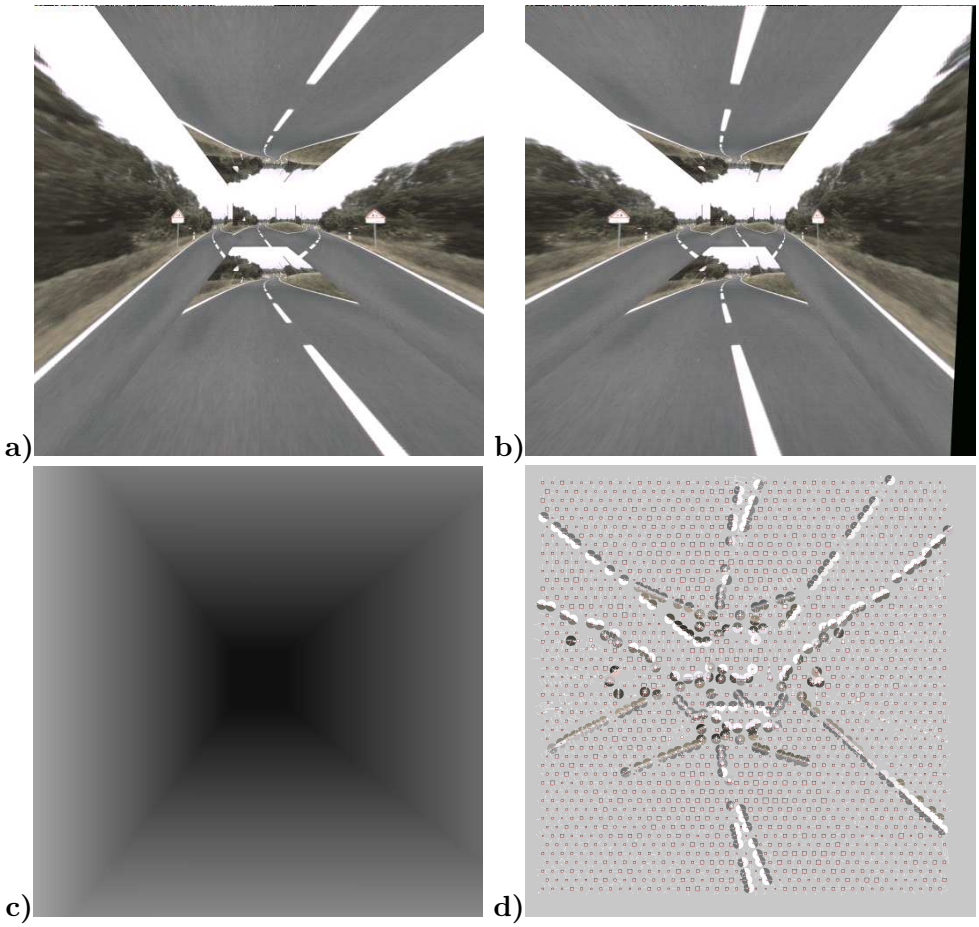


Figure 4.4: **a)** and **b)** are respectively the left and right images of the scene. **c)** shows a greymap of the disparity ground truth for this frame. **d)** shows the features extracted from this frame.

exact measure of the theoretical disparity can be computed. On the other hand, the projected textures ensure that we do work with natural structures.
By comparing the estimation of the disparity found with our method we can have a measure of the performance of the similarity function for this task. We consider sequences of 10 frames for each of those sequences, which comes to a total of 50 stereo frames of 512 per 512 pixels. Our statistics are made over a total of more than 66,000 matches.
In order to compare the relative importance of those modalities, we define a relative weighting $\alpha, \beta, \gamma \in [0, 1]$. α is the relative weight of the optic flow versus all the static modalities, β the weight of geometric information (the orientation measurement) versus

Sequence #	Chance	Cross-correlation	Multimodal	Weights of the peak performance
1	20%	26.0%	28.6%	$\alpha = 0.5, \beta = 0.4, \gamma = 0.2$
2	20%	45.0%	46.0%	$\alpha = 0.4, \beta = 0.4, \gamma = 0.3$
3	20%	56.2%	55.5%	$\alpha = 0.3, \beta = 0.2, \gamma = 0.2$
4	20%	68.2%	68.4%	$\alpha = 0.4, \beta = 0.2, \gamma = 0.3$
5	20%	65.3%	63.7%	$\alpha = 0.5, \beta = 0.4, \gamma = 0.2$
All	20%	52.1%	52.4%	$\alpha = 0.42, \beta = 0.32, \gamma = 0.24$

Table 4.1: Optimal parameters for each sequence, and comparison of performances.

structural information (phase and colour) and finally γ is the relative weight of phase versus colour.
We reformulate the distance (4.4) to use those relative parameters:

$$D'_{\alpha,\beta,\gamma}(E, E') = \alpha d_f(E, E') + (1 - \alpha)(\beta d_\theta(E, E') + (1 - \beta)(\gamma d_\phi(E, E') + 1 - \gamma d_c(E, E'))) \tag{4.5}$$

from (4.5) we define the similarity as follows:

$$Sim_{\alpha,\beta,\gamma}(E, E') = 1 - D'_{\alpha,\beta,\gamma}(E, E') \tag{4.6}$$

This formula is used to identify the best corresponding local Primitiveof the right image (maximizing (4.6)) along the epipolar line (cf. [29]). The subsequent disparity is then compared to the ground truth for the disparity of the sequence. The quality of the similarity function is then evaluated simply by the ratio of correct correspondances over all matches.

4.4.2 Performances using all Modalities

To have a performance baseline to estimate the quality of our correspondances, we calculated the chance performance (the performance using a random similarity function) and a cross-correlation over 10x10 patches (here the similarity function used is the cross correlation of the patches). Those have been calculated for our five benchmark sequences. The matching performance of our similarity function for values of $\alpha, \beta, \gamma \in [0, 1]$ is shown in figure 4.5. We can see a plateau for α close to 1. In this case, only the optic flow modality is being used, so the variations in β and γ do not affect the surface. Also, as β is close to 1, the γ parameter does not affect the curve, then reduced to a 2-dimensional curve function of α. This is consistent with formula (4.5) where the higher the value of α, the lower the impact of the two other parameters, and the higher the value of β the lower the impact of γ).
In average over all sequences the peak performance is reached for $\alpha = 0.42, \beta = 0.32, \gamma = 0.24$. The results for specific sequences are shown in table 4.1. We can see that the optimal weighting is very consistent over the different sequences, even when the quality of the disparity changes drastically. The peak performance is reached for a strong use of

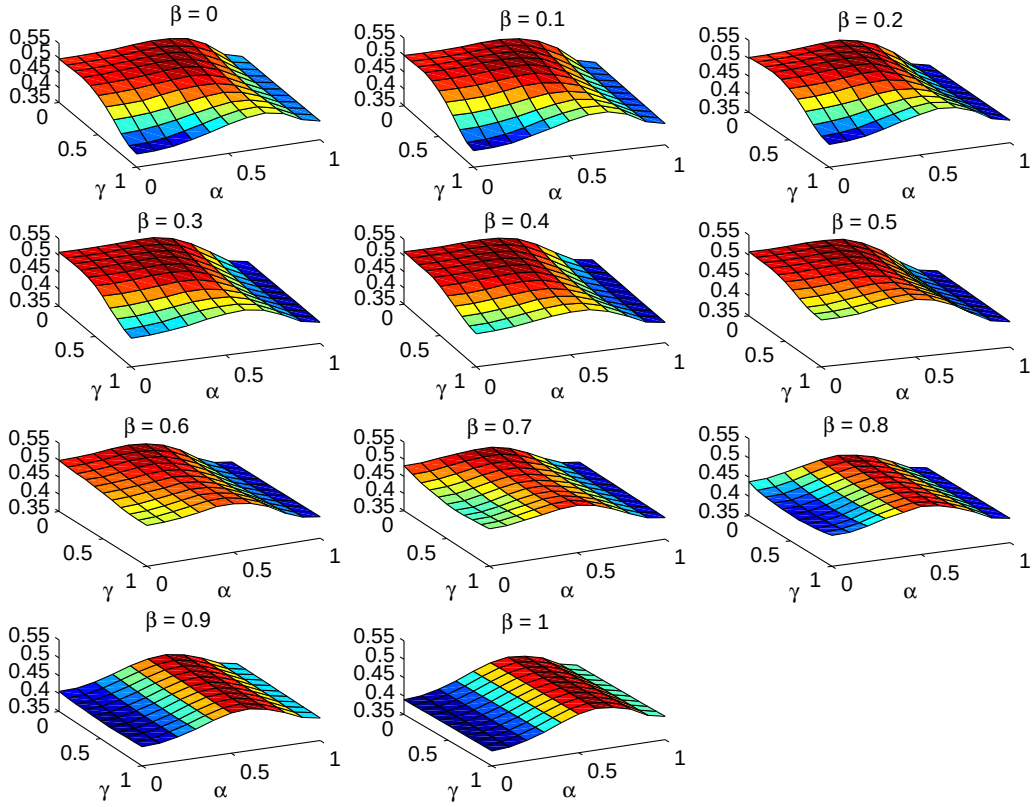


Figure 4.5: Graph of the disparity quality for different modality weights over all sequences. The different graphs are for different values of β, the α values are along the x axis and the γ along the y.

the optic flow information ($\alpha \simeq 0.4$), showing the relevance of the optic flow modality for this task. Also, the algorithm performs slightly better than the cross correlation while *using only ten parameters instead of 300.*

4.4.3 Performance without Colour or Optic Flow

The performances with grey level images (γ is then set to 1) is shown in figure 4.6 and table 4.2, third column. Again the peak performance is reached for a significant use of optic flow. The peak performance drops by 2.7% compared to colour images and again by 2.5% if the optic flow is neglected ($\alpha = 0$). As expected, on figure 4.6 the performance decreases considerably when using only one of the modalities. This shows the relevance of this multimodal matching, and more specifically of the use of optic flow for this task.

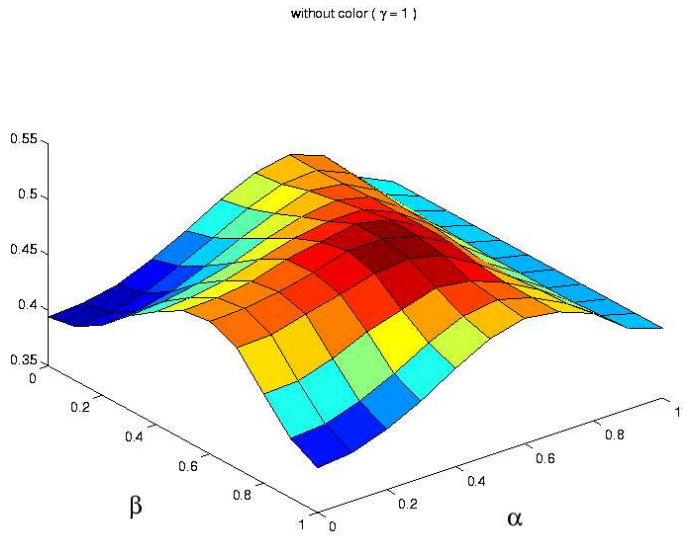


Figure 4.6: Graph of the disparity quality for different modality weights over all sequences, excluding the colour ($\gamma = 1$).

Sequence #	Chance	Multimodal without Colour	Multimodal without Optic Flow
1	20%	25.8%	26.7%
2	20%	43.3%	45.5%
3	20%	53.1%	54.6%
4	20%	64.5%	67.9%
5	20%	62.0%	63.6%
All	20%	49.7%	51.7%

Table 4.2: Performances of our function when excluding one parameter (colour or optic flow).

In table 4.2, fourth column, is shown the performance of the program on colour images without using the optic flow information (*i.e.* with the parameter α set to 0). This respresents a drop in peak performance of 0.7%. Compared to the 2.5% with greyscale images, this leads us to assume that the use of different modalities improves the robustness as well as the general performance of the method. The marginal loss of performance when ignoring one of the most weighted modalities (2.7 percents for colour, and 0.7% for the optic flow, compared to the 5.2% of loss when neglecting those two), also confirms the robustness of this multimodal similarity function.

4.5 Conclusion

In this paper we presented a multimodal similarity function and applied it to the stereo correspondance problem. We applied this method to several scenes of diverse difficulty and compared its performance with a standard normalized cross-correlation algorithm. The results clearly shown the importance of the optic flow in this method. It is also interesting to note that our data processing allows an important data reduction: this representation features *only ten parameters* (or 4 without the colour information) *instead of 300 for the cross correlation (100 without colour)*, which comes to a reduction of 96.6% (96% without the colour). In spite of this considerable condensation, we assume that no crucial information loss (relatively to the task) had happened, as the result matches the performances of the cross correlation, and even outperform it slightly on difficult scenes (emphasizing again the importance of the added optic flow information). The robustness of the method is outlined by the consistency of the optimal weights found over all sequences, while the peak performance itself varied largely..

4.6 Orientation in the Plane and Switching

The orientation θ is defined in $[0, \pi[$. Consequently if we consider the direction $d \in [0, 2\pi[$, we face the ambiguity between two interpretations: $d = \theta$ and $d' = \theta + \pi$. Those two interpretations are realistic and equivalent, orientation-wise, as no direction can be defined locally for an edge. Also the color and phase information encoded in our primitives are relative to this orientation. This leads to ambiguity in three specific cases: First when comparing two primitives in their modalities, secondly when generating a new primitive, and finally when creating 3D primitives from a stereo pair of 2D primitives.

When comparing two primitives, we face the problem that the difference between the orientations is between $[0, \pi[$. This does not make sense as geometrically two orientations cannot be more different that orthogonality, so a difference of $\frac{\pi}{2}$. Consequently it means that if $\Delta\theta > \frac{\pi}{2}$, a more accurate comparison between the primitives is achieved by condidering θ_1 and $\theta_2 + \pi$. Furthermore, the meaning our definition of the color and of the phase is orientation dependant. This also means that the phase and color information should be corrected accordingly to the chosen interpretation, for the comparison to be accurate. This means comparing left (respectively right) color with right color (left), and comparing p_1 with $-p_2$. We call this correction *switching*.

If a primitive is generated, if its direction is outside $[0, \pi[$ then the primitive needs to be permanently switched (as specified before) for the resulting orientation to be in this interval. Consequently if we consider a relation $T : [0, \pi[\times[0, \pi[\rightarrow [0, \pi[$ which estimates an orientation out of two prior orientations (typically the case of the correction of a primitive using a predictor), then this simmetry problem is two-fold: First, can we compare the two orientations, or do we need to operate a switching on one ? And secondly, do the

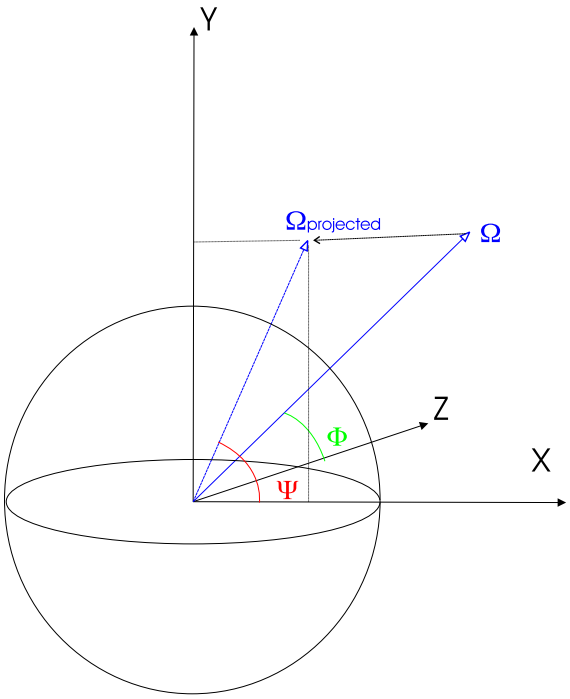


Figure 4.7: Here $\phi \in [0, \frac{\pi}{2}]$, and $\psi \in [0, 2\pi[$, so to define the half-sphere of vectors with positive z values.

resulting orientation require itself a switching ?

4.7 from stereo 2D primitives to 3D primitive

If we have a pair of calibrated images of one scene, we know that from a pair of corresponding point, one in each image, we can reconstruct a 3D position in the scene. Also, from a pair of planar orientations, we can reconstruct a 3D orientation. We want to extend this so that from two 2D Primitives$\{\mathbf{e}^L, \mathbf{e}^R\}$ we reconstruct a 2D entity $\mathbf{E}$. The reconstruction is as follows:

$$R(\mathbf{e}^L, \mathbf{e}^R) \rightarrow \mathbf{E} \tag{4.7}$$

of course we want the inverse opperation to be possible

$$P_L(\mathbf{E}) \rightarrow \mathbf{e}^L \tag{4.8}$$
$$P_R(\mathbf{E}) \rightarrow \mathbf{e}^R \tag{4.9}$$

Our local 2-dimensional Primitives were defined as follows:

$$\mathbf{e} = \{X, \theta, \phi, C, f\}$$

for respectively the 2D position $X = (x, y)$, the orientation $\theta \in [0, \pi[$, the phase $\phi \in [-\pi, \pi]$, the color $C = (c_l, c_m, c_r)$ and the optic flow $optiFlowVar = (u, v)$.
The 3D entity is the reconstruction of the 2D entities in the 3D space. Consequently it is bound to the surface that generated it. We consider that a 3D entity is defined in a plane tangent to the generative surface at this point. A 3D entity has then to define unambiguously:

- the position
- the tangent plane: which can be defined by two vectors and the position
- the orientation on this plane
- the other modalities

the 3D orientation gives us one of the vector required to define the plane, so we just need to define a second one. We define an additional vector Γ, which belongs to to the tangent plane defined by the primitive, and towards the side designated as the 'left' color (cf. figure 4.8). χ, θ and Γ define a plane tangent to the surface the 3D-primitive belongs to. If we consider the two projected points on the stereo images P_l, P_r. From those points and the projection matrice we calculate the projection vectors v_l, v_r. We define the observer direction vector as follows: $o = \frac{v_l + v_r}{2}$. Then $\Gamma = \theta \times o$
We define the 3D entities accordingly:

$$\mathbf{E} = \{\chi, \theta, \Gamma, \phi, C, F\}$$

As we defined the 2D orientation between 0 and π to remove ambiguity, we will define the orientation in the half sphere towards z positives. So, $\chi = (x, y, z)$ the position, $\theta = (\phi, \psi) \in [0, \frac{\pi}{2}] \times [0, 2\pi[$ the 3D orientation (cf. figure 4.7). Γ is the vector defining the plane, ϕ and C are still the definition of the phase and color of the entity. As a 3D entity is defined relatively at a plane tangent to the surface at this location, their meaning is identiqual to the 2D one. Finally F is the three-dimentional flow, which we will neglect in the following.
The phase and color information is then an averaged of the values of the pair of 2-dimensional primitives. We face here the first switching issue discussed in 4.6: for the stereo pair to be comparable a switching may be required. Then also, as the 3D-orientation is defined in the demi-sphere of positive z, if the resulting 3D vector is of negative z an additional switching is required.

