Statistical and Deterministic Regularities: Utilisation of Motion and Grouping in Biological and Artificial Visual Systems

Norbert Krüger and Florentin Wörgötter

Preliminary incomplete Version
Please do not distribute !!!!

to be published officially in Spring 2004

July 25, 2003

Contents

1 Introduction **3**

2 The Problem of Vagueness and Uncertainty in Vision **10**

3 Regularities in Visual Data **12**
3.1 Why does Vision need Regularities 13
3.2 Statistical and Deterministic Regularities in Visual Data 15
3.2.1 Statistical Regularities 16
3.2.2 Deterministic Regularities 17
3.3 Utilisation of Statistical and Deterministic Regularities in Biological Systems 18
3.3.1 Evidence for a large Degree of Genetic Determination of Deterministic Regularities 19
3.3.2 Evidence for an Adaptive Component in the Ontogenesis of Abilities connected to Statistical Regularities 20
3.4 Computational Differences between Statistical and Deterministic Regularities 20
3.5 Consequences for the Design of Artificial Visual Systems 23

4 Formalisation, Estimation and Application of Rigid Body Motion 23
4.1 The deterministic Regularity Rigid Body Motion applied in Stereo Processing 24
4.1.1 The projective Map 24
4.1.2 The Correspondence Problem in Stereo 24
4.2 The RBM Estimation Problem 26
4.3 Classification of Methods and Situations 30
4.3.1 Different types of Methods 30
4.3.2 Different Types of Situations 31
4.4 Using Different kinds of Entities 32
4.4.1 Entities of different Dimension 32
4.4.2 Entities of different Complexity 35
4.5 The Correspondence Problem 36
4.6 RBM Estimation and Grouping 38
4.7 Mathematical Formulation of the RBM Estimation Problem 40
4.7.1 Different kind of Optimisation Algorithms 40
4.7.2 Mathematical Formalisations of Rigid Body Motion 41
4.7.3 Parametrisation of Visual Entities 44
4.7.4 Constraint Equations 46
4.8 Properties of Rosenhahn et al's RBM estimation algorithm 49

5 The Utilisation of Statistical Regularities in Artificial Visual Systems 50
5.1 Basic Entities in perceptual Organisation 52
5.2 Approaches to ground Gestalt Laws 53
5.3 Relating Gestalt principles to the statistics of natural images 55

6 Combination of Statistical and Deterministic Regularities 56

7 Conclusion 58

1 Introduction

Vision–based devices have been entering the industrial and private world more and more successfully: Face recognition systems control the access to buildings (see, e.g., [134]); the quality of goods is controlled by visual inspection (see, e.g., [23]); airports and train stations are controlled by Video Surveillance devices (see, e.g., [110]); and cars become equipped with vision–based driver assistance systems (see, e.g., [22]). There exists even attempts to build human like–robots [52]. However, the gap between human performance and the top performance of today's artificial visual systems is considerable (which is probably an understatement). Especially, scene analysis in unfamiliar environments leading to highly reliable actions is an outstanding quality of biological systems. The easiness with which we are able to navigate in an unfamiliar building or to grasp an unknown object may lead us to the conclusion that the underlying algorithmic problems are 'easily' solvable as well, especially compared to some 'hard' tasks such as, e.g., playing chess. However, today's computers can compete with and will probably soon be better than today's chess masters [126]. However, up to now there exists no robot that could grasp a cup from a table, fill it with coffee and hand it over to Ann or Paul. These 'easy' problems are apparently much harder to model than the 'hard' chess task.

The underlying problems (recognising the cup, getting some idea of its position and orientation, recognising the pot of coffee, recognising Ann or Paul, ...) are far away from being solved. In this paper, we will discuss one of the main reasons for this 'failure' of technological ambition: It is the the *ambiguity of local visual information*. In a nutshell, most visual systems (see, e.g., [98, 103, 109, 94, 82, 139, 21, 81, 73]), and in particular the human visual system (see ,e.g., [55]), analyse in its first stages *local image areas*. However, when analysing local image areas our categorisation of structure is necessarily erroneous. For example, when we look at some local image patches in figure 1 (left) it is quite hard to say to what kind of lines or junctions they correspond. Moreover, there is no way to get an idea about the underlying 3D structure. However, taking the global context into account (see figure 1 (right)) we have a clear idea about the local line/junction as well as the 3D structure. In section 2, we will show that this local ambiguity is a property of the local signal structure that is reflected in different sub–visual structures (in the following also called visual modalities) such as, colour, local motion, binocular vision etc.

Each visual system faces the problem of ambiguity. However, the existence of biological systems that are able to act with high reliability in complex environments gives us the hope and trust that this problem is solvable. On the other hand, since one third of your brain deals with vision there is some evidence that the algorithmic problems involved are not 'easy' at all but of considerable complexity as will become obvious in this paper.

The development of computer vision during the last decades was dominated by Marr's idea of preprocessing of localised image structure descriptors [86]. Descriptors representing

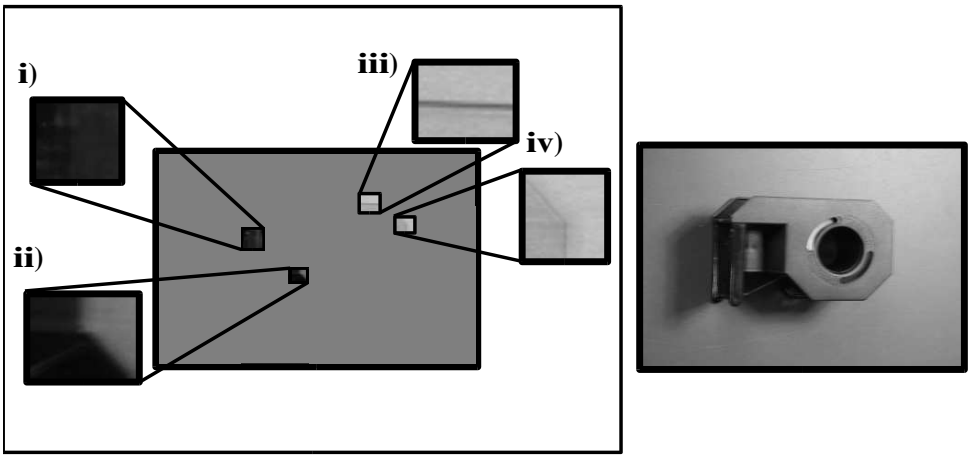

Figure 1: Ambiguity Edge and Junction Detection. Left: Local edge and junction structures in an image. A detection and classification is difficult. For example, the image patch i) consists basically of noise only. Image patch ii) may be interpreted as a junction, but whether this junction consists of 2 or 3 intersecting lines (i.e., whether it is a L–junction or Y–junction) remains unclear. Image patch iii) is interpretable reliably as a local line. However, that this line actually separates one surface from the other (i.e., represents a depth discontinuity) remains unclear. Image patch iv) we could interpret as a Y–junction. However, a more global perspective would lead to the interpretation as a L–junction. Right: Taking the global context into account a semantic description (even in terms of the underlying 3D structure) is easy.

structure in different modalities (such as, edge detection, colour, binocular vision, local motion (optic flow)) have therefore been one main focus of vision research. This has led to a better understanding of visual modalities and to the development of efficient feature extraction algorithms in the different domains that are now successfully used in applications in *controlled* environments. However, progress in this field has still not led to systems that work with the same efficiency and reliability as the human visual system. There seems to exist a '90% performance ceiling' which can not be transcended (at least under uncontrolled conditions, especially variation of illumination in outdoor scenes). From this impairement, some scientists have concluded that Marr's approach is fundamentally wrong (see, e.g., [16]).

However, here we claim that Marr's ideas can be reformulated in a way, that allows for transcending the '90% ceiling'. We claim that, although locally ambiguous, visual data is *dominated by regularities* that allow for disambiguating locally erroneous statements. These regularities allow for an understanding of vision as a *process of recurrent predictions* that initiate a process that disambiguates the locally erroneous interpretations. In this paper two main kinds of regularities in visual data are discussed that have been widely (but mostly independently) used in artificial systems for disambiguation of visual infor-

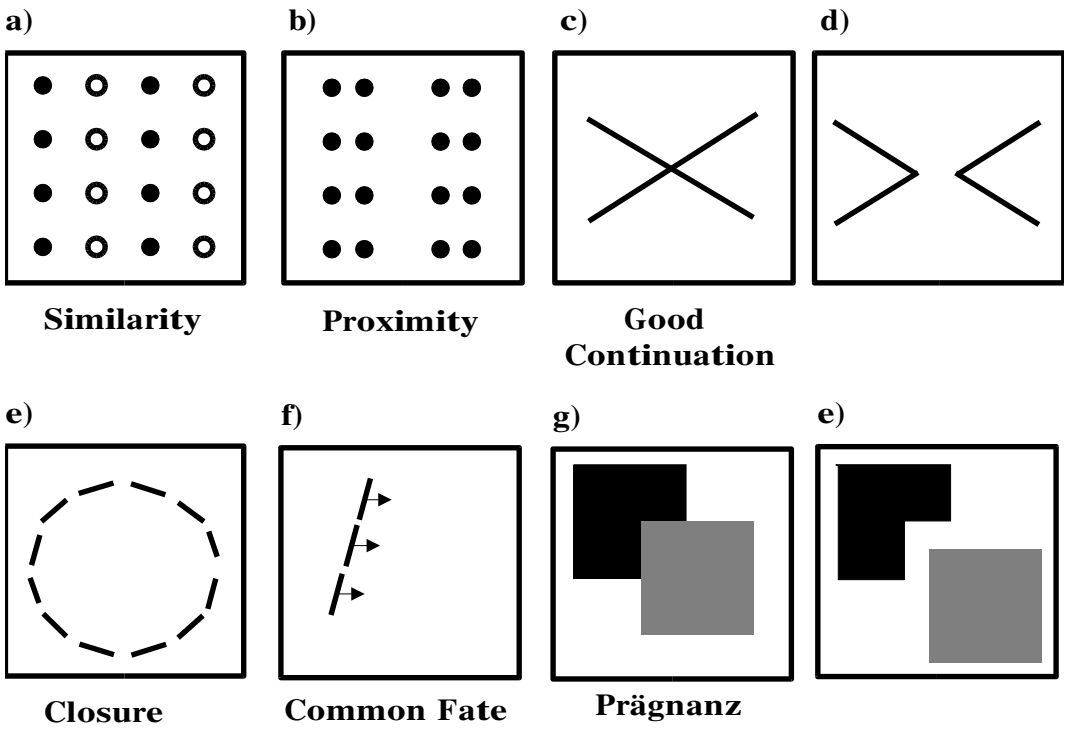

Figure 2: Six examples of classical Gestalt Laws: a) Similarity: Entities with similiar attributes are grouped together. b) Proximity: Entities with close distance are grouped together. c) Good continuation: We tend to see two intersecting lines. d) However, an interpretation as two wedges pointing at each other is possible as well but is less likely. e) Closure: Gaps are filled to perceive rather a whole than isolated units. f) Common Fate: Coherently moving objects are perceived as a whole. g) Prägnanz: The most simple interpration of two squares (one partly covering the other) is preferred although more complex but less likely constellations are possible (see, e).

mation: (1) The utilisation of the deterministic regularity 'Rigid Body Motion'. (2) The utilisation of statistical regularities on which most of the classical 'Gestalt laws' [138, 69] and 'Grouping processes' are based upon (see figure 2). The central idea of this review is to establish a ground for the joint use of both regularities. The great potential of such an approach has been stressed by Sudeep Sarkar and Kim L. Boyer [120, 14] and is also one focus of our current research.

This has been the main motivation for establishing an artificial vision system in which different modalities are represented, co–operate and stabilize (see [74, 76, 77, 80, 78, 75, 107]) that has been started in 1998 [87]. Figure 3 gives you an idea about some aspects of the multi–modal processing while in figure 4 you see an example of the resulting image representations.

The attempt of building artificial visual systems necessarily touches different scientific dis-

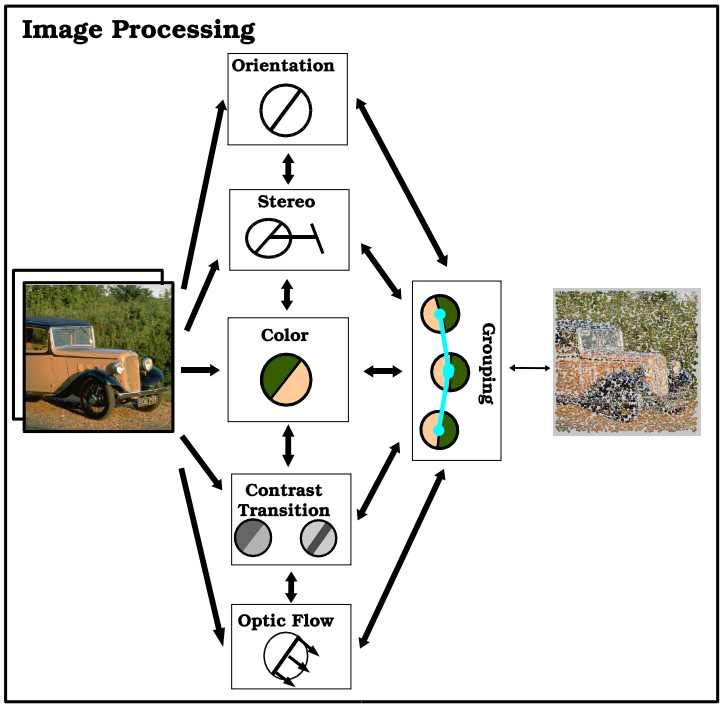

Figure 3: Different stages of image processing: First basic features in different domains (orientation, color, contrast transition, optic flow, stereo) are being processed which are then grouped. Feature processing in the different domains as well as grouping is closely intertwined. (Image from the OSU/SAMPL Database)

ciplines. Therefore, vision is essentially an interdisciplinary field. Since an efficient visual system needs to have sub–modules (or their correlate of sub–areas in the human brain) that interact with each other, we have to take an *engineering perspective*. *Mathematics and Signal Theory* gives us the framework to process visual modalities and regularities in visual scenes. *Computer Science* gives us important tools in form of today's computers and software with which we can realize artificial systems. *Biology and Psychology* allow us the study a successful system and may help us in important design decisions. Since, we think that vision can most successfully been addressed by an interdisciplinary approach, this review has a broad perspective and we hope that scientists across disciplines become interested in it. Mathematics are to a certain extent part of it but is organized such it can be skipped without loosing over all understanding.
This paper is structured as follows:

Figure 4: a) An example of our multi–modal image representations. Here optic flow and stereo are not represented.