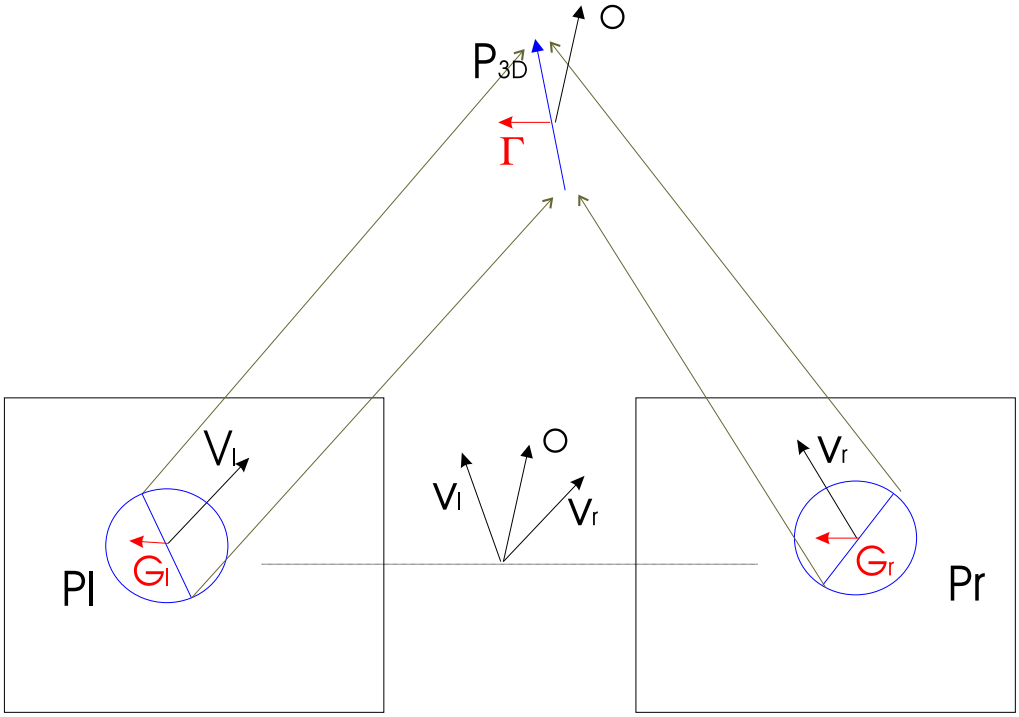


Figure 4.8: The vector Γ is calculated from the projective lines v_l and v_r and the 3D orientation θ and define the 'left' color in the 3D domain. Then when reprojecting, the vectors G_l, G_r give the reprojected left color.

4.8 Reprojection: from 3D Entities to Pseudo-Primitives

We can reproject the 3D points onto the 2D images using the projection matrices of both images. This do not cause problem for position and orientation, yet for the other modalities the case become slightly more complex. We also reproject the vector Γ onto the image plane, to check if a switching of the color and phase is required.
Finally, when reprojecting the 3D-entity onto a 2D plane, we can get unambiguous 2D pseudo-Primitives: if the reprojected orientation is outside $[0, \pi[$ then a switching occurs. Then if the projection of the vector Γ points towards the area defined as right in the primitive, another switching is required.

Chapter 5

Formalisation, Estimation and Application of Rigid Body Motion

The knowledge of ego motion and motion of other objects is an important regularity that allows for predictions across frames which can be used to disambiguate visual information. The formalisation and computation of motion has received the attention of a significant number of scientists (see, e.g., [63, 34, 29, 28, 107]). As we will see, it is the correspondence problem that is crucial in this context and that the combined utilisation of the deterministic regularity RBM and statistical regularities in grouping processes can help significantly to deal with it.

5.0.1 The projective Map

By watching a scene with a camera the 3D world is projected onto a 2D chip. This can be described (in a simplified camera model[1]) by the equation

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{X}{Z} \\ \frac{Y}{Z} \end{pmatrix} \tag{5.1}$$

where (x, y) are the image coordinates and (X, Y, Z) are the 3D–coordinates. The Z-dimension is lost, leading to a coniderable degree of ambiguity in scene analysis. However, having two cameras that look at the scene from different viewpoints we can reconstruct

[1]Note that for a real camera we have to find a set of parameters that describe the mapping between world coordinates and pixel coordinates. The RBM between the camera and the world coordinate system is one sub–set of parameters (external parameters) to be found. Internal parameters (i.e., the co–ordinates describing the position and angle of the chip in the camera, the size of the chip, the number of pixels as well as the focal length) have to be computed as well. This estimation process is called calibration and is known to be sometimes quite awkward (see, e.g., [29, 61])

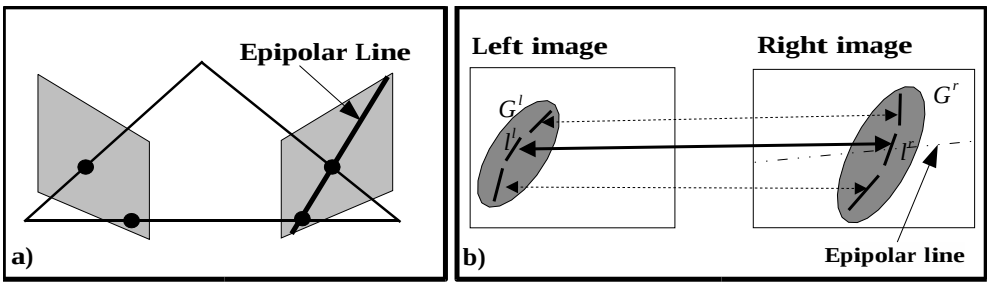


Figure 5.1: a) Epipolar Line Constraint. b) Predictions in the stereo domain based on grouping: Assuming the correspondence indicated by the solid line the correpondences indicated by the broken lines can be predicted.

the third dimension. Note that different kind of correspondences lead to different types of reconstruction. For example, two point correspondences lead to a 3D point. Two line correspondences lead to a 3D line (see, e.g., [29]), and the correspondence of two points with associated orientation lead to a 3D point with associated 3D orientation (see, e.g., [73]).

5.0.2 The Correspondence Problem in Stereo

Reconstruction presupposes a correspondence of visual entities in the left and right image. Although for humans this seems easily solvable, it is a serious problem in computer vision systems. What makes it so difficult?

- Different perspectives in the left and right image lead to differences in the projection. For example, the orientation of the projected edge is in general different in the left and the right image. Indeed, it is this difference which on the one hand makes the correspondence problem difficult and, on the other hand, makes the reconstruction possible. Furthermore, the colours of surfaces in the left and right image are different, since they depend on the viewing angle. Moreover, it may be that, because of occlusion, we see a different physical surface in the left and right image.
- There may occur repeating structures in a scene. These structures can not be distinguished by pure local matching.
- Many image areas are homogeneous or weakly structured. Thus, there is no chance to find correspondences by local comparisons since these would all give high similarities. In this case we need to apply indirect and more global methods.

However, there exist a number of constraints that reduce the correspondence problem.

- Uniqueness: An image entity in the left image can have at most one correspondence in the right image. Note, that it is possible to have zero correspondences in case of occlusion.
- Epipolar Line Constraint: The corresponding point in the left image must fall onto the so called epipolar line. The epipolar line is the intersection of the right image with the epipolar plane (see figure 5.1 and [29]). The epilolar plane is generated by the line spanned by the optical centre of the left camera, the image point and the optical centre of the right camera[2] (see figure 5.1). In this way, we can reduce the correspondence problem to a one–dimensional search problem.
- It has been shown that the use of multiple modalities enhances stereo performance (see, e.g., [65, 73]). In our system, we have utilized the modalities orientation, phase, colour and optic flow to improve stereo matching [73, 95].
- There exist further spatial constraints [29, 61]. Assuming certain assumptions about the 3D scene are made, constraints on the relative displacement of features in the left and right image can be made.
  - Ordering: *The order of points on the epipolar line is the same in the left and right image.* This contraint is valid if the objects in the scene have similar distance to the camera. This constraint is, for example, used in dynamic programming approaches (see, e.g., [19, 39]).
  - Limit of Disparity: *Difference in the position of corresponding points in the left and right image does not exceed a certain disparity value.* This contraint is fullfilled when objects have a minimal distance from the camera.
- Grouping can significantly enhance stereo matching (see, e.g., [18]). In figure 5.1b, a possible application of grouping in stereo processing is described: Assume a local line segment l^l in the left image is part of a group G^l. Furthermore, assume that this line segment has a correspondence l^r in the right image which in a similar way is part of the group G^r, then all local entities of G^l must have a correspondence in one of the local entities of G^r.

5.1 The RBM Estimation Problem

Different kind of motion patterns exist in visual scenes. For example, the motion of a bird is a complex combination of its limb movements and the movement of its elastic

[2]The same holds also from right to left.

skin and feather structure that depends on the ego–motion and on other factors such as wind and temperature. A motion with similar complexity is the motion of humans. Human motion is also a commercially interesting problem, since it leads to applications in, e.g., video surveillance. It has been addressed by many scientists (see, e.g., [15]). However, there are other motion patterns that are much simpler than that of a bird or a human. One important class of motion is pure ego–motion, that occurs, e.g., in a video taken from a car on an empty highway or in a movie of a still life taken from a moving camera. The mathematical structure of this kind of motion has been studied for a long while (see, e.g., [5, 60]) and will be described in detail below. This structure, often called 'Rigid Body Motion' (RBM)[3], can be described as a six–dimensional manifold consisting of a translation (parametrised by the three coefficients $\mathbf{t} = (t_1, t_2, t_3)$) and a rotation (parametrised by $\mathbf{r} = (r_1, r_2, r_3)$). In figure 5.2a such a parametrisation is displayed. First we perform a rotation $Rot(\mathbf{p})$ around the axis $\mathbf{r}$. The norm of this axis codes the angle of rotation $\alpha = ||\mathbf{r}||$. Then we move a point according to the translation vector $\mathbf{t}$.[4] Note that in many scenes, not only one (ego–)motion exists but in addition other rigid objects (other cars and lorries) move. Their motion is also describable by an independent rigid body motion.
An RBM describes the transformation of a 3D entity[5] $\mathbf{e}$ in the first frame to a 3D entity $\mathbf{e}'$ in the second frame[6]

$$RBM^{(\mathbf{t},\mathbf{r})}(\mathbf{e}) = \mathbf{e}'. \tag{5.2}$$

To apply equation (5.2) we need to define correspondences between visual entities $\mathbf{e}$ and $\mathbf{e}'$.[7] Each of these correspondences defines one or more constraint equations. If the RBM is applied to the entity $\mathbf{e}$ it must match $\mathbf{e}'$. Therefore, it must hold

$$||RBM^{(\mathbf{t},\mathbf{r})}(\mathbf{e}) - \mathbf{e}'|| = 0. \tag{5.3}$$

Note that the norm $||\ ||$ can vary. This especially holds for different choices of entities $\mathbf{e}$. We discuss this issue in section 5.6.4. If we have a set of constraints (based on a set of

[3]We define Rigid Body Motion of an object as a continuous movement of the object, such that the distance between any two particles of the object remains fixed at all times.

[4]There exist other ways to formalize an RBM, e.g., by Euler angles or dual quaternions (see section 5.6.2). However, it is always a six–dimensional manifold that describes the RBM

[5]In the following 3D entities are printed in boldface while 2D entities are printed normal.

[6]For the sake of simplicity we also use the notation $RBM(\mathbf{e}) = \mathbf{e}'$ if the context is clear.

[7]There exist methods that avoid an explicit coding of features or entities. In these methods, the rigid body motion problem is formulated not on derived features but on the pure image data. As a consequence, the formulation in equation (5.2) would appear only implicitly in these methods (see, e.g., [16, 120, 88, 49]). In our approach, we do not follow this implicit approach. However, we will discuss the implications of the different methods in section 5.2.1.

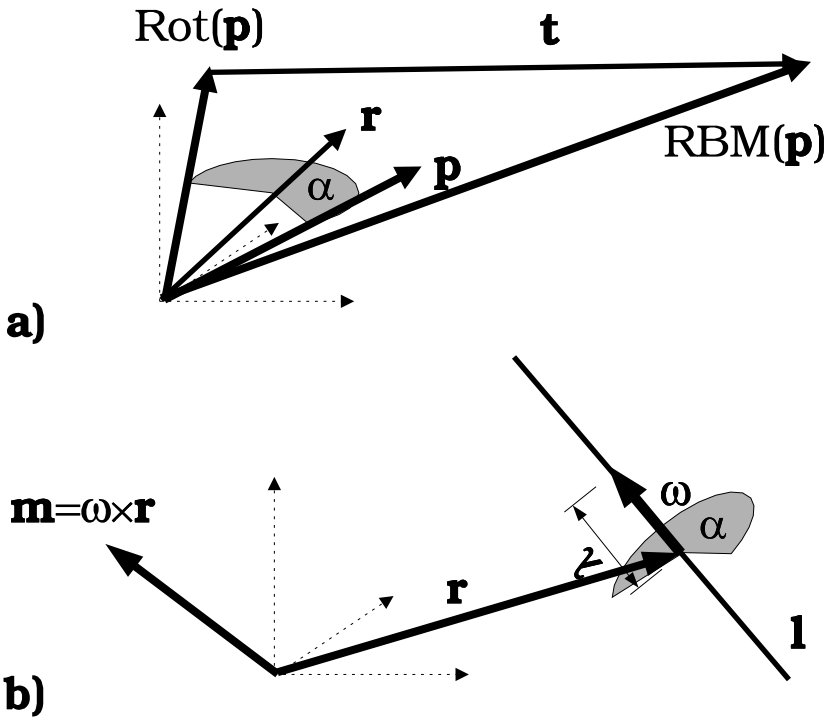


Figure 5.2: Two Representations of a Rigid Body Motion. a) Combination of rotation and translation. b) Twist representation: A rotation around a line $\mathbf{l}$ in the 3D Space with direction $\mathbf{w}$ and moment $\mathbf{m}$ and a translation along $\mathbf{w}$ with magnitude λ is performed.

correspondences) we get a system of equations that allows for computing the RBM, i.e., the underlying parameters $\mathbf{t}, \mathbf{r}$.
Up to this point the motion estimation problem may appear to be quite simple. However, there are significant problems involved that will be discussed now:

- **Dimensionality of Entities:** There occur different situations of different complexities in which RBM estimation can be performed (see section 5.2.2). For example, since in vision, a camera records a scene on a 2D chip, we only record a motion in 2D and we have to deal with 2D features extracted from images.[8] Therefore, we may not want to directly apply equation (5.3) but instead may want to embed this

[8]Note, that there exist sensors that record 3D information directly such as range finders [98]. However, they are very different from standard cameras and have specific disadvantages such as high costs and limited resolution and depth range. Furthermore, such approaches are rarely realized in biological systems.

equation in some kind of 2D context. On the other hand, in a stereo scenario, we have the possibility to extract 3D features (see section 4). However there is a high degree of ambiguity in these features which we would probably like to eliminate before addressing the rigid body problem.

- **Semantic of visual entites:** Apart from the dimensionality of the entities used for RBM estimation (see section 5.3.1), we can apply entities of different semantic (see section 5.3.2): In equation (5.2) we can bring points to a correspondence. However, one could also think of correspondences of line segments or entities of even higher complexity such as curves or circles. Therefore, we want to formulate the RBM estimation problem for different kind of visual entities.

- **Mixing of visual entities:** Through grouping, complex, extened entities can be formed by combining local entities (see figure 5.4). These groups can include different kind of entities. For example point–like or line–segment–like entities. When we want to apply such groups for RBM estimation it is advantageous to have the ability to *mix* such correspondences.

- **Correpondence problem:** For RBM estimation, we have a correspondence problem (discussed in section 5.4) that is even more serious than the correspondence problem in the stereo case (see section 5.0.2) since the epipolar constraint is not directly applicable[9]. The correspondence problem becomes even more severe in scenes with multiple independent motions. In section 5.4, we will discuss the power or value of different kind of correspondences as well as different constraints that make the correspondence problem manageable. We will see that grouping can be an important constraint that has only seldomly been used in artificial visual systems.

- **RBM representation:** There are some problems that are deeply connected to the mathematical representation of Rigid Body Motion which are discussed in section 5.6. For example,

  - the solution of equation (5.3) needs to be computed by some kind of numerical optimisation method. Different choices of numerical method may lead to different kind of solutions (see section 5.6.1).
  - the algebraic embedding of RBM may lead to systems of equations with more unknowns than necessary. For example, the standard matrix formulations work on 12 unknowns, but only 6 are needed to code an RBM. As a consequence, such approaches search in the wrong and far too large space. This leads to solutions that are no RBM anymore (see section 5.6.2).

[9]However, the epipolar line constraint can be used implicitely (see [110])

- the way we represent mathematical entities such as points and lines (see section 5.6.3) influences the formulation of our constraint equations (5.3). Their definition is not trivial, since a proper formulation of distance between such entities has to be found.
- it would be advantageous to have a geometric interpretation for the constraint equation (5.3) to ensure stability of computation. This will be discussed in section 5.6.4.

Moreover, we will see that all the above mentioned problems are deeply intertwined.

Having described basic problems of RBM estimation in section 5.2, 5.3, and 5.4, we will derive four desired requirements of RBM estimation algorithms for real world applications in section 5.5: accuracy, reliability, flexibility and minimality. We will show that grouping can be a crucial aspect in RBM estimation that is involved in all four requirements.
In the following, we will discuss the RBM estimation problem in a way that we hope is understandable for a broad range of scientists with different background. However, RBM estimation is also a mathematical problem and therefore math can not be completely avoided. However, the discussion of mathematical problems is concentrated in section 5.6 and can be skipped in a first reading.
Within this review, we will outline an RBM estimation algorithm to some mathematical detail that has been developed by our colleagues Bodo Rosenhahn, Oliver Granert and Gerald Sommer [103, 102, 104, 40, 101]. This has three reasons: First, this specific RBM estimation algorithm has certain unique advantages that will become obvious in the following discussion. Secondly, we use this algorithm in our attempt to implement artificial visual systems (see, e.g., [75]). Finally, we will use this pose estimation algorithm to exemplify general problems of RBM estimation that can be easier understood by looking at a specific mathematical formulation.

5.2 Classification of Methods and Situations

5.2.1 Different types of Methods

In RBM estimation entities used to define correspondences can be represented explicitely as features (as done in equation 5.2) or implicitly. There has been a long debate about this issue. According to the degree of expliciteness different methods can be separated into feature based, optic flow based and direct methods (see [110]).

- **Feature based methods:** In feature based methods [99, 82], at first features (e.g., junctions [94] or lines [75]) are extracted. Once these features are found, correspondences between features are defined and used in the constraint equations. These methods have to deal with the problem of feature extraction. The ambiguity

of visual data leads to erroneous or missing features. For example, it may be that the local interpretation is 'wrong'. There may exist a weak line structure in the first frame (slightly above threshold) but the corresponding structure in the second frame is below threshold (or dominated by noise). Then there is no chance to find a correspondence since the corresponding entity simply does not exist in the second image. Therefore, special mechanisms to deal with these cases need to be considered. One possibility to deal with this dilemma is to make use of confidences associated to features (see, e.g., [75, 21, 72]).