- In section 2 we give examples of the ambiguities of local visual data in different visual domains such as line extraction, colour, stereo, local motion etc.

- We claim that this ambiguity can be resolved by the integration of information based on regularities in visual data (section 3.1). This integration takes place across the temporal and spatial domain, but also across different visual modalities.

- In section 3.2, we characterise two different kinds of regularities (statistical and deterministic ones) that yet have mostly been used separately in artificial visual systems.

- Regarding results of neurophysiology and developmental psychology we show in section 3.3 that abilities based on deterministic regularities are to a wide degree genetically determined while abilities based on statistical regularities are to a considerable degree learned.

- We will discuss computational differences of these two kind of regularities in section 3.4. We claim that these differences lead to different requirements for their formalization. Both regularities have been applied with different success within artificial visual systems: While the power of deterministic regularities has been formalised with great success within the last two decades (see, e.g., [31, 44]), the full potential of statistical regularities has not yet been employed at all [14].

- From the different amount of genetical pre-coding, we draw conclusions for the design of artificial systems in section 3.5: We argue that a basic concept of a 3D space has to be part of the pre-wired structure of a successful system with similar complexity than the human visual system. However, the formalisation of classical Gestalt laws based on statistical regularities has to be based on learning.

- We discuss the formalisation and the underlying problems of the deterministic regularity 'Rigid Body motion' (RBM) in section 4. We will especially focus on the estimation of the motion between two frames. We will specify and discuss problems of motion estimation by describing a specific algorithm [116, 115, 117, 38, 114] to some detail. This algorithm is also used in our system. Once the RBM is known, we have a strong spatial–temporal relation across frames that allows for a large amount of predictions of feature events across time frames.

- The correspondence problem is essential for motion estimation: To estimate the RBM we need to know correspondences of features in different time frames. This leads to a tough combinatorical problem since the number of possible correspondences grows exponentially with the number of features extracted. We will argue that statistical regularities can help to overcome this problem, especially in complex

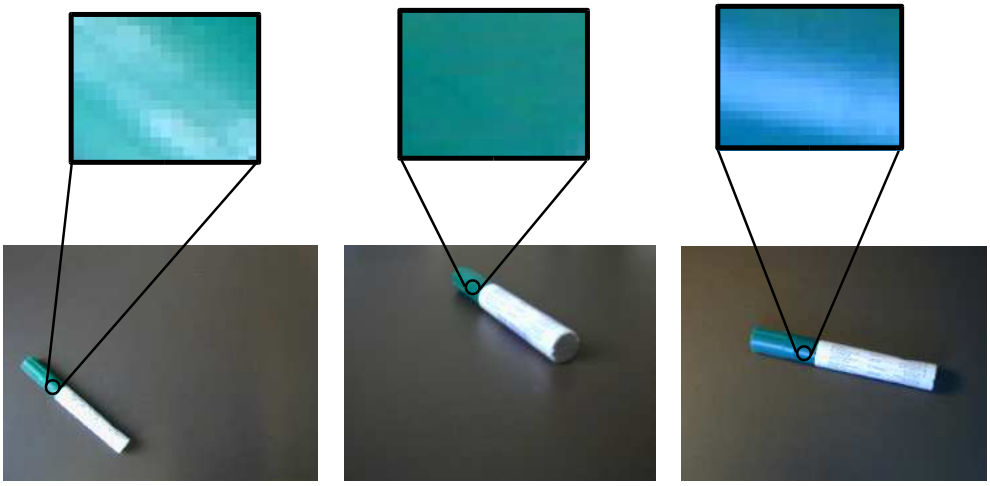

Figure 5: Ambiguity in the Colour domain: Three images of the same object under different illuminations and pose. The local pixel value depends on the object colour, the viewing angle and of the spectral distribution of the illumination. This distribution depends strongly on the light sources in the room, daytime, the amount of clouds, etc.

environments. However, to use the joint power of statistical and deterministic regularities we have identified four necessary properties of motion estimation algorithms in section 4.6. We will show that the algorithm [116, 115, 117, 38, 114] is especially useful in this context.

- Concerning the utilisation of statistical regularities, we refer in section 5 to an approach that has been first formulated by Brunswick [17] around 1950 but has only recently been justified by empirical data [72, 28, 35, 127]. Brunswick formulated the idea to relate the statistics of natural scenes to Gestalt laws. This leads to an understanding of the formalisation of Gestalt laws as essentially a *learning problem*. We suggest that by this approach the power of Gestalt laws becomes applicable in a more efficient way.

- We conclude this work by discussing the great potential for combining statistical and deterministic regularities in section 6 and we give directions for future research.

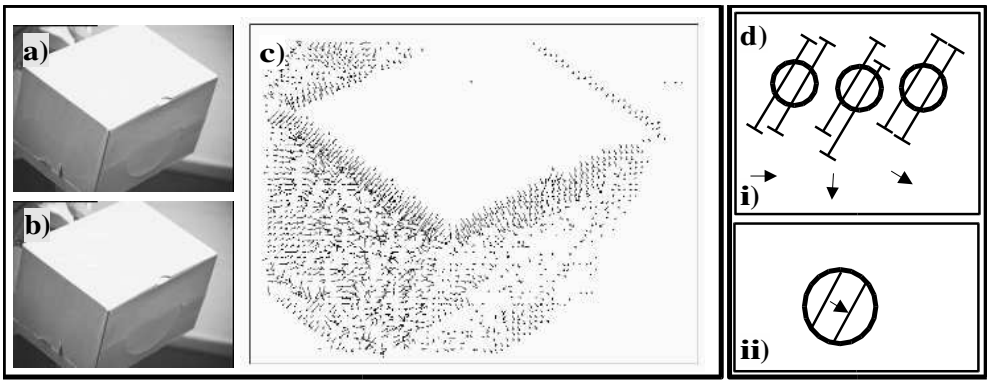

Figure 6: Ambiguity of Optic Flow: Left: a) + b) Two images showing a small downwards motion. c) Computated optic flow using the Nagel algorithm showing a considerable amount of noise and missing optic flow vectors. Furthermore, at edges only the normal flow (i.e., the flow orthogonal to the orientation of the edge) is computable because of the aperture problem. d) The three motions in i) are not distinguishable when looking only at a local patch, (or through an aperture): all three motions cause the same local pattern (ii). Therefore, only the motion in the direction of the intensity gradient is locally computable.