- In **optic flow methods** (see, e.g., [16, 49]) the optic flow with all its inherent ambiguities is used. A nice property of optic flow methods is that these methods may acquire a good solution by implicitly averaging over the ambiguous data. However, since this kind of correction process is implicit, one does have only little control about the influence of specific outliers.

- In **direct methods** no explicit representations as features or optic flow vectors are used but image intensities are directly matched [120, 88, 23]. The advantage of these methods is that all problems connected with feature extraction can be avoided. However, the drawback is that the ambiguity of local interpretations is also implicitly existent in the intensity patches.

In our system, we do feature based pose estimation. However, we are aware of the difficulties connected with such approaches.

5.2.2 Different Types of Situations

The RBM estimation problem occurs in different situations.

- **Single image:** Alignment of an existing 3D model of an object within a 2D image is a complex task since no constraints concerning the RBM can be made. This problem occurs in case of object alignment in 2D images (see, e.g., [83, 103]). In the constraint equations we therefore need correspondences between 3D object and 2D image equations (see figure 5.3b).[10]

- **Stereo:** In case of recording the scene with a stereo system we have two images that record the same RBM. Therefore, having an image entity in the left frame and a corresponding entity in the right frame $Cor(e^l, e^r) = 1$, both describe the same RBM and lead to one additional constraint equation[11]:

[10] This is also the standard problem that has to be solved in camera calibration with known calibration body.

[11] P^l or P^r is the projective map of the left or right camera respectively

$$\Big((P^l(RBM(\mathbf{e})) = e^l) \wedge (Cor(e^l, e^r))\Big) \Rightarrow (P^r(RBM(\mathbf{e})) = e^r).$$

Furthermore, we can use stereo to extract 3D information and then apply 3D-2D pose estimation even if we have no prior object knowledge (see, e.g., [75]). As a consequence, we can use correspondences between 3D object and 2D entities in our constraint equations.

- **Image sequences:** When we record a scene with a (stereo)–camera system continously we have different frames that are connected by the camera's RBM and the motions of the objects within the scene. At normally used frame rates, it is very unlikely that corresponding image coordinates have large distance in consecutive frames. This *continuity constraint* reduced the correspondence problem considerably and leads to more stable motion estimates.

5.3 Using Different kinds of Entities

In our constraint equations, we need correspondences between visual entities. These entities can have different spatial dimension (see section 5.3.1) as well as different semantic (see section 5.3.2). We will see that in the context of grouping both aspects are relevant.

5.3.1 Entities of different Dimension

Following [48], we distinguish 3 cases of RBM estimation problems that differ depending on the spatial dimension of visual entities. First, we can compute the RBM from 3D–3D correspondences (see figure 5.3a). Second, we can have a model of an object that inherits 3D aspects, either by manual design (see, e.g., [83, 103]) or by some kind of acquisition mechanism that has taken place beforehand (see, e.g., [70]). In this case, 3D aspects of the object can be brought into correspondence with 2D aspects of its projection (see figure 5.3b). Thirdly, we can deal with 2D projections only (see figure 5.3c).

3D–3D Correspondences: We can extract 3D information by stereo or by a sensor that works directly in the 3D domain (e.g., range finders [98]). Then we can define correspondences in 3D and our constraint equations have the simple form

$$RBM(\mathbf{e}) = \mathbf{e}'. \tag{5.4}$$

From a mathematical point of view, this is the easiest case since we can avoid any problems resulting from the perspective projection (see section 5.0.1).
However, working with 3D entities inherits other problems. For example, in case of extracting 3D information by stereo, we have to deal with its ambiguity since wrong

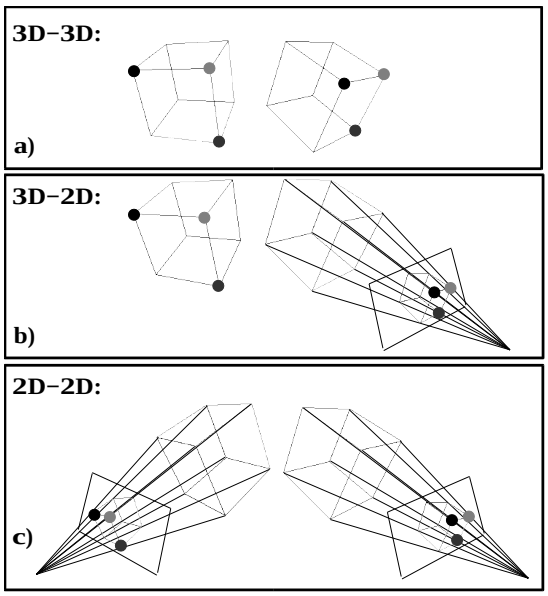


Figure 5.3: RBM–Estimation from different Correspondences. a) RBM estimation from 3D correspondences (displayed as circles). b) RBM estimation having a 3D model and 2D correspondences in an image. c) RBM estimation having 2D image coordinates in one image and its 2D correspondences in a second image.

correspondences will lead to significant distortions in the RBM estimation. In case of laser range finders, we have to deal with a type of sensor that has specific problems such as the necessity for expensive and time consuming scanning and a limited depth range. Furthermore, the determination of 3D–3D correspondences is not trivial.

RBM from 3D-2D Correspondences: A camera projects a scene to a 2D chip. Therefore, it is convenient to use entities that are extracted from a 2D image only. However, there occur many applications in which prior object knowledge does exist. For example in industrial robot applications CAD descriptions of objects may be available (see, e.g., [26]). This leads to the problem of estimation the RBM from entities of different dimensions: The 3D object knowledge needs to be aligned with 2D entities in an image of this object. The problem of computing the RBM from correspondences between 3D object and 2D image entities is commonly referred to as 3D–2D pose estimation problem [42, 101].[12] In mathematical terms we have the following kind of constraint equations:

[12]When combined with ego–motion or object–motion we can apply this approach in an iterative scheme leading to a particulary succesful approach based on the so called analysis–by–synthesis paradigm (see, [63, 25]).

$$P(RBM(\mathbf{e})) = e',$$

where P represents the perspective projection.
There exist different ways to approach the 3D-2D pose estimation problem. They differ in the way they deal with the perspective projection. The perspective projection makes the 3D–2D pose estimation problem mathematically more demanding than the 3D–3D case since the perspective projection introduces a non–linear and non–invertible function. However, one can try to deal with this problem by simplifying the projected 3D motion or by a simplified camera model. Furthermore, there are approaches that reproject 2D entities in the 3D space.
In the following we will discuss the different alternatives in more detail.

- **Orthographic formulation:** For objects with a large distance from or with similar depth to the camera, the projective map can be approximated by the so called orthographic projection

  $$O : (x, y, z) \rightarrow (x, y).$$

  This leads to the constraint equation

  $$O(RBM(\mathbf{e})) = e'.$$

  As the perspective projection, the orthographic map is not invertible, but it is much simpler. Some authors (see, e.g., [15, 115]) formulate the pose estimation problem by making use of the orthographic map.[13]

- **Simplified formulation in image coordinates:** In Lowe's pioneering work [83] an error function measures the deviation of image points $P(RBM(\mathbf{e}))$ and points e' in an iterative manner. However, the transformation of image coordinates is simplified by an affine approximation.

- **Fully projective formulation in image coordinates:** Both approaches mentioned above have the serious drawback that their approximations are not necessarily exact. Therefore, it is advantageous to deal with the full perspective projection. This has been done by [2], who generalise Lowe's algorithm [83] to a fully perspective formulation.

- **Formulation in 3D Space:** Instead of formalising the pose estimation problem in the image plane, we can associate a 3D entity to each 2D entity: For example a 2D image point together with the optical center of the camera spans a 3D line (see

[13] Note that Bregler and Malik [15] use some kind of scaling to minimise the effect of approximating of the projective function with the orthographic map.

figure 5.5b) and an image line together with the optical center generates a 3D plane (see figure 5.5c). We denote the 3D entity that is generated in this way from a 2D entity e' by $\mathbf{e}^{P^{-1}(e')}$. Now the RBM can be applied to 3D entities

$$RBM^{(\mathbf{t},\mathbf{r})}(\mathbf{e}) = \mathbf{e}^{P^{-1}(e')}.$$

The Euclidian formulation has been applied by, e.g., [93, 40, 103]. This formulation is elegant, since it deals with the full perspective projection. It works in the space where the RBM takes place (i.e., the Euclidian space) and also allows for nicely interpretable constraint equations. However, one problem of this formulation is that the constraints are defined in 3D. This approach inherits problems since error measurements of 3D entities depend on the depth: The estimation of feature attributes of entities with large depth has a higher uncertainty than that of entities at a close distance. Thus, correspondences of entities with large distance would have higher influence in the constraint equations (see [78]).

Structure from Motion using 2D-2D Correspondences: In the structure from motion problem only 2D entities occur and the problem reads:

$$P(RBM^{(\mathbf{t},\mathbf{r})}(\mathbf{e}^{P^{-1}(e)})) = e'$$

A considerable amount of literature is concerned with this problem (see, e.g., [44]) and reconstruction of complex 3D–scenes can be performed by this approach (see, e.g., [107, 63, 94]). However, 3D information can only be computed up to a scaling factor since a small object with close distance and low speed would lead to the same pattern than a big object that is identical except its size with high speed. In the following, we will mainly concentrate on the first two cases, i.e., RBM estimation from 3D–3D and 3D–2D correspondences. However, we want to point out that RBM is also the underlying regularity in structure from motion algorithms. For overviews about structure from motion algorithms we refer to [117, 44].

5.3.2 Entities of different Complexity

Visual Entities can not only be characterised by their spatial dimension but also by other attributes such as, e.g., orientation or curvature. This has been also reflected in the RBM estimation literature: There exist a large number of RBM estimation algorithms for points (see, e.g., [42, 93, 83]) and lines (see, e.g., [48, 109]) and also for higher entities such as circle-like structures (see, e.g., [62, 101]).
At this point we face a general problem. What are the entities we want to use for pose estimation? We must be careful not to make assumptions that are motivated by the

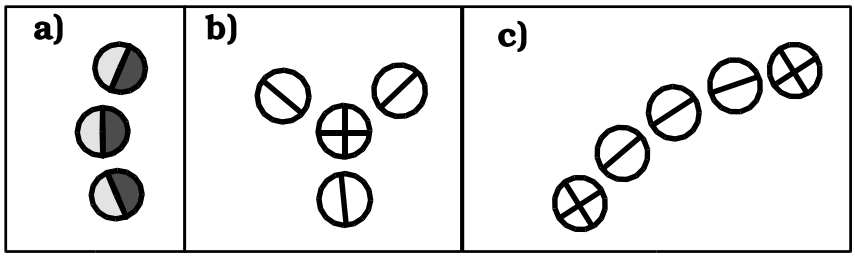


Figure 5.4: Examples of groups: a) Constellation of collinear line segments. b) A junction as a combination of an intrinsically two–dimensional and 3 intrinsically one–dimensional primitive. c) A collinear group with two defined endpoints.

mathematical framework we use but may not be in accordance with our problem. Since geometry usually deals with points and lines these entities are not necessarily good visual entities. For example, each point–feature in an image (such as a junction) has additional attributes: in case of a junction there are oriented edges that are directed towards that point and most line–like features have some kind of start and end point, i.e., are not of infinite length such as mathematical lines are. Therefore, *there are no ideal points and lines in images.*
In this work we suggest to use *groups of multi-modal local entities as basic entities for RBM estimation.* Groups can be interpreted as 'Gestalts' generated by specific joint properties. For example, by similar colour or collinear orientation. Figure 5.4 shows some examples of possible groups. A particular property of groups is

- that they consist of local entities of possibly different type (for example a line with its end points or a junction point with its lines intersecting), and
- that they can not pre-defined but self-emerge dynamically depending on the actual scene (see, e.g., [119]).

An RBM estimation algorithm that uses the power of grouping must have the property to use different kinds of visual entities since groups may consist of entities of different structure. However, mixing entities within one system of equation is not easy from a mathematical point of view since the RBM may have different formalisations for different entities. For example, the RBM of a point can be described straightforwardly by a matrix [29] while dual quaternions are also suited to describe the RBM of a line (see, e.g., [109] and 5.6.2). It is an important step forward to be able to mix these kind of correspondences and it has been shown that this can be done by e.g., [40, 101]. A specific

algebraic formulation in 'conformal algebra' (see, e.g., [45]) that allows for dealing with different kind of entities at the same time was helpful to derive such a formulation.

5.4 The Correspondence Problem

When we want to estimate the RBM, we face a correspondence problem that is even more serious than in the stereo case. The correspondence problem for RBM estimation depends on the situation we have to deal with (see section 5.2.2). For example, when we deal with image sequences, we can apply a continuity constraint, i.e., we can assume that corresponding pixels in consecutive frames have a small distance (see, e.g., [94]). However, for 3D–2D pose estimation from a single image (see, e.g., [82]) we can not apply this constraint. If we have multiple motions, e.g., as in our car scenes, the correspondence problem becomes much more severe since we have, on top of the correspondence problem for single motion estimation, to find a separation of the data set that corresponds to the different RBMs.
We will further see in section 5.6.4, that correspondences of different kind of entities have 'different weight' in the sense that they lead to different number of constraint equations. As a consequence, different number of correspondences are needed for different visual entities to be able to compute the RBM. For example,

- a correspondence of a 3D point with a 3D point gives us three independent constraint equations and we need at least three independent 3D/3D point correspondences to compute an RBM,

- a correspondence of a 2D point with a 3D point gives us two independent constraint equations and we need again three 2D point/3D point correspondences to compute the RBM,

- a 2D point / 2D line correspondence gives us only one constraint equations. Then we need six 2D point / 2D line correspondences to compute the RBM.

Note that in case of more complex entities (that are formed by combinations of more primitive entities) less correspondences are needed since the constraints of each of the more primitive entities can be combined. For example in case of a 3D junction with three outgoing lines that is brought to correspondence with a similar 3D junction in the second frame only 1 correspondence is needed since we have one 3D/3D point constraint and three constraints in the outgoing lines.
If we have, e.g., a feature set of 1000 image features and 1000 3D features and we would need 3 correspondences to compute an RBM then we have approximately $1000^3 = 10^9$

possible correspondences to consider. Even when we neglect the problem that corresponding features may not be extracted because of the ambiguity in visual data this space is not computable in any real time scenario.
There is one 'easy way' to solve the correspondence problem and that is to label correspondences by hand (as done e.g., in the standard 3D extraction software [51]). However, this is not satisfying since a manual intervention would be necessary in each situation. Thus, it has turned out that it is the correspondence problem that is crucial in the context of RBM estimation (see, e.g., [9]).
From the discussion in 5.0.2 about the correspondence problem in the stereo domain it became clear that constraints are essential to reduce the correspondence problem and in the following we will discuss such constraints for RBM estimation. It will turn out that *grouping in addition to other constraints can be an essential way to deal with the combinatorial explosion.*

- **Multiple Modalities:** As in the stereo case it is advantageous to use different modalities for the elimination of wrong matches. The power of this constraint depends on the situation and the modality. E.g., in case that markers of different colour are associated to an object, colour alone can solve the correspondence problem (see, e.g., [102]). However, these situations are in some sense artificial and in natural scenes a combination of different modalities (weighted according to the current situation) will give the best performance. This is why we represent different modalities in our object representations (see, e.g, [79]). It has been shown that also the human visual system makes use of different modalities to improve matching performance (see, e.g., [46]).

- **Initial Estimate based on few Correspondences:** For RBM estimation we only need a small number of correspondences (see section 5.4). Therefore, we can compute an RBM by using only this small set of correspondences and then check whether there exist other entities that can be brought to correspondence by the computed RBM. This is the underlying principle in the so called RANSAC (Random Sample Consensus) algorithm [34].

- **Continuity:** The continuity constraint is applicable in image sequences. It is very powerful since it reduces the correspondence problem to a small area. Furthermore, optic flow can give information where the corresponding entity is supposed to be (see, e.g., [75]). Finally, correspondences need not to be defined in a two frame scheme only but can be verified over a number of frames for which a similar RBM can be assumed. In the last decade, it has turned out that the continuity constraints is sufficient to solve the structure from motion problem in quite complex scenarios (see, e.g., [44]).

- **Epipolar Constraint:** For RBM estimation no epipolar line constraint can be used since it is the RBM that establishes the epipolar geometry. However, once an RBM is computed we can use the epipolar constraint to decrease the search space for finding further correspondences (see, e.g., [94, 110]).

5.5 RBM Estimation and Grouping

In section 5.1 we have introduced the RBM estimation problem. For feature beased methods (see section 5.2.1) we have the option to formulate correspondences for entities of different dimension (see 5.3.1) and different complexity (see 5.3.2). As discussed in section 5.4 the correspondence problem is cruical in the context of RBM estimation.
From this discussion can now identify four desired properties in the context of RBM estimation algorithms. All these properties are connected to the grouping problem.

- **Accuracy:** We want to have a high degree of precision in the estimation of parameters associated to the entities brought to correspondence in equation (5.2) and (5.3) since any deviation from the truth leads to distortions within the constraint equations and subsequently distorts the computed RBM.
- **Reliability:** Different kind of visual entities may be extracted with different reliability. For example, an edge and its associated orientations can be extracted with higher reliability in case of high contrast compared to a low contrast patch and also 3D points can be computed by stereo matching with different degree of reliability. In the context of RBM estimation, we are interested in preferably using entities that are reliable. Therefore, we want to code features as well as their reliability. Note that this presupposes some degree of explicitness in our representations since a distinction between reliable and unreliable features is not possible for implicit representations.
- **Flexibility:** We want to make flexible use of correspondences, i.e., we want to mix them. Therefore, we are looking for RBM estimation methods that can deal with several kinds of entities at the same time. For example, if we have found a reliable point correspondence and two reliable line correspondences, we want to use these 3 correspondences to estimate the RBM, i.e., we want to apply and mix them within one system of equations.
- **Minimality:** As will be discussed in section 5.4, different kind of correspondences have different value in the sense that they lead to a different number of constraint equations. Since the space of possible correspondences increases exponentially with the number of features we are interested in estimating an RBM with as few correspondences as possible. Therefore we are after descriptors of high complexity.

Grouping, in addition to the other constraints, can play an important role to reduce the RBM estimation problem. Grouping addresses three of the above–mentioned properties: Accuracy, Reliability and Minimality. However, grouping demands Flexibility.

- **Accuracy:** Within a group semantic properties of entities can be estimated with higher accuracy. For example, the orientation and position of a line can be interpolated by taking a number of points into account (see, e.g., [50]).
- **Reliability:** Groups of entities have higher reliabilty than single entities since they are confirmed by their context. For RBM estimation, we can start in a natural way with correspondences of larger groups, i.e., we can make functional use of correspondences of different reliability (see, e.g., [75]).
- **Flexibility:** Since groups may consist of different kinds of entities (e.g., points and line–like features, see figure 5.4) the utilised RBM estimation algorithm needs to allow for dealing with different kinds of entities.
- **Minimality:** The number of necessary correspondences to compute one RBM is much smaller if entities are combined into groups. If, for example, a group is constituted by a corner point and the three lines intersecting in this point (see figure 5.4b), one correspondence is sufficient.