2 The Problem of Vagueness and Uncertainty in Vision

There is a large amount of evidence that the human visual system processes a number of aspects of visual data in its first cortical stages (see, e.g., [55, 95]). These aspects, in the following called visual modalities, cover, e.g., the local orientation [55, 56], junction structure [124] or colour [56] at a certain retinal position. Other aspects cover the relation across two or more frames. For example, the local motion describes the change of a certain visual event (e.g., the occurrence of a line) over consecutive frames [56]. In stereo processing the relation between visual events in the left and right image are processed and used to estimate depth information [8]. Accordingly, in many artificial computer vision systems, in the first stages one or more of the above–mentioned aspects are processed (see, e.g., [86, 121, 82]).
However, an important problem the human visual system as well as any artificial visual system has to cope with is an extremely high degree of ambiguity in these low level processes [2]. This has a number of reasons:

> **Noise and limited resolution:** Some reasons are associated with image acquisition: owing to noise in the acquisition process along with the limited resolution of cameras (in the enlarged frames in figure 1 these effects are clearly visible), only

rough estimates of semantic information (e.g., orientation) are possible.

Influence of illumination on colour: Illumination variation heavily influences the locally measured colour values and is hard to model analytically (see, e.g., [57]). For example, in figure 5, the same part of the same object is shown under different illuminations and pose. As can be seen, the change of the colour pixel values is considerable. Thus, when we want to use colour for, e.g., object recognition we need to overcome these changes.

Local feature processing: The processing of edges and junctions has been intensively investigated in computer vision during the last decades. However, the developed algorithms (see, e.g., [113, 58]) mostly work locally on image patches. In figure 1 (left) you see local image patches that correspond to edge– and junction–structures. Even for humans, a correct classification for many image areas is difficult if not impossible when looking locally at images. However, taking the global context into account (see figure 1 (right)), classification becomes feasible and even 3D attributes can be associated to the local areas.

Correspondence problem in optic flow: In optic flow estimation we want to estimate the local motion across consecutive frames. Computer Vision has developed a large number of optic flow algorithms (for an overview see, e.g., [9]). Figure 6 shows an image sequence and the optic flow derived by the well known Nagel algorithm [90]. It demonstrates some fundamental problems of optic flow estimation that come on top of the above mentioned problems:

- At homogenous patches either no motion or instead noise is computed. The underlying problem is that there is no way to find correspondences across such image patches.
- At edges only the normal flow (i.e., the motion component orthogonal to the local orientation) is computed. The underlying problem is the so called 'aperture problem': the fact that looking at a moving edge locally (i.e., through an aperture) all the different motions shown in (figure 6d) look the same.
- Only at junction like structures the correct motion can be computed. However, these occur only at very few image locations.

Loosing depth information: Probably the most serious problem is the fact that in vision the world is perceived by a projection onto a 2D surface: the retina in case of the human and a chip in case of a camera (see figure 7a). In this way, we loose the third dimension, i.e., depth information. However, there exist many cues to regain depth (so called depth cues). One important depth cue is stereo. In stereo processing (elaborately discussed in section 4.1) we can recontruct depth

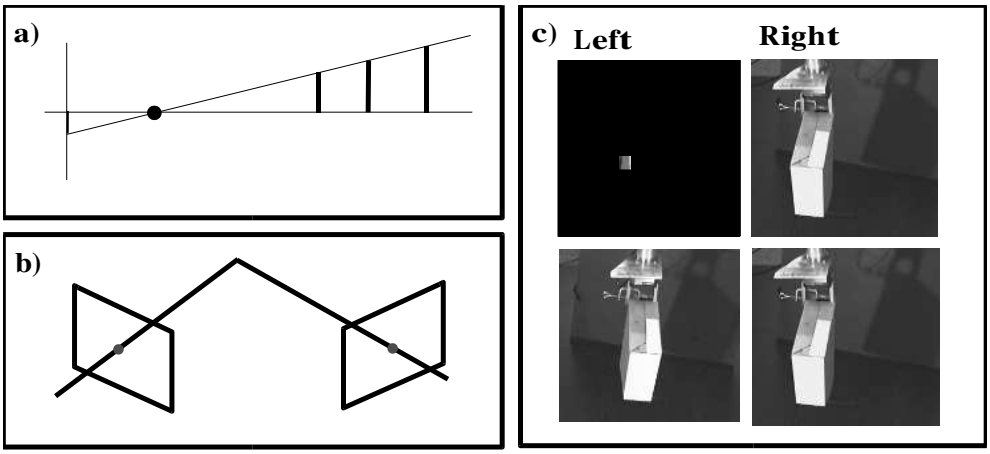

Figure 7: Ambiguity of Stereo: a) Image projection onto the retina or onto a camera chip leads to the loss of the depth dimension: All objects on the right produce the same projection. b) By means of corresponding points in two images taken from different perspectives we can reconstruct the depth information. c) However, finding correspondences can be extremely awkward when comparing image patches locally (top) but becomes feasible when context is taken into account.

information when we have correspondences in the left and right image recorded with a two camera system (see figure 7b). Accordingly, we need to find correspondences between pixels in the left and right image (see, e.g., [66]). Again, as can be seen in figure 7c (top), looking locally at an image pair this problem can be especially awkward. However, by looking at the context 7c (bottom) it becomes feasible.

3 Regularities in Visual Data

In section 2, we have described the ambiguity of local visual information as a fundamental problem of vision and have exemplified this problem in different visual domains. In the following, we will show that despite this problem, the human visual system can acquire visual representations which allow for actions with high precision and certainty even under rather uncontrolled conditions. The human visual sysem does this by *disambiguating locally erroneous information through integration across these visual modalities (see, e.g., [47]) and across spatial and temporal context* (see, e.g., [99, 21]). Integration and disambiguation is possible since there exist *regularities in visual data*. We now will take a closer look at such regularities in visual data. Then we will discuss aspects of their formalisation in human and artificial visual systems in section 4 and 5.

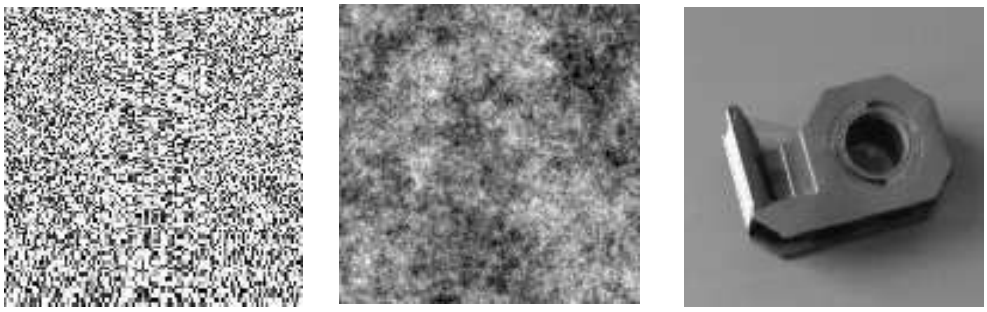

Figure 8: Image showing different amount of structure. a) White noise. b) 1/f noise. c) Natural image.

3.1 Why does Vision need Regularities

Natural images represent only a tiny subset of the set of possible images. Indeed, if an image would be produced by a random operator the chances that it would look like something which resembles a natural image are as good as zero. The image would look like white noise most of the time (see figure 8a). In natural images, there is a lot of structure that can not be found in white noise images. For example, David Field found a law (the so called $\frac{1}{f}$ law) about the distribution of energy in different frequency domains [32]. Images for which this $\frac{1}{f}$ law holds are still noisy but differ significantly from white noise (see figure 8b). There exists a good amount of work addressing specific structural properties of natural scenes (see, e.g., [128, 123, 54]). For example, local orientation plays an important role in natural images (see, e.g., [28]) and the local orientation of pairs of oriented pixels has a specific statistical distribution that corresponds to the Gestalt law of good continuation (see, e.g., [72, 27, 35, 127]). One can, for example, show that the visual event 'occurrence of a line segment with a certain orientation' increases the likelihood of the visual event 'occurrence of a collinear line segment nearby' (see figure 9a). Taking the temporal context into account, i.e., looking at image sequences, we find additional regularities. For example, it would be extremely unlikely that an object suddenly appears and then simply disappears. Usually, we approach an object or the object approaches us with a certain speed such that it becomes observable in consecutive frames. Therefore, events in one frame can be used to predict events in a second frame provided we have some information about the motion.
That leads us to a definition of regularities in visual data. We define a regularity in visual data as a *structural property of natural scenes that allows for the prediction of visual events.* Since regularities allow for predictions we can use them to define recurrent processes that trigger a disambiguation process that lead to stable percepts computed

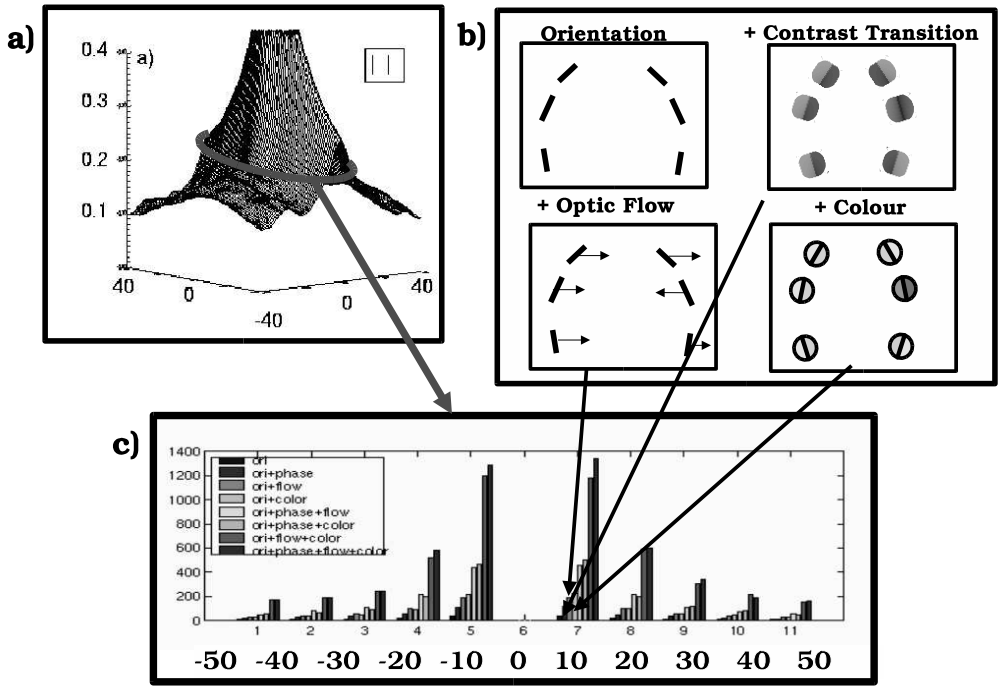

Figure 9: a) The cross–correlation of pairs of filter responses of same oriemtation. The x– and y–axes represent the distance of the kernels in x– and y– direction and the z-axis represents the correlation. Parallelism and collinearity are clearly visible: Collinearity is detectable as a ridge in the first diagram and parallelism appears as a global property expressed as offset of the surface in the first diagram that is missing for the surfaces corresponding to non–parallel orientations. In contrast to the high correlation of similar oriented line–segments, the correlation to non–similar orientations is low (detailed results can be found in [72]). b) Grouping becomes enforced through joint use of multiple modalities. c) Statistical Interdependencies for the collinearity ridge when multiple modalities are taken into account. The main axis represents the pixel distance for collinear line segments. Each sub–column represents the statistical interdependencies for different combinations of visual modalities. As can be seen, the principal form of the ridge is preserved also when using multiple visual modalities. However, the strength of the statistical interdependencies increases sigificantly (for details see [80]).

from the ambiguous inputs.

3.2 Statistical and Deterministic Regularities in Visual Data

As discussed in section 3.1, the power of modality fusion arises from intrinsic regularities in visual data. In this paper, we suggest a distinction between two kind of regularities:

1) **Deterministic regularities** (most importantly rigid body motion, RBM) applied in, e.g., stereo and motion processing (see figure 10a,b) and,

2) **Statistical regularities** between features as applied in, e.g, Gestalt laws such as good continuation or collinearity (see figure 10c) and in pictorial depth cues (see figure 10d,e).

Deterministic regularities allow for deterministic predictions that are in general based on *analytically describable geometrical* relations grounded on different perspectives of a scene or an object (see figure 10b). For example, knowing the RBM between two frames and knowing the 3D structure of the scene, we can explicitely compute the occurrence of features in consecutive frames (see, e.g., [31, 122, 74]). In this case, the occurrence of an event in the first frame makes the occurrence of a certain event in the second frame mandatory and therefore allows for *deterministic predictions*.[1] While the regularity RBM leads to deterministic predictions, statistical interdependencies occur as statistical correlations between events which only lead to *probabilistic predictions* about the occurrence of other events.
In this section,

1) we will give a precise definition of deterministic and statistical regularities in subsection 3.2.1 and 3.2.2.

2) we will summarize psychological and neurophysiological evidence supporting the assumption that abilities based on deterministic and statistical regularities develop *with a different amount of genetical pre-structuring* in the human visual system (section 3.3).

3) we will discuss computational differences between these two kinds of regularities that lead to the requirement of a different kind of mathematical framework in their formalization (section 3.4).

4) will discuss the consequences for the design of artificial systems that make use of these two regularities in section 3.5.

[1]Note that, since feature extraction necessarily is ambigiuous (see section 2) algorithms that make use of RBM also have to deal with uncertainty. However, the underlying regularity is deterministic.