5.6 Mathematical Formulation of the RBM Estimation Problem

So far we have addressed underlying problems of RBM estimation (such as, e.g., the correspondence problem and the problem of choosing and mixing of visual entities) without looking at concrete mathematical formulations of RBM and the RBM estimation problem. This will be addressed now. We will see that the mathematical formalization of RBM estimation is to a certain extent crucial and that all problems defined so far are deeply intertwined with the mathematical representation.
This part necessarily has to deal with a mathematical framework of considerable complexity. However, the reader who is not interested in this issue might directly skip to section 5.7.

5.6.1 Different kind of Optimisation Algorithms

The constraint equations (5.2) and (5.3) lead to a set of equations for which an optimal solution has to be found. The set of equations generally is overdetermined and a best solution has to be found by numerical optimization methods.

We distinguish between linear and non–linear optimisation methods that both have different advantages and disadvantages. For example, when we formulate an RBM as a matrix, our system of equations is linear and we can use standard optimisation methods to find the best matrix that minimizes the error

$$||RBM(\mathbf{p}) - \mathbf{p}'|| = ||A^{RBM}\mathbf{p} - \mathbf{p}'|| \tag{5.5}$$

where A^{RBM} is the matrix that represents the RBM.
However, what we get does not need to be an RBM since not all matrices represent an RBM[14]. Therefore, additional (non–linear) constraints need to be defined to make sure that the matrix represents an RBM (see, e.g., [29]).
Using non–linear methods (see, e.g., [121]) we can make sure that we formalise the RBM estimation problem in the appropriate space. It has been shown that with these methods often also a higher accuracy can be achieved (see, e.g, [110]). However, the theory of systems with non–linear equations is much more complex and statements about uniqueness of solutions, convergence etc. are much harder to establish.
As will be shown in section 5.6.4, the pose estimation algorithm [103, 102, 104, 40, 101] combines some of the advantages of linear and non–linear optimization methods.

5.6.2 Mathematical Formalisations of Rigid Body Motion

A Rigid Body Motion $RBM^{(\mathbf{t},\mathbf{r})}$ as well as visual entities can be formalised in different ways. For example, an RBM of a 3D point $\mathbf{x} = (x_1, x_2, x_3)$ that is represented in homogeneous coordinates as the 4D vector $(x_1, x_2, x_3, 1)$ can be formalised by a 4×4 matrix [29] and an RBM of a line as dual quaternions [109]. In the following, we will give a description of different possible formalisations of RBM.

- **Matrix Formulation.** The most common formulation of RBM is in matrix form (see, e.g., [29]). A $RBM^{(\mathbf{t},\mathbf{r})}$ can be written as

$$RBM^{(\mathbf{t},\mathbf{r})} = \begin{pmatrix} r_{11} & r_{21} & r_{31} & t_1 \\ r_{12} & r_{22} & r_{32} & t_2 \\ r_{13} & r_{23} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} A(\mathbf{r}) & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix} \tag{5.6}$$

  The 4×4 matrix consists of a rotational part that can be described by the 3×3 matrix $A(\mathbf{r})$ (that has orthogonal columns and determinant 1) and a translation vector $\mathbf{t}$. $\mathbf{r}$ codes the axis of rotation as well as the angle of rotation in its length

[14] In general when using matrices, an RBM is coded as a 4×4 matrix. In this case the optimization method would search in a 16–dimensional space instead of a 6–dimensional.

($||\mathbf{r}|| = \alpha$). Note that $A(\mathbf{r})$, although spanned by the 3–dimensional, vector $\mathbf{r}$ has 9 dimensions.

This formulation has different advantages. First, matrix algebra is very common and well understood. Each matrix represents a linear map and the well derived theory of linear systems can be applied. However, one fundamental problem of the matrix formulation is that it formulates the RBM estimation problem in a space with too mnay degrees of freedom. An RBM is described by 6 parameters and not by 12 or 16. So there are at least 6 degrees of freedom too much. This leads to problems when we want to optimise our system of linear equations (see section 5.6.1): First, the solution might not correspond to an RBM. Second, due to the large search space such an approach is noise sensitive.

- **Quaternions and Dual Quaternions:** A more compact representation of rotation of points can be realized by the use of quaternions. A quaternion is a four dimensional vector

$$\mathbf{q} = (q_1, q_2, q_3, q_4) = p_1 + iq_2 + jq_3 + kq_4$$

for which a multiplication $\mathbf{q_1 q_2} = \mathbf{q_3}$ is defined by $i^2 = j^2 = k^2 = ijk = -1$ (see, e.g., [12]). The rotation of a point

$$\mathbf{p} = (0, p_1, p_2, p_3)$$

around an axis $\mathbf{w} = (w_1, w_2, w_3)$ with angle α can be described by the unit quaternion

$$\mathbf{q} = (cos(\frac{\alpha}{2}), sin(\frac{\alpha}{2})w_1, sin(\frac{\alpha}{2})w_2, sin(\frac{\alpha}{2})w_3)$$

and the final rotation can be described by

$$\mathbf{p}' = \mathbf{q}\mathbf{p}\bar{\mathbf{q}}$$

where $\bar{\mathbf{q}}$ is the conjugate of $\mathbf{q}$. This kind of formulation has been used, e.g., by [93]. In contrast to the matrix formulation of rotation that has 6 degrees of freedom to much, for the quaternion formulation we have only one additional degree of freedom.

Dual Quaternions are an extension of quaternions (see, e.g., [12]) that can be used to describe the RBM of lines (see, e.g., [109]). They represent an eight–dimensional formulation of the 6 dimensional problem. By introducing additional constraints on the norm of dual quaternions the problem can be reduced to 6–dimensions.

- **Exponential Representation (Twists):** The pose estimation algorithm [103, 102, 104, 40, 101] makes use of a formulation of RBM based on twists. We therefore

describe twists in more detail now. Twists have a straightforward linear approximation (using a Taylor series expension) and lead to a formalization that searches in the 6 dimensional space of RBMs. Our description is motivated by (and close to) the description given by Oliver Granert [40]. A formalization of the very same approach using geometric algebra is given in [103, 102, 104, 101].

The rotation matrix $A(\mathbf{r})$ can also be defined as the limit of a Taylor series. A rotation of a point $\mathbf{p}$ around an axis $\mathbf{w} = (w_1, w_2, w_3)$ with an angle α can be described by

$$\mathbf{p}' = e^{\tilde{w}\alpha}\mathbf{p} = \mathbf{A}(\mathbf{r})\mathbf{p}.$$

$e^{\tilde{w}\alpha}$ is the matrix that is constituted by the limit of the Taylor series

$$e^{\tilde{w}\alpha} = \sum_{n=0}^{\infty} \frac{1}{n!}(\tilde{w}\alpha)^n \tag{5.7}$$

with

$$\tilde{w} = \begin{pmatrix} 0 & -w_3 & w_2 \\ w_3 & 0 & -w_1 \\ -w_2 & w_1 & 0 \end{pmatrix}, \text{ with } ||\mathbf{w}|| = 1.$$

The exponential representation allows for a straightforward linearisation by using only the first two terms of (5.7), i.e.,

$$e^{\tilde{w}\alpha} \approx I_{3\times 3} + \tilde{w}\alpha. \tag{5.8}$$

On the other hand, having $\tilde{w}$ and α we can compute $\mathbf{A}(\mathbf{r})$ by the formula of Rodriguez (see, e.g., [86]):

$$\mathbf{A}(\mathbf{r}) = I + sin(\alpha)\tilde{w} + (1 - cos(\alpha))\tilde{w}\tilde{w}. \tag{5.9}$$

The exponential representations can be extended to an RBM. However, for this we need to apply another understanding how the RBM is constituted. In figure 5.2b an RBM is understood as a rotation of angle α around a line l in 3D space with direction $\mathbf{w}$ and moment $\mathbf{w} \times \mathbf{q}$ (see section 5.6.3). In addition to the rotation a translation with magnitude λ along the line $\mathbf{l}$ is performed. According to Chasles' theorem, each RBM can be expressed in this way (see, e.g., [86]).

Then an RBM can be represented as

$$\mathbf{p}' = e^{\tilde{\xi}\alpha}\mathbf{p} = RBM\mathbf{p}$$

with

$$e^{\tilde{\xi}\alpha} = \sum_{n=0}^{\infty} \frac{1}{n!}(\tilde{\xi}\alpha)^n \tag{5.10}$$

with $\tilde{\xi}$ being the 4×4 matrix

$$\tilde{\xi} = \begin{pmatrix} \tilde{w} & -\tilde{w}\mathbf{q} + \lambda\mathbf{w} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -w_3 & w_2 & w_3q_2 - w_2q_3 + \lambda w_1 \\ w_3 & 0 & -w_1 & w_1q_3 - w_3q_1 + \lambda w_2 \\ -w_2 & w_1 & 0 & w_2q_1 - w_1q_2 + \lambda w_2 \\ 0 & 0 & 0 & 0 \end{pmatrix} =$$

$$\begin{pmatrix} 0 & -w_3 & w_2 & v_1 \\ w_3 & 0 & -w_1 & v_2 \\ -w_2 & w_1 & 0 & v_3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

with

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} w_3q_2 - w_2q_3 + \lambda w_1 \\ w_1q_3 - w_3q_1 + \lambda w_2 \\ w_2q_1 - w_1q_2 + \lambda w_2 \end{pmatrix}$$

In analogy to (5.8) a straight forward linearisation is given by

$$e^{\xi\alpha} \approx (I_{3\times 3} + \tilde{\xi})\alpha. \tag{5.11}$$

Having $\mathbf{w}$, α, and $\mathbf{v}$, we can apply the formula of Rodriguez for the RBM to get the matrix representation:

$$\mathbf{t} = (I - \mathbf{A}(\mathbf{r}))\tilde{w}\mathbf{v} + \alpha\mathbf{w}\mathbf{w}^T\mathbf{v}$$

and $\mathbf{A}(\mathbf{r})$ is computed as in equation (5.9).

At this point, we have expressed an approximation of an RBM as a 4×4 matrix. Up to now nothing seems to be won compared to the matrix formulation in (5.6), since we still deal with a 12 dimensional description. However this representation expresses the motion parameters directly and, as will be shown in 5.6.4, can be used to derive a formulation that is very compact and efficient.

5.6.3 Parametrisation of Visual Entities

When we want to estimate an RBM we need not only to choose a representation for the RBM but we also need to formalize entities on which the RBM operates. There exist different representations for points and lines that are relevant for the RBM estimation problem.

Explicit Representation: A point can be described explicitely as a vector (p_1, p_2, p_3) and a line $\mathbf{L}$ can be described explicitely by

$$L(\lambda) = \mathbf{p} + \lambda \mathbf{r}$$

with $\mathbf{p}$ being a point on the line and $\mathbf{r}$ its direction. This representation is well established. However, in the context of the RBM estimation problem in our system we make use of an implicit representation. This implicit representation allows for a *direct representation of the distance of correponding entities* that will be cruical for RBM estimation.

Implicit Representation: In the formulation of the RBM estimation problem [103, 102, 104, 40, 101] that we use in our system [75], an implicit representation of entities as null spaces of equations is applied.

- **Implicit Representation of 3D Points:** We can represent a 3D point $\mathbf{p} = (p_1, p_2, p_3)$ by the null space of a set of equations

$$\mathbf{F^p}(\mathbf{x}) = \begin{pmatrix} p_1 - x_1 \\ p_2 - x_2 \\ p_3 - x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{5.12}$$

  If (x_1, x_2, x_3) fullfills this equation it is identical with $\mathbf{p}$. We can write the very same expression in matrix notations by[15]:

$$\mathbf{F^p}(\mathbf{x}) = \begin{pmatrix} 1 & 0 & 0 & -p_1 \\ 0 & 1 & 0 & -p_2 \\ 0 & 0 & 1 & -p_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{5.13}$$

  Note that the value $||\mathbf{F^p}(\mathbf{x})||$ represents the Euclidian distance between $\mathbf{x}$ and $\mathbf{p}$. This will be important to derive interpretable constraint equations (see section 5.6.4).

[15]Note that it must be ensured that the fourth component is equal to one (i.e., $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}$) to let (5.13) be identical to (5.12).

- **Implicit Representation of 3D Lines:** A 3D line $\mathbf{L}$ can be expressed as two 3D vectors $\mathbf{r}, \mathbf{m}$. The vector $\mathbf{r}$ describes the direction and $\mathbf{m}$ describes the moment which is the cross product of a point $\mathbf{p}$ on the line and the direction

$$\mathbf{m} = \mathbf{p} \times \mathbf{r}.$$

$\mathbf{r}$ and $\mathbf{m}$ are called Plücker coordinates. If we assume that $\mathbf{r}$ has length 1 this representation is unique up to a sign[16].

The null space of the equation

$$\mathbf{x} \times \mathbf{r} - \mathbf{m} = \mathbf{0}$$

is the set of all points on the line.

In matrix form this reads

$$\mathbf{F}^{\mathbf{L}}(\mathbf{x}) = \begin{pmatrix} 0 & r_x & -r_y & -m_x \\ -r_z & 0 & r_x & -m_y \\ r_y & -r_x & 0 & -m_z \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = 0 \tag{5.14}$$

Note that the value $||\mathbf{F}^L(\mathbf{x})||$ can be interpreted as the Euclidian distance between the point (x_1, x_2, x_3) and the closest point on the line to (x_1, x_2, x_3) [56, 101].

- **Implicit Representation of 3D Planes:** A 3D plane $\mathbf{P}$ can be parametrised by the unit normal vector $\mathbf{n}$ and the Hesse distance d_H using the equation:

$$\mathbf{n} \cdot \mathbf{p} = d_H.$$

In matrix formulation this reads:

$$F^{\mathbf{P}}(\mathbf{x}) = \begin{pmatrix} n_1 & n_2 & n_3 & -d_H \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{5.15}$$

Note that $F^{\mathbf{P}}(\mathbf{x})$ describes the Euclidian distance between the closest point on $\mathbf{P}$ to $\mathbf{x}$.

In section 5.6.4, we will see that this implicit representation of entities in combination with the twist representation of an RBM (see section 5.6.2) and the formulation of the pose estimation problem in the Euclidian space (see section 5.3.1) allows for defining suitable and geometrically interpretable constraint equations.

[16] The uniqueness can be easily proven: Let $\mathbf{p}_1$ and $\mathbf{p}_2$ be two points on the line then $\mathbf{p}_2 = \mathbf{p}_1 + \lambda\mathbf{r}$. Therefore, $\mathbf{p}_2 \times \mathbf{r} = (\mathbf{p}_1 + \lambda\mathbf{r}) \times \mathbf{r} = \mathbf{p}_1 \times \mathbf{r} + \lambda\mathbf{r} \times \mathbf{r} = \mathbf{p}_1 \times \mathbf{r} + \mathbf{0} = \mathbf{p}_1 \times \mathbf{r}$

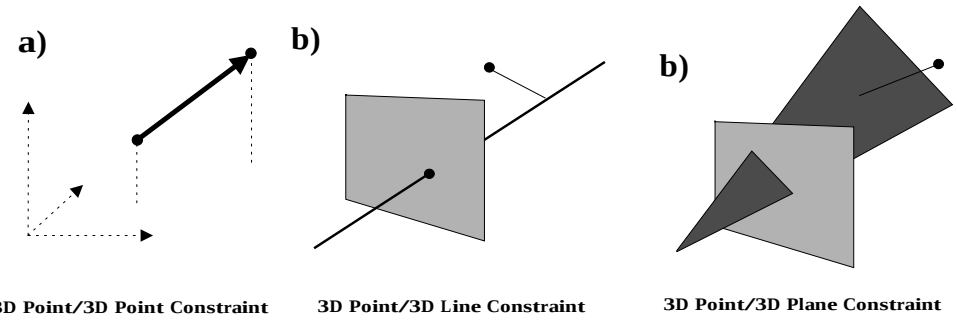


Figure 5.5: Geometric Interpretation of constraint equations. a) The 3D-3D point constraint realized the Euclidian distance between the two points. b) The 3D point/3D line constraint realizes the shortest Euclidian distance between the 3D Point and the 3D line. c) The 3D Point/3D Line constraint realizes the shortest Euclidian distance between the 3D Point and the 3D Plane.

5.6.4 Constraint Equations

After having formalized an RBM as a twist transformation in section 5.6.2 and geometric entities in section 5.6.3 we can now define constraint equations for different kind of correspondences.

3D-point/3D-point constraint: One can express the constraint equation (5.4) for the case that our corresponding entities are 3D points by using the linear approximation (5.11) of the twist $\tilde{\xi}\alpha$ and the implicit representation of points (5.12) by

$$\mathbf{F}^{\mathbf{p}'}((I_{3x3} + \tilde{\xi}\alpha)\mathbf{p}) = \mathbf{0}.$$

In matrix form this reads

$$\begin{pmatrix} 1 & 0 & 0 & -p'_1 \\ 0 & 1 & 0 & -p'_2 \\ 0 & 0 & 1 & -p'_3 \end{pmatrix} \begin{pmatrix} 1 & -\alpha w_3 & \alpha w_2 & \alpha v_1 \\ \alpha w_3 & 1 & -\alpha w_1 & \alpha v_2 \\ -\alpha w_2 & \alpha w_1 & 1 & \alpha v_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Any deviation from $\mathbf{0}$ describes a vector whose norm is the Euclidian distance from p, i.e, it describes a geometrically interpretable measure (see figure 5.5a).

By simply re-ordering the system we get:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & p_3 & -p_2 \\ 0 & 1 & 0 & -p_3 & 0 & p_1 \\ 0 & 0 & 1 & p_2 & -p_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha v_x \\ \alpha v_y \\ \alpha v_z \\ \alpha w_x \\ \alpha w_y \\ \alpha w_z \end{pmatrix} = \begin{pmatrix} p'_1 - p_1 \\ p'_2 - p_2 \\ p'_3 - p_3 \end{pmatrix}.$$

Note that our optimisation method now directly acts on the parameters of the RBM. Since $||\mathbf{w}|| = 1$, α represents the angle of rotation.