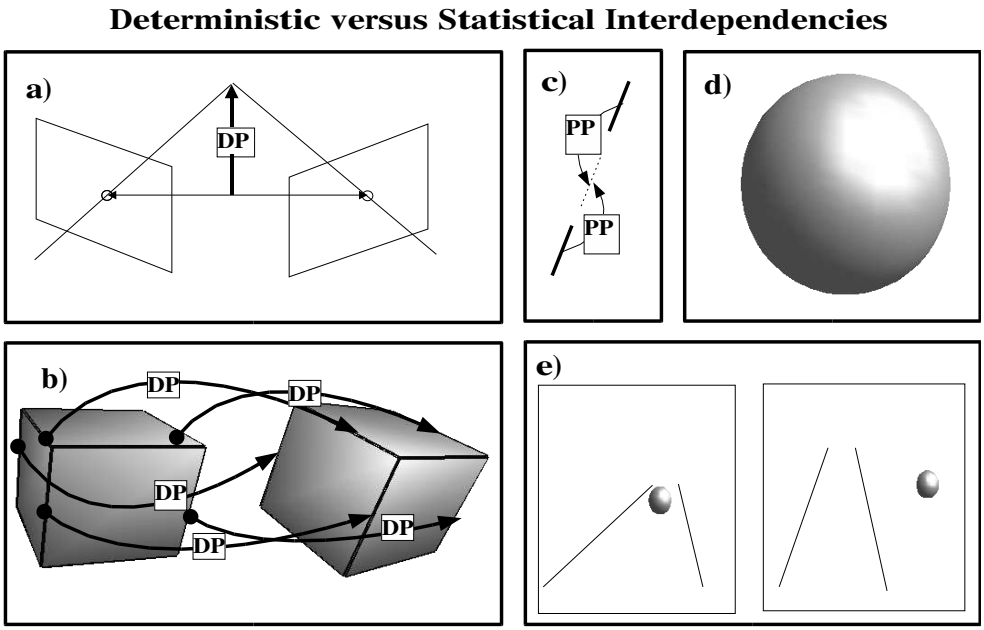

Figure 10: Examples of deterministic predictions based on geometrical regularities (a,b) and probabilistic predictions based on statistical regularities (c,d,e): a) In stereo processing two image correspondences lead to a deterministic 3D-prediction. b) Knowing the RBM between two frames for each feature in the first frame deterministic feature predictions for the consecutive frame can be made. c) The two collinear line segments (drawn bold) lead to an increase of likelihood for the existence of a third line segment inbetween. However, only a probabilistic prediction (PP) is possible. d) We tend to see a convex sphere illuminated from the top. However, this shading pattern could also be caused by a planar surface with appropriate texture or a concave surface illuminated from the bottom. The convex surface hypothesis is only the *most likely* in our visually experienced world. e) Using linear perspective, we tend to see in the left image two parallel lines in 3D and a ball on top. However, this pattern could also be caused by a different geometric structure (see right).

3.2.1 Statistical Regularities

To be more explicit, we give a definition and discuss one specific example of a statistical regularity in visual data.

Definition: There exists a statistical regularity between two visual events e and e' if the occurrence of the event e influences the likelihood of occurrence of the event e' but does not make the occurrence or non–occurrence of e' mandatory.

The most straightforward example of a statistical interdependency is Collinearity or Good

Continuation (see figure 2c and 10c): It has been shown by [72, 27, 35] that the occurrence of a line segment e influences the likelihood of the occurrence of other line segments e' in natural images in a very specific way: The likelihood of the occurrence of a collinear line segment e' increases up to a factor 6 while the occurrence of a parallel line segment increases approximately by a factor of 2 (detailed statistical investigations can be found in [72]). For line–segments with different orientation there is no significant statistical interdependency detectable (see [72, 27, 35]). Figure 9a shows the measured interdependencies on a large set of natural images.
There are many other examples in which statistical regularities are involved. For example, in Shape from Shading we want to compute 3D structures from 2D images making use of the grey level variation. The image shown in figure 10d appears to most people to be a convex sphere illuminated from the top. However, this is only the *most likely* interpretation in natural scenes since the very same grey level distribution is generated by a concave sphere illuminated from the bottom or simply a specifically textured planar surface. All pictorial depth cues, such as, e.g., linear perspective, are based or involve statistical regularities in visual data. The image in figure 10e (left) makes us perceive a sphere on a road-like surface. However, the same projective pattern can be also produced by a completely different 3D structure (see figure 10e (right)).

3.2.2 Deterministic Regularities

Definition: There exists a deterministic regularity between two events e, e' if the occurrence of the event e makes the occurrence or non–occurrence of e' mandatory.

The most important deterministic regularity in visual scenes is Rigid Body Motion (RBM): Assume the following is known:

- the 3D Position of a point e at time t_0 belonging to an object O in a first frame, and
- O is a rigid object, i.e. an object that does not change its form over a short interval of time, and
- the motion RBM of the object O from the first to the second frame, and
- that no external forces damage the object or influence its trajectory,

$\rightarrow$ than the event occurrence of a 3D point $e' = RBM(e)$ is mandatory[2].

Indeed, the RBM leads to predictions for each object feature (see figure 10b). Therefore, RBM is a strong regularity in visual scenes and its estimation is important. We will

[2] $RBM(e)$ is the point e to which the Rigid Body Motion RBM is applied.

discuss analytical properties of RBM as well as problems involved in the estimation of RBM in detail in section 4.

3.3 Utilisation of Statistical and Deterministic Regularities in Biological Systems

In section 2, we have discussed the problem of ambiguity of visual data. In section 3.1 and 3.2 we have argued that, since visual data is dominated by regularities, ambiguity can be reduced by utilizing these regularities in a system of recurrent predictions. An open question remains whether the existence, structure and utilization of such regularities is learned during the development of a human being or whether this is essentially hard–wired (i.e., learned by evolution). Similarly, as a designer of an artificial visual system we have to decide about what structural knowledge we want to build into our system and what structural knowledge we have to learn by experience.
The human visual system neither is a completely hard–wired structure nor a 'blanc table' or 'tabula rasa'. However, it is a system with the ability to adapt to and to learn essential aspects from the environment. This leads to the so called *bias variance dilemma* (see, e.g., [36]) that is faced by all systems with the ability to learn: If the starting configuration of the system has many degrees of freedom, it can learn from and specialise to a wide variety of domains, but it will in general have to pay for this advantage by weak generalization —the "variance" problem. This results in bad convergence and instability of the trained system. On the other hand, if the initial system has few degrees of freedom it may be able to learn efficiently but there is great danger that the structural domain spanned by those degrees of freedom does not cover the given application domain at all —the "bias" problem. This results in a system that may be well applicable in very special domains (such as, e.g., traffic sign recognition) but lack the quality of generalisation to other domains (such, e.g., tracking of vehicles). As a conclusion, Geman and Bienenstock [36] argue that a certain amount of "bias needs to be *designed* to each particular problem". However, each concrete choice of *a priori* knowledge is a crucial point: A wrong choice may lead to the exclusion of good solutions in the search space. A choice of predetermined structural knowledge that is too restricted may result in an increase of the search space, leading to unrealistic learning time and bad generalisation.
Within a biological system, bias can be established by genetic coding. The question of predetermined components is also most essential for the design of any artificial visual system that is able to learn, since this predetermined knowledge helps the system to focus on essential aspects in the huge amount of data it has to cope with. However, to actually find out what the genetically determined component is can be a difficult undertaking since learning and a priori knowledge may be deeply intertwined and difficult to separate by any kind of observation (for a detailed discussion see [73]).
How can we escape the bias/variance dilemma? The existence of the human visual system

with its ability to deal with its surroundings efficiently *and* with sufficient adaptivity raises hope that this problem can be solved. The predefined structural constraints applied in the human visual system have been evolved during evolution and appear to be well suited to organise visual experience: They seem to cover essential structures of the physical world. Thus, it is a valuable opportunity to look at the results from biology to become inspired for suitable definitions of constraints. In this sense, nowadays the Kantian idea [62] to establish a *table* of *a priori* constraints which organise perception can be supported, guided and justified by a good amount of neuropysiological and psychophysical data[3] We discuss such data in the next two subsections.

3.3.1 Evidence for a large Degree of Genetic Determination of Deterministic Regularities

At the end of the 19^{th} century, William James [59] characterised the world of the newborn as a "blooming, buzzing confusion". Imagine that there would not be any innate concept of depth, the idea that objects come into or leave existence when they appear or dissappear from the visual field would be inescapable. However, there exists a good amount of evidence that the newborn's world is not as confusing as assumed by James. Indeed, psychophysical research indicates that certain geometric relations of the Euclidian space are very likely not *learned* but are to a considerable degree *genetically determined.*
Depth information can be acquired by different cues. Cues based on deterministic dependencies are for example stereo (see figure 10a), and convergence of two eyes during fixation (see figure 11a). Statistical regularities are used by pictorial cues such as occlusion (see figure 11b), shading (see figure 10d), familiar size (see figure 11b) and linear perspective (see figure 10e) are applied for static depth extraction. Concerning the question of genetic prestructuring, it is interesting in which order these different cues develop and whether there is a percept of 3D established in newborns.
Kellman and Arterberry [64] state that 3D information is acquired even by the newborn:

> *Achieving accurate size perception ... implies that at least one source of egocentric distance information ... is functional at birth'.[64]*

In [64], it is also claimed that convergence must be the cue first applied. The stereo cue is used by babies after approximately 12 weeks and the whole stereo machinery starts rather instantly instead of showing steady increase of performance [45] probably caused by 'maturational change in cortical disparity-sensitive units' [64].
The start of utilising motion information (also based on deterministic regularities) for extracting 3D information is not fully clear. Some work indicates that one month old

[3]See also [73] where a priori constraints for object recognition have been motivated by neurophysiological and psychophysical investigations.

babies already can use motion information to extract depth [91]. In general, it is assumed that 'motion carried information about space appears to operate at the beginning ...' [64]. Neurophysiological research indicates that our concept of space (used, e.g., for navigation tasks) is realized in cortical maps as well as in maps in the hippocampus (probably relating to different competences on different evolutionary stages). There seems to be different brain areas genetically provided in which our geometrical representation of the world is realized and this representation is multi–sensorial (see, e.g., [4]), i.e., in these maps information of multiple sensors (e.g., vision, sound, touch, ...) is coded.
Note that the Gestalt law Common Fate (see figure 2f) which is based on motion (although not necessarily motion in 3D space) plays a special role in the development of the human visual system. It has been demonstrated by [129] that Common Fate is the Gestalt law that is used first by the human visual system. Spelke suggests that Common Fate is then used to establish other Gestalt laws. This has also been demonstrated in computational models in the group of Christoph von der Malsburg [104, 105].
The most likely conclusion we can draw from these findings is that a basic concept of depth (realized in genetically determined maps) is existing from birth on but that this idea is first (coarsely) realized by the depth cue convergence, then in addition by stereo and motion cues. All these cues are based on deterministic regularities in visual data. The use of pictorial cues (that are based on statistical regularities) evolve later (see the next sub–section).

3.3.2 Evidence for an Adaptive Component in the Ontogenesis of Abilities connected to Statistical Regularities

In contrast to the early use of deterministic interdependencies in depth perception, the use of pictorial depth cues more likely involves visual experience since these cues are used by 7 months old babies but not by 5 months old ones [64]. This has been independently shown for several pictorial depth cues: linear perspective [96] (see figure 10e), familiar size [142] (see figure 11b), occlusion [39] (see figure 11c) and shading [40] (see figure 10d). Besides the relative late occurrence of the ability to use pictorial cues there is conceptional evidence [72, 27, 35] and evidence from computational neuro–science [106] for an adaptive component in the ontogenesis of the ability to use statistical regularities which will be discussed in section 5.3.

3.4 Computational Differences between Statistical and Deterministic Regularities

Deterministic and statistical regularities are already widely used in artificial systems to stabilise uncertain and vague image information (for applications of deterministic interdependencies see, e.g., [130, 31, 44]; for applications of statistical interdependencies see, e.g.,

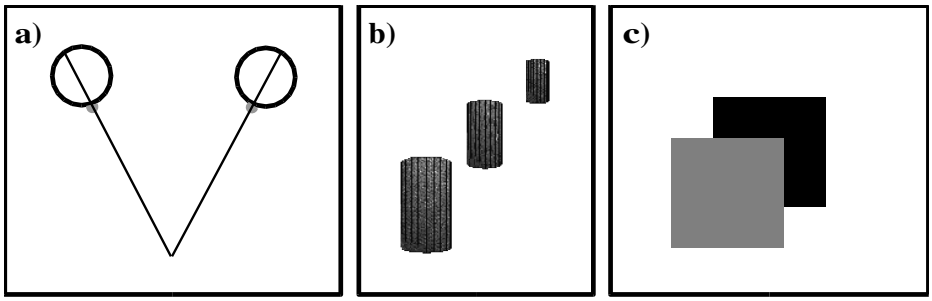

Figure 11: Variety of Depth Cues: a) Convergence: When a 3D point is fixated we can compute its 3D position from the angles of rotation of the two eyes by a simple geometric law. b) Familiar Size: Since we assume that the 3 objects having identical size we impose a relative distance of the objects. c) Occlusion: By assuming that the two patches with different grey level structure correspond to two objects that have the form of a square we impose a relative distance on the objects by assuming that they are superimposed.

[120, 43, 14]). As it will be shown below, these two regularities have different properties. As a consequence they have mostly been treated independently [120, 14]. Our central goal, however, is to design our system such that both regularities will support each other and we will discuss how this could be achieved in section 6.

There is a distinct difference in the success of usage of deterministic and statistical regularities in artificial visual systems. While the potential of geometrical constraints has been very successfully utilized within the last two decades (see, e.g., [44, 68, 66]) , the potential of statistical regularities has only been exploited to a much smaller degree. This holds even more for their combined exploitation [120, 14].

We argue that one reason for the different success in exploitation of deterministic and statistical interdependencies lies in their structural differences: Deterministic regularities can be modeled with a framework of reasonable complexity since they are analytically describable. In contrast, statistical interdependencies are based on specific statistical patterns in natural scenes that can not be modelled analytically but have to be learned.

RBM reflects a *geometric dependency* in the time-space continuum: The transformation of a non-deformable (rigid) object from one position to another. It is describable by six parameters, three for translation and three for rotation (see, e.g., [31, 89, 44]). The motion of a camera, the motion of a car within a static scene and also the motion of a rigid object on an assembly line can be fully captured by their RBMs. RBM is also the underlying regularity in stereo processing since it is the RBM between left and right camera that makes reconstruction possible.