3D point/2D point constraint: We now want to formulate constraints between 2D image entities and 3D object entities. Given a 3D point $\mathbf{p}$ and a 2D point p we first generate the 3D line $\mathbf{L}(\mathbf{r}, \mathbf{m})$ that is generated by the optical center and the image point (see figure 5.5a).[17] Now the constraint reads:

$$\mathbf{F}^{\mathbf{L}(p)} \left((I_{3\times 3} + \tilde{\xi}\alpha)\mathbf{p} \right) = 0.$$

Using the implicit representation of 3D lines in (5.14) we get:

$$\begin{pmatrix} 0 & r_1 & -r_2 & -m_1 \\ -r_3 & 0 & r_1 & -m_2 \\ r_2 & -r_1 & 0 & -m_3 \end{pmatrix} \begin{pmatrix} 1 & -\alpha w_3 & \alpha w_2 & \alpha v_1 \\ \alpha w_3 & 1 & -\alpha w_1 & \alpha v_2 \\ -\alpha w_2 & \alpha w_1 & 1 & \alpha v_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Once again we can make use of the intuitive geometrically interpretable measure coming along with the implicit representation of our geometric entities introduced in section 5.6.3 (see also figure 5.5b).
Simple reordering gives:

$$\begin{pmatrix} 0 & -r_3 & r_2 & -p_3 r_3 - p_2 r_2 & p_1 r_2 & p_1 r_3 \\ r_z & 0 & -r_x & p_2 r_1 & -p_1 r_1 - p_3 r_3 & p_2 r_3 \\ -r_2 & r_x & 0 & p_3 r_1 & p_3 r_2 & -p_2 r_2 - p_1 r_1 \end{pmatrix} \begin{pmatrix} \alpha v_x \\ \alpha v_y \\ \alpha v_z \\ \alpha w_x \\ \alpha w_y \\ \alpha w_z \end{pmatrix} =$$

$$\begin{pmatrix} p_3 r_2 - p_2 r_3 + m_1 \\ p_1 r_3 - p_3 r_1 + m_2 \\ p_2 r_1 - p_1 r_2 + m_3 \end{pmatrix}.$$

[17] Note that the line $\mathbf{L}$ depends on the camera parameters.

Given a 3D point 2D point correspondence we have now a different set of constraints *that work on the very same RBM parameters*. Therefore we can simply combine these correspondences by adding the set of equations derived from the 3D point/3D point correspondence to the set of equations derived from the 3D point/2D point correspondences.

3D Point/2D Line constraint: Given a 3D point and a corresponding 2D image line l we can construct the 3D Plane $\mathbf{P}(l)$ that is spanned by the image line and the optical center of the camera (see figure 5.5c). We can then define the constraint

$$\mathbf{F}^{\mathbf{P}(l)}((I_{3\times 3} + \tilde{\xi}\alpha)\mathbf{p}) = \mathbf{0}.$$

Using the implicit representation of 3D planes we get the equations

$$\begin{pmatrix} n_1 & n_2 & n_3 & -d_H \end{pmatrix} \begin{pmatrix} 1 & -\alpha w_3 & \alpha w_2 & \alpha v_1 \\ \alpha w_3 & 1 & -\alpha w_1 & \alpha v_2 \\ -\alpha w_2 & \alpha w_1 & 1 & \alpha v_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ 1 \end{pmatrix} = 0.$$

Reordering leads to the constraint equations:

$$\begin{pmatrix} n_1 & n_2 & n_3 & -n_3p_2 - n_2p_3 & -n_1p_3 - n_3p_1 & -n_2p_1 - n_1p_2 \end{pmatrix} \begin{pmatrix} \alpha v_x \\ \alpha v_y \\ \alpha v_z \\ \alpha w_x \\ \alpha w_y \\ \alpha w_z \end{pmatrix} =$$

$$\begin{pmatrix} -d_H - n_1p_1 - n_2p_2 - n_3p_3 \end{pmatrix}.$$

Figure 5.5c shows the geometric interpretation of the 3D point/2D line constraint.

5.7 Properties of Rosenhahn et al's RBM estimation algorithm

In this section, we have discussed different aspects of the RBM estimation problem. We have especially addressed the problem of choosing good entities for RBM estimation and we have seen that this is cruical in terms of the correspondence problem. It turned out that these issues are deeply intertwined with the mathematical representation of the RBM and the estimation problem.
The representation of the RBM estimation problem introduced by [103, 102, 104, 40, 101] that has been described in section 5.6.3 and 5.6.4 has several advantages:

Searching in the space of RBMs: It leads to a set of equations that (although approximated) directly acts on the RBM parameters. The final RBM is computed iteratively. Twists have been proven to be an efficient representation of RBM enabling such a formalization. Twists have been also used by [15], although for an orthographic formulation of the RBM estimation problem.

Geometric Interpretation: The constraint equations give a geometrically interpretable intuitive measure in terms of Euclidian distance. This has become possible by making use of an implicit representation of geometric entities introduced in section 5.6.3. Implicit representations of geometric entities had also been used by [56] but had not been applied to the pose estimation problem before.

Mixing of different Entities: Correspondences of different kinds of entities can be mixed. This concerns differences in dimension as well as in complexity. This issue has also been adressed by, e.g., [124].

In the discussion, we have also seen that grouping can play an important role to overcome problems of RBM estimation in terms of four properties: Accuracy, reliability, flexibility and minimality. In the next section, we therefore address grouping in more detail.

Chapter 6

Time-Space Gestalts

Based on the multi-modal Primitives (the Primitives as well as their biological motivation by hypercolumns are described in [77]) we have developped temporal-spatial Gestalts. Rigid body motion (RBM) is the underlying regularity that binds Primitives derived from single frames together (a detailed review about RBM, its formalisation, estimation and utilisation as well as its potential combination with grouping processes is described in [80]). The process that leads to temporal-spatial Gestalts is schematically described in figure 6.1. In this scheme the change of visual entities across different frames is predicted and correspondences lead to an incresae of confidences (while non–correspondences lead to an decresae of confidences) as well as to an interpolation of parameters of entities. This scheme has already been described in the last report and in [75]. However, this scheme has not been used within our framework of multimodal Primitives. This we have achieved now. By doing this we have realised some problems that needed to be solved as descrined below.
We will first give a short description of the application of the scheme, the specific problems we have encountered as well as their solution. Then we will describe the results that we have achieved.

6.1 Formalization of Spatial-Temporal Gestalts and their Utilization for Disambiguation of Stereo Information

This scheme is of rather generic nature. However, for its application a number of crucial details were to be solved:

1) **Change of entities across frames:** The transformation of entities across frames must be formalized. In our scenes the dominated change is caused by ego-motion which can be described by an RBM (see the circled 1 in figure 6.2). However, the

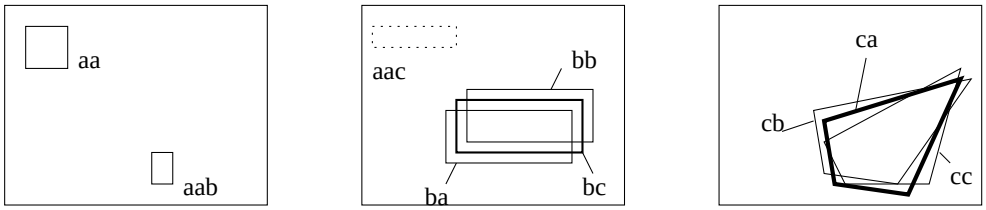


Figure 6.1: The accumulation scheme. The entity e^1 (here represented as a square) is transformed to $T^{(1,2)}(e^1)$. Note that without this transformation it is barely possible to find a correspondence between the entities e^1 and e^2 because the entities show significant differences in appearance 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.

RBM can only be applied directly to 3D entities and not to image entities such as the Primitives (see also [80]). Therefore 3D entities must be reconstructed from different perspective views of a stereo pair of images (see the circled 2 in figure 6.2). For this we make use of the different modalities coded in the Primitives. Furthermore, after applying the RBM to the 3D entity, this entity must be reprojected to the stereo image pair of the the second frame to be comparable to the extracted Primitives. That means that beside the formalisation of the change of entities during an RBM also the reconstruction and reprojection problem needs to be addressed (see the circled 3 in figure 6.2). Reconstruction is done from stereo correspondences that have been found by using a multi-modal stereo matching that makes use of all aspects coded in the Primitives. We have also investigated the importance of the different aspects for stereo matching (see [95, 73, 71]). Reprojection addresses beside the geometric information also all other modalities coded in the Primitives.

2) **Comparison of entities:** When we want to find correspondences of transormed entities and Primitives in a frame a comparison of entities according to some metric is required. Here we have a couple of choices that may lead to quite different results. For example we can perform a comparison of 3D entities, i.e., we formalise the process in Euclidian space. However, we found out that such a formalisation

(as done, e.g., in [75]) leads to problems since reconstruction acuity depends on depths, i.e., we would need to apply a metric in an unhomogeneous space. Entities that have large depths would tend to find less likely correspondences then entities that are close to the camera. The solution to this problem that we have chosen is a formalisation of the metric in the image space in which errors reflect a more homogenous behaviour (see the circled 4 in figure 6.2).

3) **Handling of different Modalities:** The visual Primitives carry beside the geometric information position and orientation also non-geometrical infomation such as phase and colour. However, since Rigid Body Motion only describes only the change of the geometric components we need to approximate the change of phase and colour. Furhermore, in the comparison step (see above) we have so far only used position and orientation. However, the comparison becomes more efficient when we also use the other modalities. For example a transformed red/green edge might be similar in orientation and position to an extracted Primitive but very different in ist colour attributes and should then not be seen as a correspondence. Therefore, we now use a comparison that takes all these modalities into account (see the circled 4 in figure 6.2).

4) **Update Rule:** When a correspondence has been found it needs to be decided how the parameters of the entities influence each other. Moreover, in the scenes entities can be out of frame after a motion when the objects have been passed by the camera. The naive application of the scheme described inb 6.1 would lead to a decrease of confidences of such hidden entities and valuable already generated knowledge would be lost. We therefore adpated the scheme to these out-of-frame situation such that entities for which the position is predicted as being out of frame are not altered once they have achieved a certain confidence.

6.2 Results

We have applied spatial-temporal Gestalts to stabilize ambigious stereo information for artificial and natural scenes. Figure 6.4, 6.5, and 6.6 shows the results.

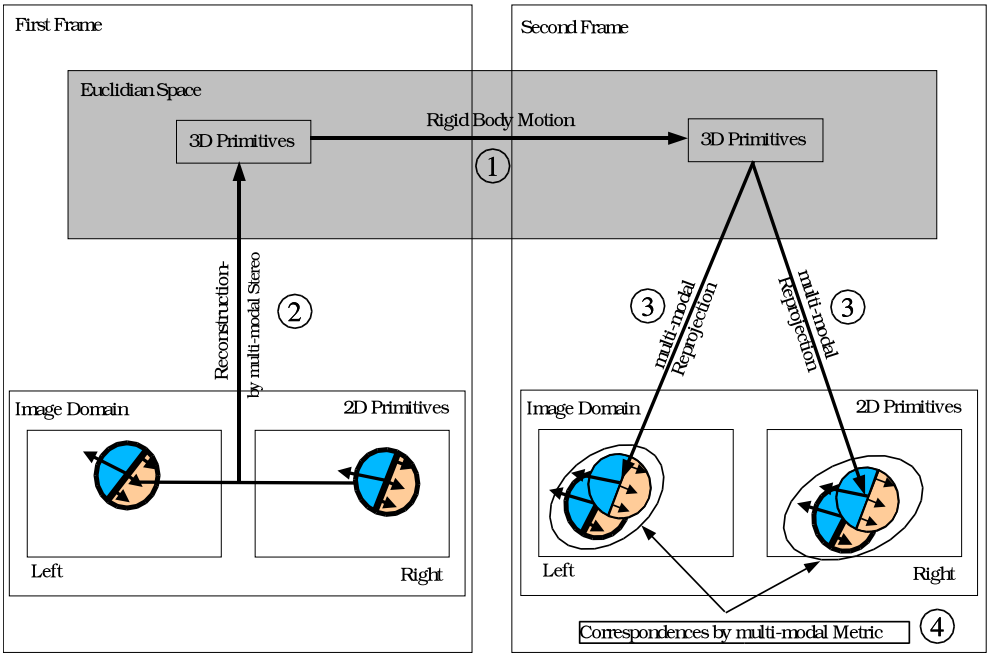


Figure 6.2: A more detailed description of one iteration of the scheme shown in figure 6.1 that points to problems concerning specific subaspects.

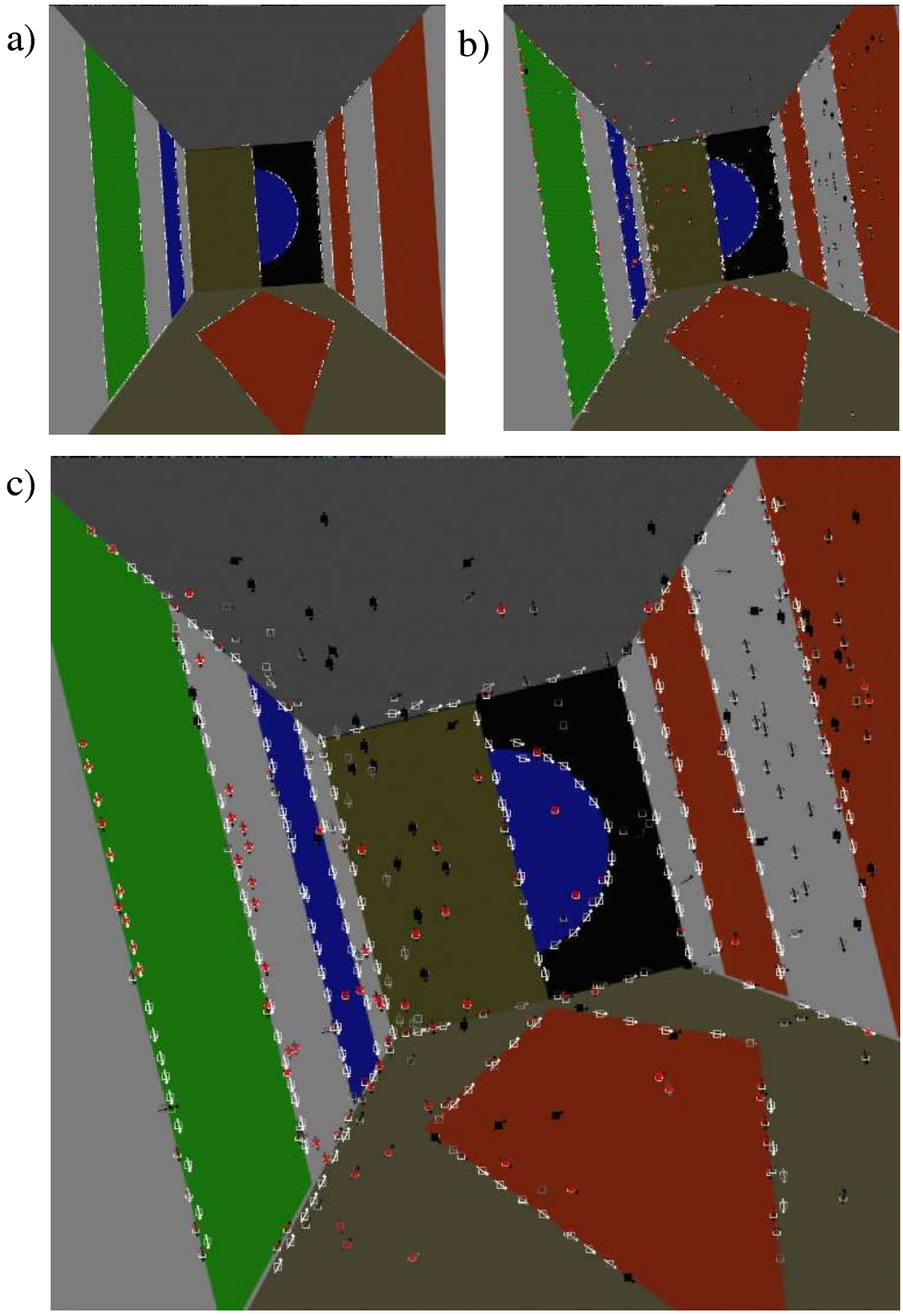


Figure 6.3: Development of spatial-temporal Gestalts across frames for the artificial sequence for frame 1, 5, and 10. The RBM in the artificial sequence is described by a translation that has significant term in x,y, and z direction and a rotation around the z axis. All visual spatial-temporal Gestalts that have been generated in the process described in section 6.1 are shown on the left side while only the spatial-temporal Gestalts with high confidence are displayed on the right side. All displayed entities have been developping over different frames by transforming the entities according to the computed RBM and updating according to the scheme. The graylevel of the orientation bar represents the confidence that has been accumulated over time. Entities which display a black square are "dead" entities, i.e. entities that have not been updated for a certain number of frames. Red entities are newly generated entities, i.e. entities for which no correspondences have been found and therefore a new hypothesis has been created.

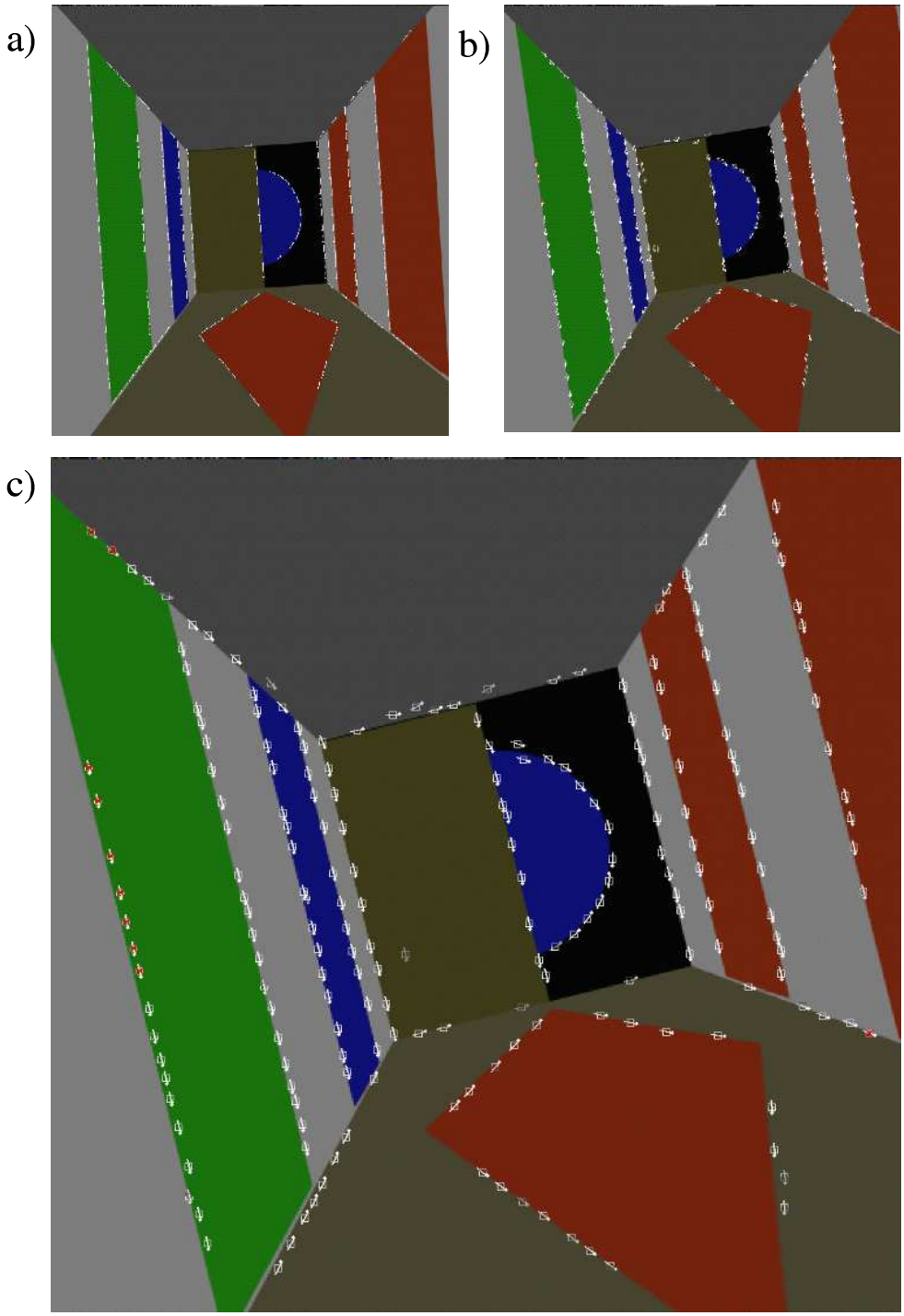


Figure 6.4: Development of spatial-temporal Gestalts across frames for the artificial sequence for frame 1, 5, and 10 showing high confidence entities only.

Figure 6.5: Development of spatial-temporal Gestalts across frames for the natural sequence for frame 1 and 10. In the natural sequence recorded in co-operation with Hella there is a translation with dominant z-component. All visual spatial-temporal Gestalts that have been generated in the process described in section 6.1 are shown.

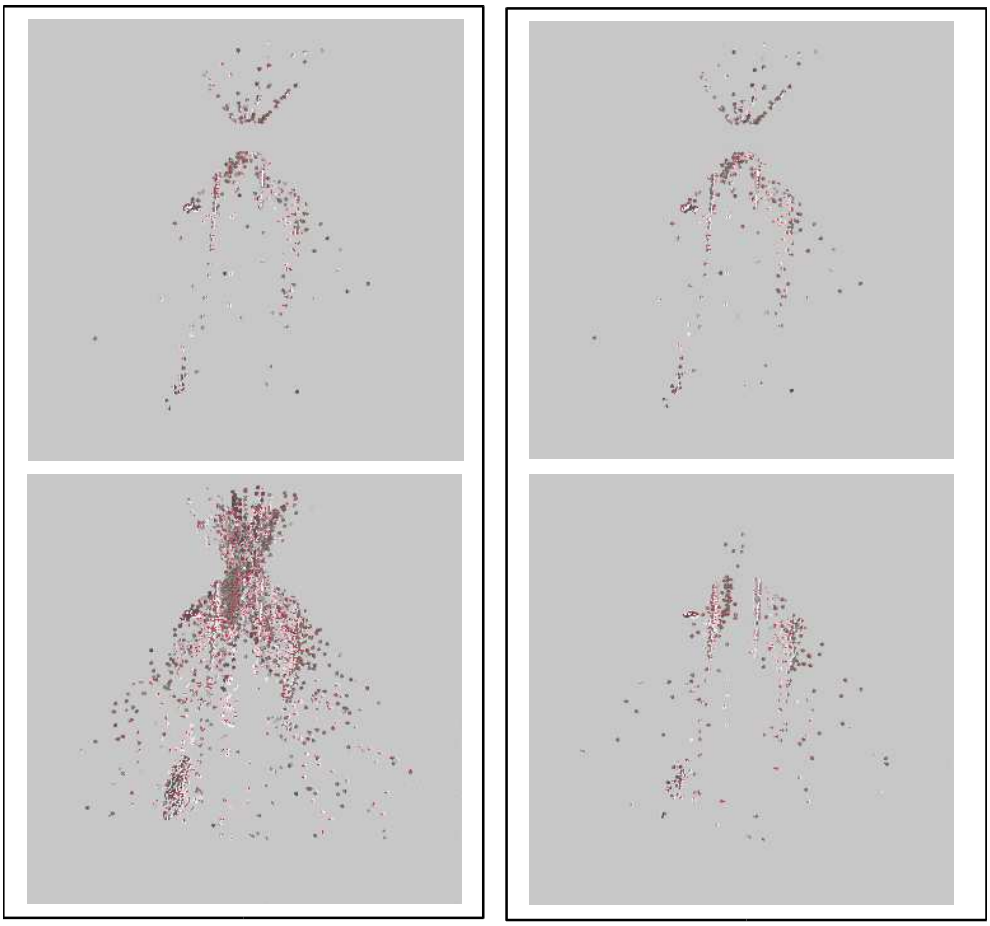


Figure 6.6: Top view of 3D spatial-temporal Gestalts without (left) and with (right thresholding. the first (top) and fifteenth (bottom) frame is shown.

Chapter 7

Prelimenary steps on higher level segments: GRouping and Stereo

Introduction

Vision, although widely accepted as the most powerful sensorial modality, faces the problem of an extremely high degree of vagueness and uncertainty in its low level processes such as edge detection, optic flow analysis and stereo estimation [1]. This arises from a number 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 rough estimates of semantic information (e.g., orientation) are possible. The severeness of these problems increases for higher semantic information, such as curvature or junction detection and interpretation. Furthermore, illumination variation heavily influences the measured grey level values and is hard to model analytically (see, e.g., [54]). Extracting information 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 awkward way (see, e.g., [4, 61]).
However, by integrating information over context [92, 47] the human visual systems acquires visual representations which allows for actions with high precision and certainty within the 3D world even under rather uncontrolled conditions. The power of modality fusion arises from the huge number of intrinsic relations in visual data. The aim of the European project ECOVISION (see [24]) is to use such regularities to achieve robust and more complete descriptions of the visual scene.
In this paper, we address a specific context in which aspects of 2D and 3D feature processing become combined. In human vision local visual entities become organised into more complex entities. This processes is usually called grouping (see, e.g., [119]). In computer vision such grouping processes are mostly treated within the image domain [106, 13]. In

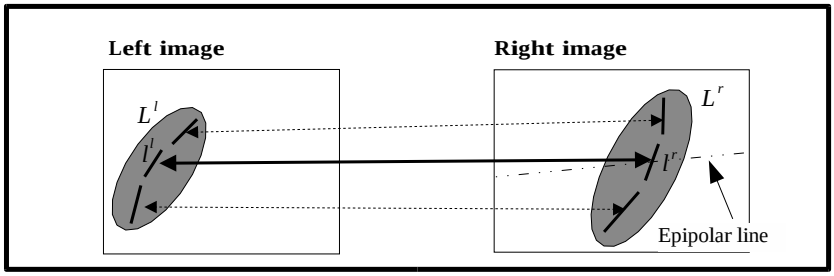


Figure 7.1: Stereo Grouping Constraint

this paper, we start with a grouping process in the 2D image domain. However, this process becomed combined with stereo processing such that coherent 3D groups emerge. The constraint on which this combination is based is the following (see also figure 7.6):

> **Stereo Collinearity Constraint:** Primitives constituting a group in the left image have stereo correspondences in the same group in the right image.

In this paper, we use this constraint to improve stereo processing. Stereo is necessarily ambiguous when based on local comparisons since the correspondence problem leads to mismatches. Using multiple modalities (such as colour or optic flow) improves but can not solve this problem (see [46, 73, 95]).
In this paper, we introduce an artificial visual system in which differerent processes are realized that support each other:

> **2D Feature Extraction:** We have developed an image representation in form of 2D Primitives. These Primitives are multi-modal local descriptors that carry information about aspects such as orientation, contrast transition, colour and optic flow in a condensed way (see figure 7.2 and [77]).
>
> **2D Grouping:** The 2D Primitives are local descriptors that become organised into higher entities in form of collinear groups. In the grouping process a linking structure is establised that makes use of a criterion that utilises collinearity as well as similarity in colour and contrast transition.
>
> **3D Feature Extraction by Stereo:** We use the 2D-Primitives to find stereo correspondences. In this way we compute 3D Primitives from the 2D Primitives. The 3D Primitives carry information about 3D position and 3D orientation in addition to the information of the generating 2D Primitives.

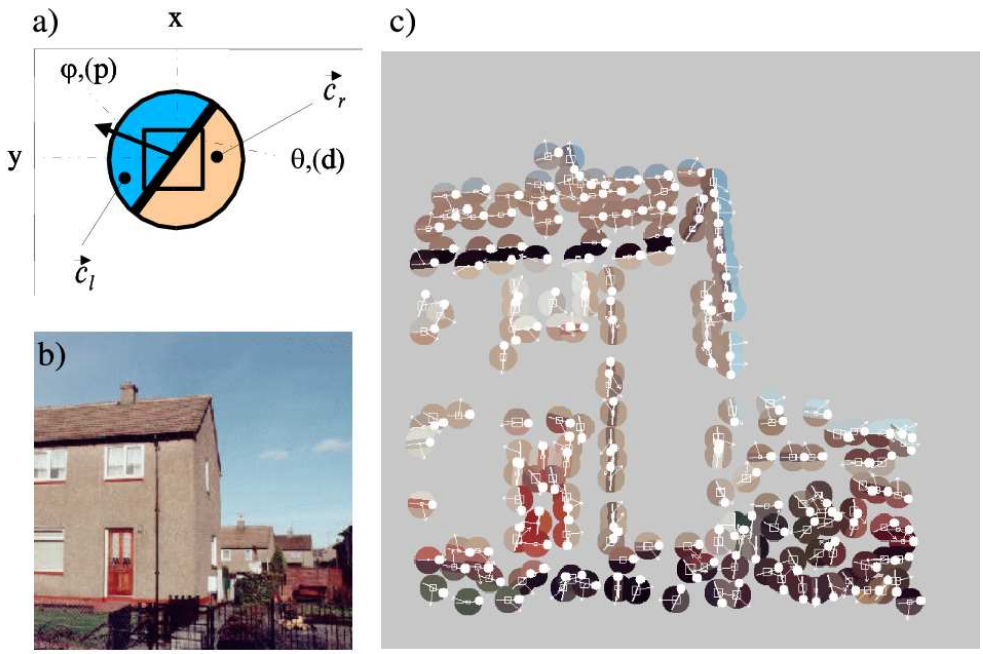


Figure 7.2: **Top left:** Schematic representation of a basic feature vector. Position is coded by (x, y), orientation by θ (or direction as d respectively), phase by φ (or p when associated with a direction), and color by $(\mathbf{c}_l, \mathbf{c}_r)$. **Bottom left:** Frame in an image. **Right:** Extracted feature vectors.

Interaction of Stereo and Grouping: Finally, the group structures are used for improving stereo leading to coherent groups in 3D using the Stereo Collinearity Constraint.

The paper is structured as follows: In section 7.1 we shortly describe our processing of multi-modal Primitives. A more detailed description can be found in e.g., in [77]. The 2D grouping process is described in section 7.2. The multi-modal stereo is described in section 7.3 (further details can be found in [73, 95]) and the integration of grouping and stereo is described in secction 7.4. Results on artificial and real scenes are given in section 7.5.

7.1 Feature Processing

In this section we briefly describe the coding of information (orientation, phase and color) in terms of multi-modal Primitives.

Position, Orientation and Phase: We use a systematic mathematical description of geometric and structural information of grey level images based on the monogenic signal [33]. The monogenic signal performs a *split of identity*, i.e., it orthogonally divides the signal into energetic information (indicating the likelihood of the presence of a structure), its orientation θ and its structure (expressed in the phase φ). Features are extracted at energy maxima in local image patches where the position is parameterized by $\mathbf{x}$ (see figure 7.2).

The phase can be used to interpret the kind of contrast transition at this maximum [67], 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.

Color: The distribution of phases in natural images has been investigated in [79]. There exist clear peaks at $\varphi = \pi/2$ and $\varphi = -\pi/2$ which show that edges (i.e., intrinsic 1-dimensional signals with odd symmetry) are the dominant one-dimensional structure in natural images while line structures (i.e., intrinsic 1-dimensional signals with even symmetry) are less dominant. Our model for an intrinsically one–dimensional signal patch (see figure 7.2) therefore describes edges.[1]

To integrate the modality color at intrinsically one–dimensional image structures we perform an averaging in the RGB color space over the left and right part ('left' and 'right' defined by the associated line segment) of the image patch (see figure 7.2).

We get two vectors $\mathbf{c}_l = (c_r^l, c_g^l, c_b^l)$ and $\mathbf{c}_r = (c_r^r, c_g^r, c_b^r)$, representing the red, green and blue values of the left and right side of the edge.

Therefore, the basic feature vector represented by our Primitives has the form

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

7.2 Establishing Groups by a multi-modal Collinearity Criterion

We want to define group of locally consistent Primitives in the image. We are interested in Primitives outlining major structures of the scenery, and subsequently of the images

[1] Although there is significantly more edge like structures than line like structures in natural images we can also make use of an extra line model to describe intrinsically one-dimensional image patches with phase close to 0 or π. The introduction of this model makes only small difference for stereo matching (but is important in other contexts). We neglect this issue here.

processed. We assume that any structure of the scene having a projective manifestation in the image, has a representation involving a set of consistent Primitives (in the following called group). From this assumption follows naturally that Primitives showing inconsistency with their neighbourhood might be considered as ambigious information likely to be caused by erroneous feature extraction. Now, we want to define the meaning of this *consistency* in the multi-modal space of the features.
In this work, we consider Primitives defining local oriented structures (e.g., lines and step edges). Therefore, we are looking for constellations defining global contours. Consistency between two Primitives is defined by two criterions: Collinearity and Modality Consistency (using the modalities colour and contrast transition). Inconsistency according to these two criterions indicates that the two Primitives are either expressions of independent structures or caused by the erroneous feature extraction process. In the following formulas we will consider a pair of Primitives $\mathbf{e}_1, \mathbf{e}_2$ such as $\mathbf{e}_2 \in N(\mathbf{e}_1)$, N being a large enough neighbourhood. We will consider the coordinate system centered in $\mathbf{e}_1$ and oriented so that $\theta(\mathbf{e}_1) = 0$. We want to define relationships between $\mathbf{e}_1$ and $\mathbf{e}_2$ defining possible structures for $\mathbf{e}_1$ and we code them as links $\mathcal{L}(\mathbf{e}_1, \mathbf{e}_2)$ between them. We associate a confidence $c[\mathcal{L}(\mathbf{e}_1, \mathbf{e}_2)]$ to a link which is an estimate of the probability for the two primitives to be part of the same structure.

7.2.1 Collinearity Criterion

Our collinearity criterion is based on two factors: Proximity and good continuation.

Proximity

Our priximity criterion take evaluate how given the position of the primitive $\mathbf{e}_2$ relatively to the primitive $\mathbf{e}_1$ a link $\mathcal{L}(\mathbf{e}_1, \mathbf{e}_2)$ is likely to exist. The idea here is that the closer the second primitive is to the first, the closer it has to be to the line defined by the orientation of $\mathbf{e}_1$: parallel segments cannot be collinear for example. Also at this very local level we only want to consider low curvatures between the two primitives. To take these aspects into account we define a distance function between two Primitives by

$$C_{position}(\mathbf{e}_1, \mathbf{e}_2) = \frac{1}{1 + e^{\lambda(|x| - max(|y|, 0.3))}} \frac{1}{1 + e^{\lambda(|y| - 0.7)}} \tag{7.1}$$

with $\lambda = 30$ being the steepness parameter
a distance of 1 in the axis means twice the size of the patch generating the Primitives, and zero meaning the generating image patches of the two Primitives are in contact or overlapping.
Figure 7.3 displays the distance function.

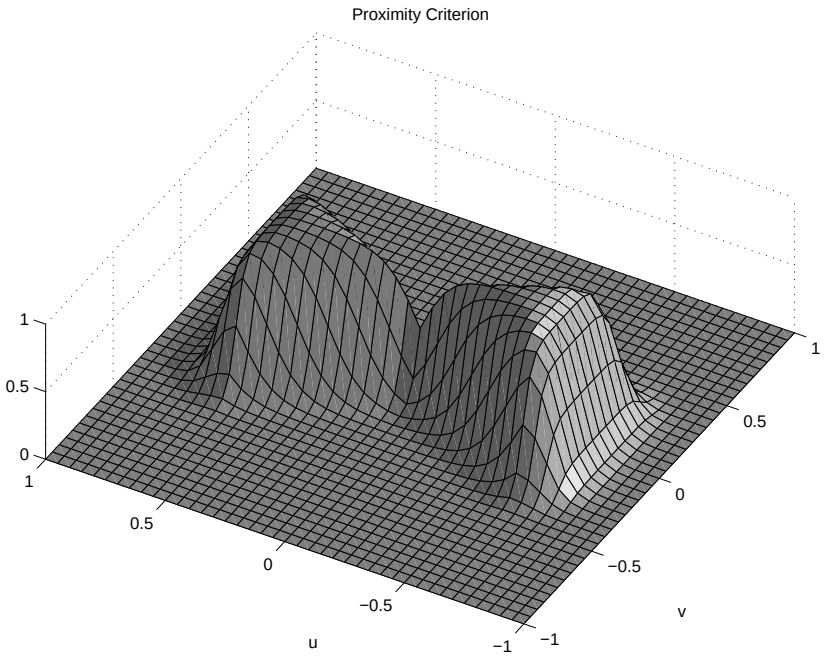


Figure 7.3: Proximity Criterion: Surface of the decrease of confidence in consistency, with the position of the second Primitive relative to the first

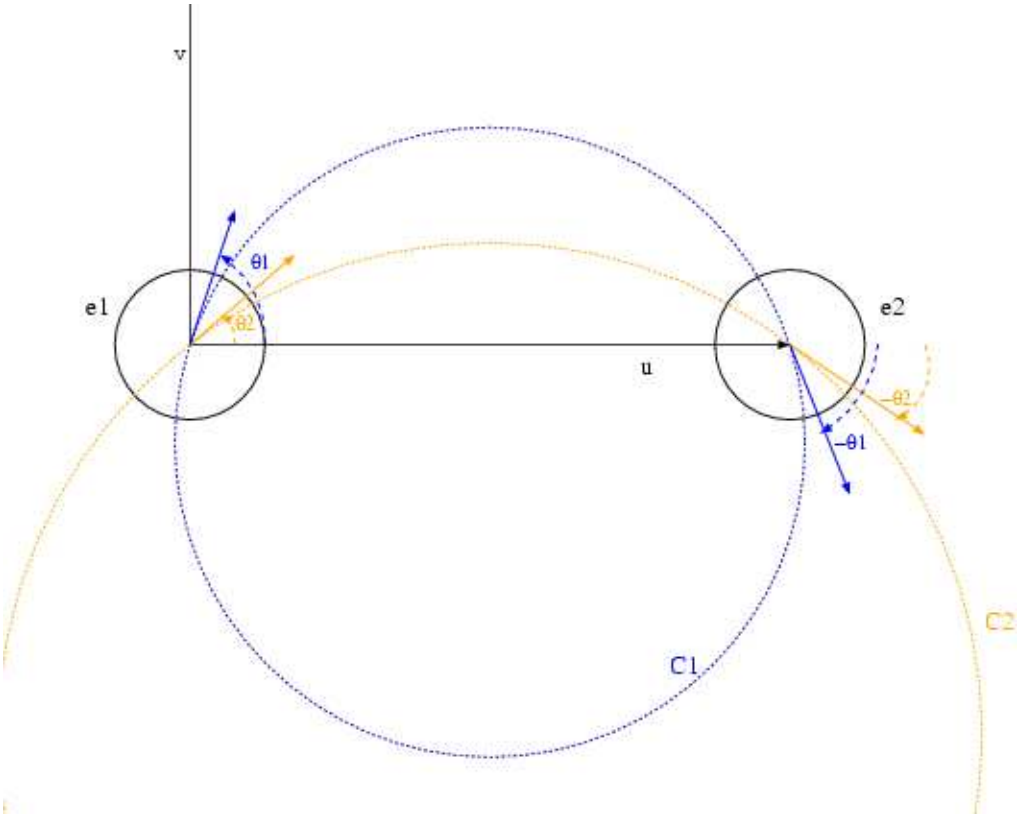


Figure 7.4: Good Continuation criterion: here we see that we can define an unique circle from the positions of $\mathbf{e}_1$ and $\mathbf{e}_2$ and the orientation of $\mathbf{e}_1$. This circle gives us an estimate for the orientation of $\mathbf{e}_2$

Good Continuation Criterion

If we consider the two modalities $\mathbf{e}_1$ and $\mathbf{e}_2$, the continuity in terms of orientation can be defined as a minimal curve joining $\mathbf{e}_1$ and $\mathbf{e}_2$. This curve ideally joins the positions A and B, and is tangent to the orientation of $\mathbf{e}_1$ and $\mathbf{e}_2$ in those points.
In the following we consider the coordinate system O, u, v (see also figure 7.4) such as:

- O being the location of the first Primitive $\mathbf{e}_1$,
- u the vector from $\mathbf{e}_1$ to $\mathbf{e}_2$
- v normal to u

The axis are normalized so that a distance of 1 is the distance between $\mathbf{e}_1$ and $\mathbf{e}_2$ in the image.
Consequently, the position of $\mathbf{e}_1$ is defined by the vector $(0, 0)$ and $\mathbf{e}_2$ by $(1, 0)$ We can define a unique circle from the positions of $\mathbf{e}_1$ and $\mathbf{e}_2$ and the orientation of $\mathbf{e}_1$. This circle gives us an estimate for the orientation of $\mathbf{e}_2$ (see figure 7.4). An estimation of the likelihood of the curve defined by the two Primitives is then the difference between this estimated orientation and the measured one (see figre 7.4).

$$v'(1) = -v'(0) = -tan(\theta_1) \tag{7.2}$$

$$C_{ori}(\mathbf{e}_1, \mathbf{e}_2) = |tan(\theta_2) + tan(\theta_1)| \tag{7.3}$$

7.2.2 Modality Continuity Criterion

The consistency over the color and phase modalities is calculated using the similarity functions for phase $sim_\phi(\mathbf{e}_1, \mathbf{e}_2)$ and colour sim_C already used in [73, 95]. Here we assume that modalities are continuous over a given 3D feature. Consequently, they should be continuous over their manifestation in the image. We say that a link $\mathcal{L}(\mathbf{e}_1, \mathbf{e}_2)$ exists when their modalities are close enough. Consequently we define an estimate for the consistency of the pair $c[\mathcal{L}(\mathbf{e}_1, \mathbf{e}_2)]$ by

$$c[\mathcal{L}(\mathbf{e}_1, \mathbf{e}_2)] = Coll(\mathbf{e}_1, \mathbf{e}_2) \times (sim_\phi(\mathbf{e}_1, \mathbf{e}_2) + sim_C(\mathbf{e}_1, \mathbf{e}_2)) \,. \tag{7.4}$$

An example of the links confidences for our test sequences can be seen in 7.5 (show for artificial and ecovision sequence)

7.3 Multi-modal stereo

To create 3D information from the 2D Primitives by stereo we need to match Primitives in the left and right image. In [73, 95]) we have derived a matching function makes use

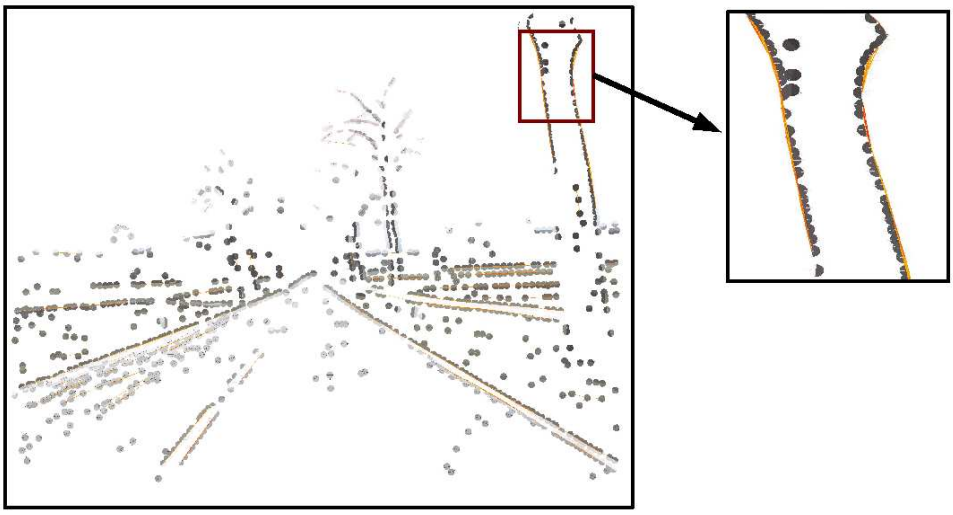


Figure 7.5: The potential links between the primitives are shown by the orange lines. The darker the line, the higher the confidence in the link.

of information in all modalities. A pair $(\mathbf{e}^l, s(\mathbf{e}^l))$ represents the correspondence found between a Primitive in the left and the right image ($s(\mathbf{e}^l)$ being the matched Primitive in the right image). From such a correspondence we can compute a 3D Primitive $\mathbf{E}$ by a reconstruction function R:

$$\mathbf{E} = R(\mathbf{e^l}, s(\mathbf{e^l})) \tag{7.5}$$

Every Primitive has a list of potential stereo-correspondences containing all Primitives of the second image intersecting the epipolar line drawn from the first Primitive. In [73, 95] only the best correspondence is used to generate the 3D-entity. The decision between several potential matches is made comparing similarities in local modality measurements of both primitives. We will call this estimation of the quality of a match the *internal confidence* and note it $c[s(\mathbf{e})]$: it is all that can be estimated using the locally available information of the Primitive.

Stereo Matching based on the internal confidence is naturally ambiguous, for example repetitive structures may occur in a scene leading to similar Primitives for distinct scene elements. Also due to projective distortion between both images the actual similarity might be misleading: for example differences in orientation and colour can be expected in both images according to the different perspective views of the left and right image. This difference of course cannot be anticipated in a local way leading to sub-optimal similarity

estimation. Consequently, the internal confidence on its own is a naturally inaccurate and ambiguous measure.

7.4 Combining Grouping and Stereo

In this paper, we want to improve the decision based of local information by taking into account the consistency over the Primitive's neighbourhood utilising the grouping process defined in section 7.2. The core idea is to compare how similar neighbourhood of the potential matches are to the neighbourhood of the original Primitive to define an *external confidence* in the match (written $c_{ext}[s(\mathbf{e})]$). The neighbourhood is here considered as the network of links associated to the Primitive.

7.4.1 Stereo-Consistency Element

We considered that consistency in Primitives was not incidental but a consequence of the scene structure and therefore this consistency should be conserved by stereo (except, of course in case of stereo occlusion). We want to define a stereo correspondence mechanism handling this external confidence based on the following principles:

Postponement of early hard Decision: Differing from (7.5) we want postpone the decision of a succesful match and allow for multiple correspondences leading to mutiple potential 3D matches. The final decision is done after the grouping process considering the Stereo Collinearity Constraint.

Uniqueness leads to Competition: As stereo correspondences are mutually exclusive competition needs to be included in any correction/adaptation process

Weighting according to Group Consistency: Over one Primitive neighbourhood, the relative weight of the stereo correspondence of a neighbour is proportional to the consistency of the Primitive with this neighbour (i.e. to the link confidence).

Weighting according to Stereo Consistency: The influence of a Primitive over its neighbours is proportional to the confidence in its stereo-correspondences (consequently a Primitive with only poor stereo correspondences will do little to help stereo decisions).

We then define that the *minimal* stereo event involving a primitive neighbourhood, is: Given two Primitives $\mathbf{e}_1^L$ and $\mathbf{e}_2^L$ in the left frame such as a link $\mathcal{L}(\mathbf{e}_1^L, \mathbf{e}_2^L)$ can be defined between them, if we consider the hypothesis that $s_i(\mathbf{e}_1^L)$ is the correct stereo-correspondence for $\mathbf{e}_1^L$ in the right image:
if exists a link $\mathcal{L}(s_i(\mathbf{e}_1^L), s(\mathbf{e}_2^L))$ between this stereo-correspondence and the public stereo-correspondence $s(\mathbf{e}_2^L)$ of the second primitive $\mathbf{e}_2^L$

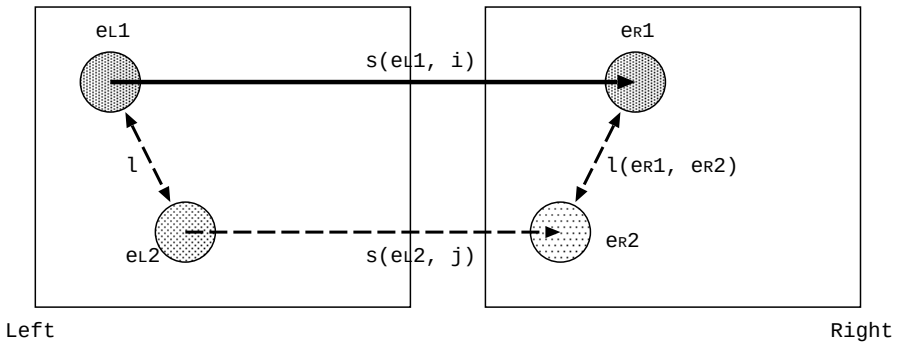


Figure 7.6: The BSCE criterion: Given a stereo correspondence $s_i(\mathbf{e}_1)$, the BSCE can be calculated for a primitive $\mathbf{e}_2$ in the neighbourhood, depending on $\mathcal{L}(\mathbf{e}_1, \mathbf{e}_2)$, $s_j(\mathbf{e}_2)$, and $l'(s_i(\mathbf{e}_1), s_)(\mathbf{e}_2 j)$. The bold line represent the event we want to confirm, and the dashed lines the external events which, in conjonction, confirms it.

then the hypothesis $s(\mathbf{e}_1^L)$ is confirmed (and also, if no corresponding link exist in the right image this hypothesis is then contradicted).
We call this trial the *Basic Stereo Consistency Event* (BSCE).

7.4.2 BSCE confidence

We want to associate a confidence to the BSCE event. Here we are not working with certainties, but with potential links and stereo-correspondences. Consequently we want a continuous formulation of the BSCE trial, giving us a confidence in its realization. We propose in this section to draw this from the previous confidences in the simple events involved.
First, we define a set of function that are used at different places:

$$\begin{array}{lcll} f^g(a_1, \ldots, a_n) & = & (a_1 \cdot \ldots \cdot a_n)^{\frac{1}{n}} & \text{Geometric Mean} \\ f^a(a_1, \ldots, a_n) & = & \frac{a_1 + \ldots + a_n}{n} & \text{Arithmetic Mean} \end{array}$$

The geometric mean represents a harder connection between events than the arithmetic mean. The multiplication works like a "logical and" ($\wedge$) while the arithmetic mean is a softer connection. We apply the arithmetic mean when different cues co-operate "democratically" while the geometric mean is used when the non-occurence of one event supresses all others.
Now the confidence associated to a BSCE can be estimated from the known confidences as follows:

$$c[BSCE_i(\mathbf{e}_1^L, \mathbf{e}_2^L)] = f^g\left(c[\mathcal{L}(\mathbf{e}_1^L, \mathbf{e}_2^L)], c[s(\mathbf{e}_2^L)], c[\mathcal{L}(s_i(\mathbf{e}_1^L), s(\mathbf{e}_2^L))]\right) \tag{7.6}$$

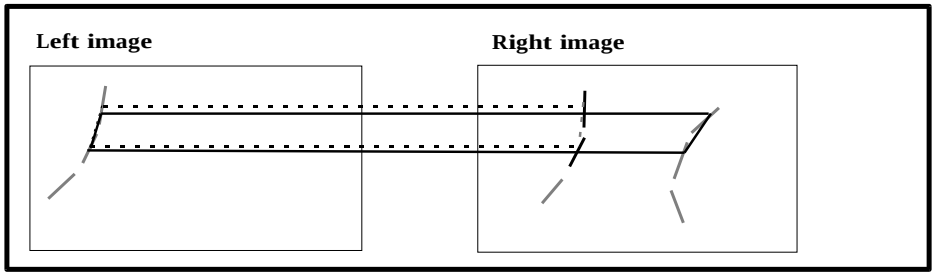


Figure 7.7: Since one Primitive can have multiple matches it can be verified by multiple BSCEs. The dotted and bold structures each represent one BSCE.

7.4.3 Neighbourhood Consistency Confidence

This formula gives us how a Primitive stereo correspondence is consistent with our beliefs on another Primitive stereo properties. We now want to estimate how this correspondence is consistent with the *whole neighbourhood* of the Primitive. Now if we consider a primitive $\mathbf{e}_1^L$ and an associated stereo-correspondence $s_i(\mathbf{e}_1^L)$, we can integrate this BSCE confidence over the neighbourhood of the primitive ($N_{\mathbf{e}_1^L}$). We call this confidence the *external confidence* in the stereo-correspondence:

$$c_{ext}[s_i(\mathbf{e}_1^L)] = \frac{1}{|N_{\mathbf{e}_1^L}|} \sum_{\mathbf{e}_k^L \in N_{\mathbf{e}_1^L}} c[BSCE_i(\mathbf{e}_1^L, \mathbf{e}_k^L)] \tag{7.7}$$

This gives us a confidence on how consistent is a stereo-correspondence with the stereo of the primitive neighbourhood.

7.4.4 Outlier Removal Process

In the outlier removal process we are after the reliable matches (i.e., we want to eliminate possibly false matches). The outlier removal process can be used where a small number of reliable features is used to compute the motion between frames (in this case we need reliable 3D-2D matches).
Our actual system rank the potential correspondences of a primitive depending on their similarity (over all modalities) with this primitive, and the best one (or public one) is assumed to be the correct correspondence. We propose here to threshold the external confidence of those potential correspondences in order to remove those in contradiction with their neighbourhood current assumptions (i.e. neighbours public correspondences).

We expect this way to remove number of wrong correspondences, otherwise impossible to discern from correct ones using local modalities.
Figures **??** show the result of the outlier removal process.

7.5 Results

We have applied this outlier removal process to two stereo sequences. The first one (fig. 7.8) is a simple artificial scene generated using openGL. The second scene has been recorded near Lippstadt (Germany) from a pair of calibrated cameras fixed to a car (with the cooperation of HELLA). This second scene (fig. 7.9) represent more accurately the standard conditions in which a natural system has to operate (low saturation, highly textured surfaces, etc...). Both figures show the left and right images on the top row. On the middle row, the images show the primitives extracted. The red lines reach to the position of their current public correspondence in the right image. Those pairs (from the public correspondences of each primitive) are used to reconstruct 3D entities. The lower figure show a reprojection of those 3D entities on the horizontal plane (the horizontal axis is the Z axis, and the vertical axis is the X axis here). The left two pictures show the original public correspondences and reconstruction using only internal confidence. The two pictures on the right of the figures show the public correspondences and the resulting 3D reconstructed entities after a thresholding of the correspondences over their external confidence (the threshold is of 0.075 in both cases).
On the figure 7.8 most of the wrong correspondences are being removed through this process. More interestingly on the figure 7.9, showing the difficult natural scene, a considerable amount of noise is being removed. In the magnified view of the correspondences, we can see that most random correspondences from primitives extracted from texture artifact are being removed, while consistent correspondences are preserved. Note that in both case this improvement is gained *only by thresholding the external confidence*, and without any additional thresholding on the actual similarity of the primitives.

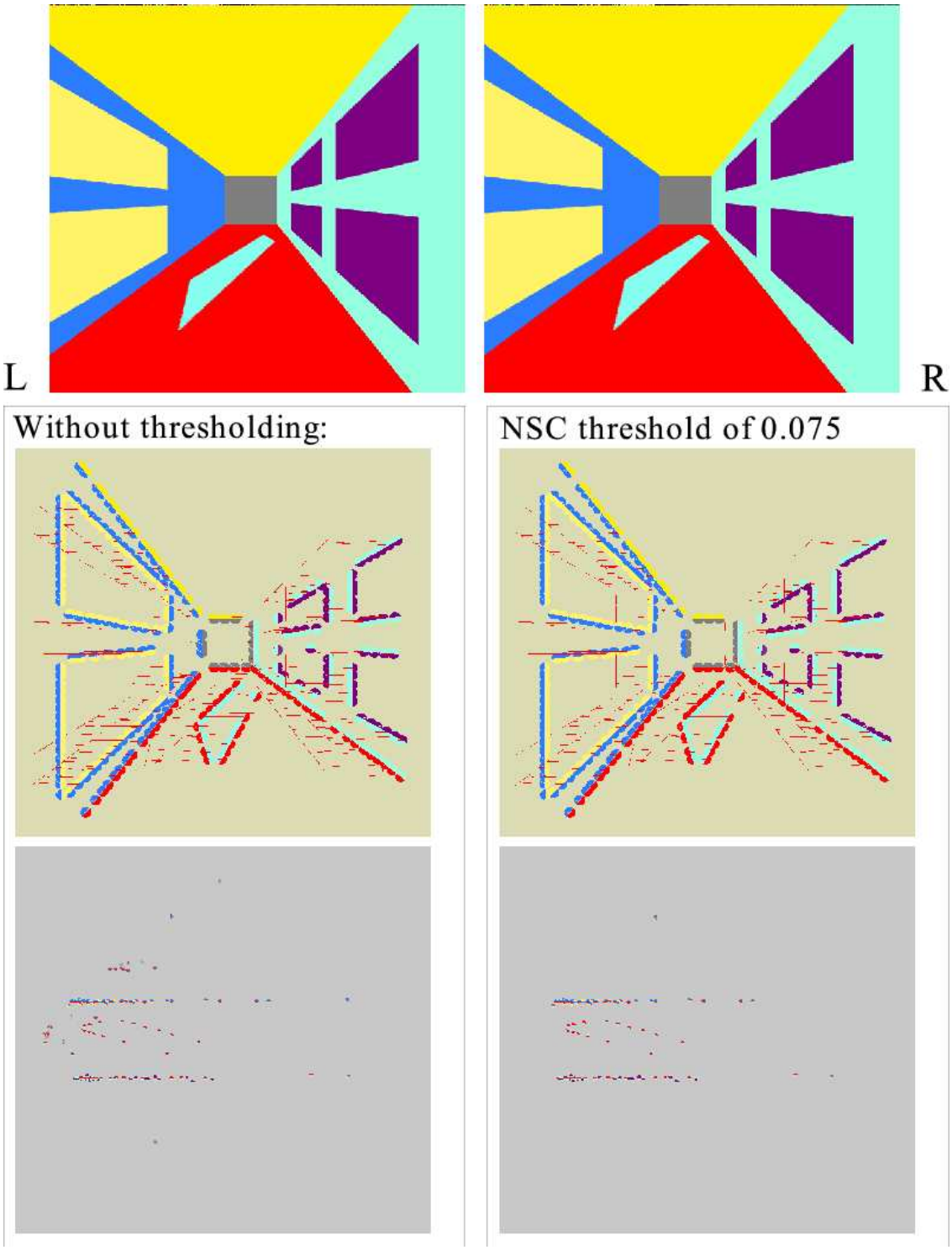


Figure 7.8: We apply our external confidence thresholding to this artificial scene. The left two images represent the results without thresholding, and the right ones with thresholding. In both case, the middle image show the primitives extracted by our program, and the lines reach to the position of their current public correspondence in the right image. The lower one show a orthographic reprojection of the reconstructed 3D entities (from those public stereo pairs). This shows the XZ (horizontal) plane.

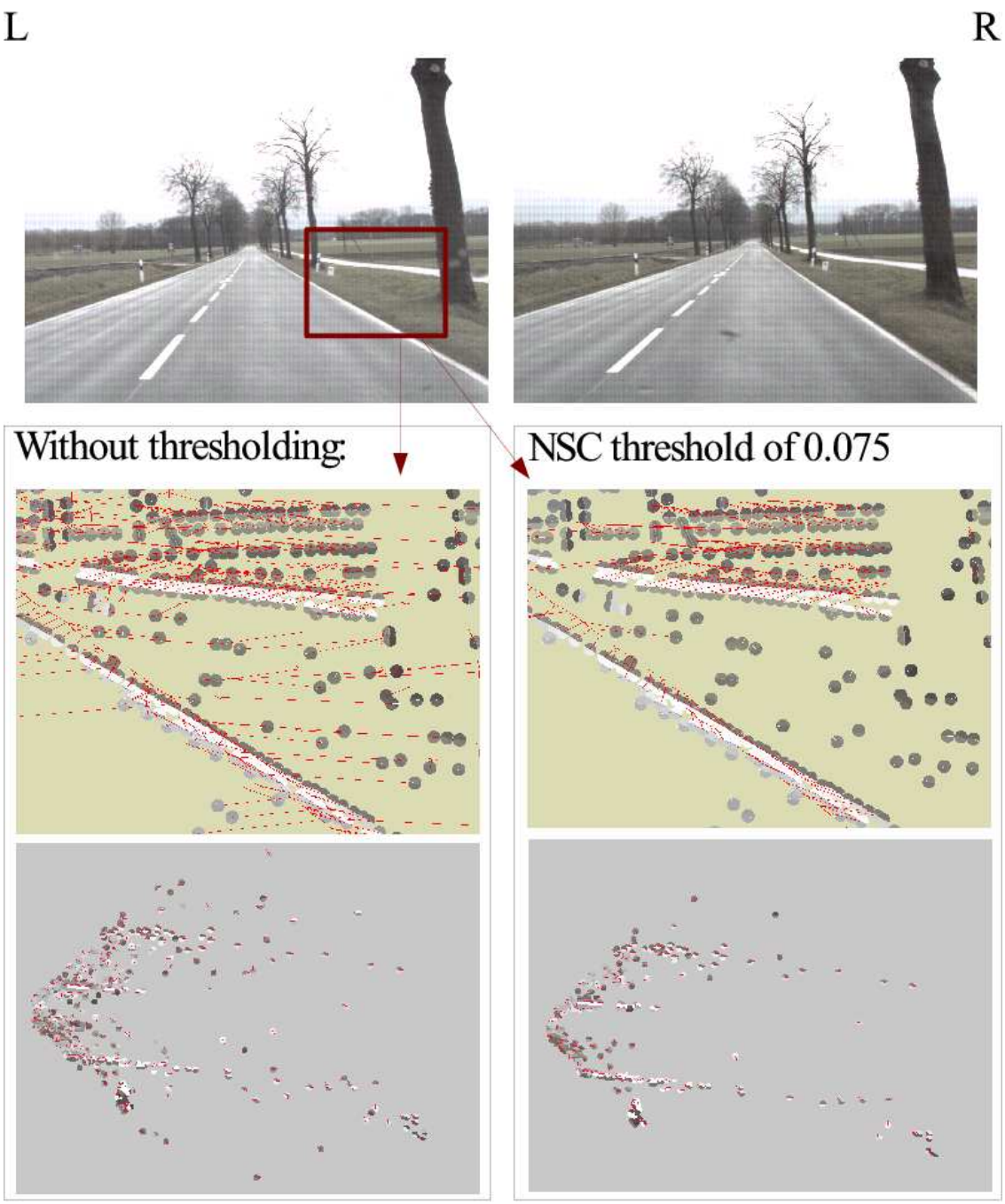


Figure 7.9: This figure show the same results as 7.8 with a natural stereo scene this time. The middle images show here a zoom in of the primitives. Here the correspondences of the primitives created by the texture are being removed, while consistent lines are being preserved.

Bibliography

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

[2] H. Araujo, R.J. Carceroni, and C.M. Brown. A fully projective formulation to improve the accuracy of lowe's pose–estimation algorithm. *Computer Vision and Image Understanding*, 70(2):227–238, 1998.

[3] D. Attwell and S.B. Laughlin. An energy budget for signalling in the grey matter of the brain. *Journal of Cerebral Bloodflow and Metabolism*, 21:1133–1145, 2001.

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

[5] R.S. Ball. *The theory of screws*. Cambridge University Press, 1900.

[6] H. Barlow, C. Blakemore, and J.D. Pettigrew. The neural mechanisms of binocular depth discrimination. *Journal of Physiology (London)*, 193:327–342, 1967.

[7] H. Bårman, G. H. Granlund, and H. Knutsson. Tensor Field Filtering and Curvature Estimation. In *Proceedings of the SSAB Symposium on Image Analysis*, pages 175–178, Linköping, Sweden, March 1990. SSAB. Report LiTH-ISY-I-1088, Linköping University, Sweden, 1990.

[8] 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.

[9] J.R. Beveridge. Local search algorithms for geometric object recognition: Optimal correspondence and pose. *PhD Thesis, University of Massachusetts at Amherst, available as Technical Report CS 93-5*, 1993.

[10] J. Bigün and G. H. Granlund. Optimal orientation detection of linear symmetry. In *Proceedings of the IEEE First International Conference on Computer Vision*, pages 433–438, London, Great Britain, June 1987. Report LiTH-ISY-I-0828, Computer Vision Laboratory, Linköping University, Sweden, 1986.

[11] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1995.

[12] W. Blaschke. *Kinematik und Quaternionen*. VEB Deutscher Verlag der Wissenschaften, 1960.

[13] K.L. Boyer and S. Sarkar. Perceptual organization in computer vision: Status, challenges, and potential. *Special Issue on Perceptual Organization in Computer Vision, October*, 76(1):1–5, 1999.

[14] R. N. Bracewell. *The Fourier transform and its applications*. McGraw Hill, 1986.

[15] C. Bregler and J. Malik. Tracking people with twists and exponential maps. *IEEE computer Society conference on Computer Vision and Pattern Recognition*, pages pp.8–15, 1998.

[16] A.R. Bruss and B.K.P. Horn. Passive navigation. *Computer Vision, Graphics, and Image Processing*, 21:3–20, 1983.

[17] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6), 1986.

[18] 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.

[19] I. Cox, S. Hingoraini, and S. Rao. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63:542–567, 1996.

[20] H.S.M. Coxeter. *Introduction to Geometry (2nd ed.)*. Wiley & Sons, 1969.

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

[22] S. C. Dakin. Local orientation variance as a quantifier of structure in texture. *Spatial Vision*, 12:1–30, 1999.

[23] B. Girod E. Steinbach. An image-domain cost function for robust 3-d rigid body motion estimation. *15th International Conference on Pattern Recognition (ICPR-2000)*, 3:823–826, 2000.

[24] ECOVISION. Artificial visual systems based on early-cognitive cortical processing (EU–Project). *http://www.pspc.dibe.unige.it/ecovision/project.html*, 2003.

[25] P. Eisert and B. Girod. Illumination compensated motion estimation for analysis sythesis coding. *3D Image Analysis and Sythesis*, pages 61–66, 1996.

[26] C. Fagerer, D. Dickmanns, and E.D. Dickmanss. Visual grasping with long delay time of a free floating object in orbit. *Autonomous Robots*, 1(1):53–68, 1991.

[27] G. Farnebäck. Fast and accurate motion estimation using orientation tensors and parametric motion models. *Proc. ICPR*, 2000.

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

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

[30] M. Felsberg. *Low-Level Image Processing with the Structure Multivector*. PhD thesis, Institute of Computer Science and Applied Mathematics, Christian-Albrechts-University of Kiel, 2002.

[31] M. Felsberg and N. Krüger. A probablistic definition of intrinsic dimensionality for images. *Pattern Recognition, 24th DAGM Symposium*, 2003.

[32] M. Felsberg and G. Sommer. Image features based on a new approach to 2D rotation invariant quadrature filters. Proc. of ECCV, pages 369–383, 2000.

[33] M. Felsberg and G. Sommer. The monogenic signal. *IEEE Transactions on Signal Processing*, 49(12):3136–3144, December 2001.

[34] R. Fischler and M. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):619–638, 1981.

[35] W. Förstner. *Statistische Verfahren für die automatische Bildanalyse und ihre Bewertung bei der Objekterkennung und -vermessung*. Number 370 in C. Verlag der Bayerischen Akademie der Wissenschaften, 1991.

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

[37] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *ISPRS Intercommission Workshop, Interlaken*, pages 149–155, June 1987.

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

[39] G. Gimel'farb and U. Lipowezky. Accuracy of the regularised dynamic programming stereo. In *ICPR02*, pages III: 619–622, 2002.

[40] O. Granert. Posenschättzung kinematischer ketten. *Diploma Thesis, Universität Kiel*, 2002.

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

[42] W.E.L. Grimson, editor. *Object Recognition by Computer*. The MIT Press, Cambridge, MA, 1990.

[43] C. G. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conference*, pages 147–151, 1988.

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

[45] D. Hesteness and G. Sobczyk. *Clifford Algebra to Geometric Calculus*. D. Reidel Public. Comp., Dordrecht, 1984.

[46] P.B. Hibbard, M.F. Bradshaw, and R.A. Eagle. Cue combination in the motion correspondence problem. *Proceedimgs of the Royal Society London B*, 267:1369–1374, 2000.

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

[48] H.H. Homer. Pose determination ftom line-to-plane correspondences: Existence condition and closed form solutions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):530–541, 1991.

[49] B.K.P. Horn, editor. *Robot Vision*. MIT Press, 1994.

[50] P.V.C. Hough. Methods and means for recognizing complex patterns. *U.S. Patent 3,069,654, Dec. 18*, 1962.

[51] http://www.photomodeler.com. 2000.

[52] D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Phyiology*, 160:106–154, 1962.

[53] D.H. Hubel and T.N. Wiesel. Anatomical demonstration of columns in the monkey striate cortex. *Nature*, 221:747–750, 1969.

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

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

[56] Selig J.M. Some remarks on the statistics of pose estimation. *Technical Report SBU-CISM-00-25, South Bank University, London*, 2000.

[57] J.P. Jones and L.A. Palmer. An evaluation of the two dimensional Gabor filter model of simple receptive fields in striate cortex. *Journal of Neurophysiology*, 58(6):1223–1258, 1987.

[58] 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.

[59] G. Kanizsa. Subjective contours. *Scientific American*, 1976.

[60] K. Klein. *Vorlesungen über nicht–Euklidische Geometrie*. AMS Chelsea, 1927.

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

[62] H. Klingspohr, T. Block, and R.-R. Grigat. A passive real–time gaze estimation system for human-machine interfaces. *CAIP Proceedings, LNCS 1298*, pages 718–725, 1997.

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

[64] J.J Koenderink. What is a feature? *J. Intell. Syst.*, 3(1):49–82, 1993.

[65] A. Koschan. Chromatic block matching for dense stereo correpondence. *Proceedings of ICIAP*, 1993.

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

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

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

[69] N. Krüger. Collinearity and parallelism are statistically significant second order relations of complex cell responses. *Proceedings of I&ANN 98*, 1998.

[70] 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.

[71] N. Krüger and M. Felsberg. An explicit and compact coding of geometric and structural information applied to stereo matching. *submitted to Pattern Recognition.*

[72] N. Krüger and M. Felsberg. A continuous formulation of intrinsic dimension. *Proceedings of the British Machine Vision Conference*, 2003.

[73] 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.

[74] N. Krüger, M. Felsberg, and F. Wörgötter. Processing multi-modal primitives from image sequences. *Fourth International ICSC Symposium on ENGINEERING OF INTELLIGENT SYSTEMS*, 2004.

[75] 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*, pages 92–100, 2002.

[76] N. Krüger, M. Lappe, and F. Wörgötter. Biologically motivated multi-modal processing of visual primitives. *Proceedings of the AISB 2003 Symposium on Biologically inspired Machine Vision, Theory and Application, Wales*, pages 53–59, 2003.

[77] N. Krüger, M. Lappe, and F. Wörgötter. Biologically motivated multi-modal processing of visual primitives. *The Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour*, 1(5), 2004.

[78] N. Krüger and B. Rosenhahn. Uncertainty and RBM–estimation. *in progress.*

[79] N. Krüger and F. Wörgötter. Multi modal estimation of collinearity and parallelism in natural image sequences. *Network: Computation in Neural Systems*, 13:553–576, 2002.

[80] N. Krüger and F. Wörgötter. Statistical and deterministic regularities: Utilisation of motion and grouping in biological and artificial visual systems. *Advances in Imaging and Electron Physics*, 131, 2004.

[81] V. Krüger and G. Sommer. Wavelet networks for face processing. *JOSA*, 19:1112–1119, 2002.

[82] D.G. Lowe. Three–dimensional object recognition from single two images. *Artificial Intelligence*, 31(3):355–395, 1987.

[83] D.G. Lowe. Fitting parametrized 3D–models to images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):441–450, 1991.

[84] ModIP. Modality Integration Project. *www.cn.stir.ac.uk/ComputerVision/Projects/ModIP/index.html*, 2003.

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

[86] Murray, Li, and Sastry. *A mathematical introduction to Robotic Manipulation*. CRC Press, 1994.

[87] H.-H. Nagel. On the estimation of optic flow: Relations between different approaches and some new results. *Artificial Intelligence*, 33:299–324, 1987.

[88] S. Negahdaripour and B.K.P. Horn. Direct passive navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):168–176, 1987.

[89] C. C. Pack and R. T. Born. Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature*, 409:1040–1042, 2001.

[90] P. Parent and S.W. Zucker. Trace inference, curvature consistency, and curve detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):823–839, 1989.

[91] A.J. Parker and B.G. Cumming. Cortical mechanisms of binocular stereoscopic vision. *Prog Brain Res*, 134:205–16, 2001.

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

[93] 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.

[94] 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.

[95] N. Pugeault and N. Krüger. Multi–modal matching applied to stereo. *Proceedings of the BMVC 2003*, 2003.

[96] N. Pugeault, F. Wörgötter, and N. Krüger. A non-local stereo similarity based on collinear groups. *Fourth International ICSC Symposium on ENGINEERING OF INTELLIGENT SYSTEMS*, 2004.

[97] E. Ribeiro and E.R. Hancock. Shape from periodic texture using the eigenvectors of local affine distortion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1459–1465, 2001.

[98] M. Rioux, F. Blais, and J. A. Beraldin. Laser range finder development for 3D vision. *Vision Interface '89, London, Ont.*, pages 1–9, 1989.

[99] J.W. Roach and J.K. Aggarwall. Determining the movement of objects from a sequence of images. *IEEE Transactions on Patterm Analysis and Machine Intelligence*, 2(6):554–562, 1980.

[100] K. Rohr. Recognizing corners by fitting parametric models. *International Journal of Computer Vision*, 9(3):213–230, 1992.

[101] B. Rosenhahn. *Pose Estimation Revisited (PhD Thesis)*. Institut für Informatik und praktische Mathematik, Chrsitian–Albrechts–Universität Kiel, 2003.

[102] B. Rosenhahn, O. Granert, and G. Sommer. Monocular pose estimation of kinematic chains. In L. Dorst, C. Doran, and J. Lasenby, editors, *Applied Geometric Algebras for Computer Science and Engineering*, pages 373–383. Birkhäuser Verlag, 2001.

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

[104] B. Rosenhahn and G. Sommer. Adaptive pose estimation for different corresponding entities. In L. van Gool, editor, *Pattern Recognition, 24th DAGM Symposium*, pages 265–273. Springer Verlag, 2002.

[105] S.J. Sangwine and R.E.N. Horne. *The Colour Image Processing Handbook*. Chapman & Hall, 1998.

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

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

[108] I.A. Shevelev, N.A. Lazareva, A.S. Tikhomirov, and G.A. Sharev. Sensitivity to cross–like figures in the cat striate neurons. *Neuroscience*, 61:965–973, 1995.

[109] F. Shevlin. Analysis of orientation problems using Plücker lines. *International Conference on Pattern Recognition, Bisbane*, 1:65–689, 1998.

[110] E. Steinbach. *Data driven 3-D Rigid Body Motion and Structure Estimation*. Shaker Verlag, 2000.

[111] P. Stumpf. über die Abhängigkeit der visuellen Bewegungsrichtung und negativen Nachbildes von den Reizvorgangen auf der Netzhaut. *Zeitschrift fur Psychologie*, 59:321–330, 1911.

[112] K. Tanaka. Neuronal mechanisms of object recognition. *Science*, 262:685–688, 1993.

[113] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, London, 1999.

[114] A. Thiele, K.R. Dobkins, and T.D. Albright. The contribution of color to motion processing in macaque area mt. *J. Neurosci.*, 19:6571–6587, 1999.

[115] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[116] G. V. Trunk. Representation and analysis of signals: statistical estimation of intrinsic dimensionality and parameter identification. *General System*, 13:49–76, 1968.

[117] S. Ullman. The interpretation of structure from motion. In *MIT AI Memo*, 1976.

[118] R. von der Heydt, E. Peterhans, and G. Baumgartner. Illusory contours and cortical neuron responses. *Science*, 224:1260–62, 1984.

[119] R.J. Watt and W.A. Phillips. The function of dynamic grouping in vision. *Trends in Cognitive Sciences*, 4(12):447–154, 2000.

[120] A.M. Waxman and S. Ullman. Surface structure and 3-D motion from image flow: A kinematic analysis. *International Fournal of Robot Research*, 4(3):72–94, 1985.

[121] J. Weng, N. Ahuja, and Huang T.S. Optimal motion and structure estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):864–884, 1993.

[122] R.H. Wurtz and E.R. Kandel. Central visual pathways. In E.R. Kandell, J.H. Schwartz, and T.M. Messel, editors, *Principles of Neural Science (4th edition)*, pages 523–547. 2000.

[123] R.H. Wurtz and E.R. Kandel. Perception of motion, depth and form. In E.R. Kandell, J.H. Schwartz, and T.M. Messel, editors, *Principles of Neural Science (4th edition)*, pages 548–571. 2000.

[124] Hel-Or Y. and Werman M. Pose estimation by fusing noisy data of different dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 17(2), 1995.

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