A considerable amount of literature is concerned with RBM estimation from different

kinds of feature correspondences (see, e.g., [50, 100, 118]) which will be discussed in detail in section 4. Once the parameters of motion are known, RBM can be used for feature integration and robust feature extraction (see e.g., [6, 74]) since RBM allows for the *deterministic prediction* of a large number of feature events in the following camera frame *based on simple mathematical transformations* (see also figure 10b).
In contrast, collinearity and parallelism are two examples of *statistical regularities* in visual data, which are also associated with the so called 'Gestalt laws' (see figure 2) formulated by Gestalt psychologists (see, e.g., [138, 69]). These occur as statistical correlations between events which only allow for *probabilistic predictions* about the occurrence of other events. Take for example the Gestalt law 'collinearity' (or 'good continuation'): The occurrence of collinear line segments makes the existence of other collinear line segments *more likely* (see figure 10c and figure 9). In contrast to RBM, statistical relations between features *cannot normally be described analytically but require a statistical framework for their formalization.*
A lot of work has focused on the usage of statistical regularities to achieve robust feature extraction in different domains, e.g., edge detection (see, e.g., [43]) or stereo estimation ([19]). Another important application domain is the grouping of local entities into higher entities to achieve stable and fast matching ([88]). However, in most of these contributions the relation between features, i.e., the applied Gestalt principle, has so far only been *heuristically* defined based on semantic characteristics such as orientation or curvature (e.g., two line segments are defined to be collinear when they lie on a contour with slowly changing curvature ([140])). However, in section 5.3 we argue that by relating statistical regularities to statistics in visual data we can overcome such heuristic settings. To achieve this, the visual system has to be equipped with the ability *to adapt according to the statistical structure of visual data.*
Both kind of regularities, deterministic as well as statistical, can be used to extract depth information. However, they work in a complementary way. For example, stereo cues only work at close distances since the basis width of the camera system (often called baseline) has to be sufficiently large in relation to the depth range to be measured. Through ego motion we can increase the basic width and we can extend the depth range for which reconstruction is possible. However, for this we always need *different perspective views of the scene.* Statistical regularities are applied in 'pictorial cues' (see figure 10d,e for two examples). Pictorial cues allow extracting 3D information from 2D images without direct geometric experience of the 3D space. In section 3.3.2 we have seen that the ability to use pictorial cues evolves later in the ontogenetical development of the visual system. We argue, that since pictorial depth cues are based on statistical regularities there is a need to acquire knowledge of such statistical patterns and how they are related to depth information. As a consequence, there is a need for a certain amount of adaptivity of the system and therefore pictorial depth cues are applied rather late in the development of

the human visual system.

3.5 Consequences for the Design of Artificial Visual Systems

As discussed above, the human visual system faces two problems: Firstly, it has to deal with a considerable amount of uncertainty in its low–level modalities while at the same time it has to support actions with high reliabilty. It is widely agreed that this precision is achieved by integration based on the regularities in the visual data (see, e.g., [2]). Secondly, it faces the *bias–variance* dilemma. Thus, in order to be able to learn it already has to know something about structures of the environment in form of *predefined structural constraints*. This directly addresses the formalization of regularities since as a designer of an artificial system we have the awkward task to decide about the specific structural knowledge we want to be built into the system to realize abilities based on such regularities.
Taking the results described in section 3.3.1 into account, we think it is *justified to equip an artificial (human like) visual system with basic mechanisms for depth extraction* from stereo based on geometrical regularities (see, e.g., [66, 74]). Furthermore, we find it justified to equip the system with a basic mechanism to estimate the RBM between frames [116] as well as a mechanism that uses the estimated RBM to disambiguate locally erroneous visual estimates [68, 74].
However, although using deterministic relations based on RBM as largely hardwired components *we want to make use of the statistical interdependencies by a mechanism which relies on visual experience with real world data*. This kind of approach, although already formulated by Brunswick in the fifties [17], has only been recently become an intensively discussed issue [72, 27, 35, 127, 80, 28]. We will come back to this approach in section 5.3.

4 Formalisation, Estimation and Application of Rigid Body Motion

As discussed in section 3.1, 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., [68, 33, 31, 30, 122]). 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.

4.1 The deterministic Regularity Rigid Body Motion applied in Stereo Processing

In stereo processing we use the different appearance of features in two images taken from different positions in a scene to extract 3D information (see figure 7b). The RBM between the two cameras is essential in stereo processing since it produces the different appearances of image structures. Then simple geometric laws can be used to extract depth information.

4.1.1 The projective Map

By watching a scene with a camera the 3D world is projected onto a 2D chip (see figure 7a). This can be described (in a simplified camera model[4]) by the equation

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{X}{Z} \\ \frac{Y}{Z} \end{pmatrix} \tag{1}$$

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 (see figure 7a) in scene analysis. However, having two cameras that look at the scene from different viewpoints (see figure 7b) we can reconstruct 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., [31]), and the correspondence of two points with associated orientation lead to a 3D point with associated 3D orientation (see, e.g., [76]).

4.1.2 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 (see figure 7c). Indeed, it is this difference which on the one

[4]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., [31, 66])

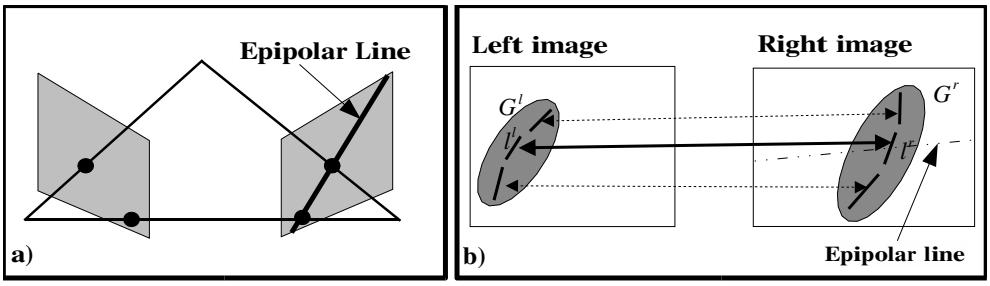

Figure 12: a) Epipolar Line Constraint. b) Predictions in the stereo domain based on grouping: Assuming the correspondence indicated by the solid line the correpondences indicated by the broken lines can be predicted.

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 (see figure 7c).

- 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 12 and [31]). 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[5] (see figure 12). In this way, we can reduce the correspondence problem to a one–dimensional search problem.

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

- It has been shown that the use of multiple modalities enhances stereo performance (see, e.g., [71, 76]). In our system, we have utilized the modalities orientation, phase, colour and optic flow to improve stereo matching [76, 107].

- There exist further spatial constraints [31, 66]. 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., [20, 37]).
  - 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., [19]). In figure 12b, 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.

4.2 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 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., [7, 65]) and will be described in detail below. This structure, often called 'Rigid Body Motion' (RBM)[6], 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

[6]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.

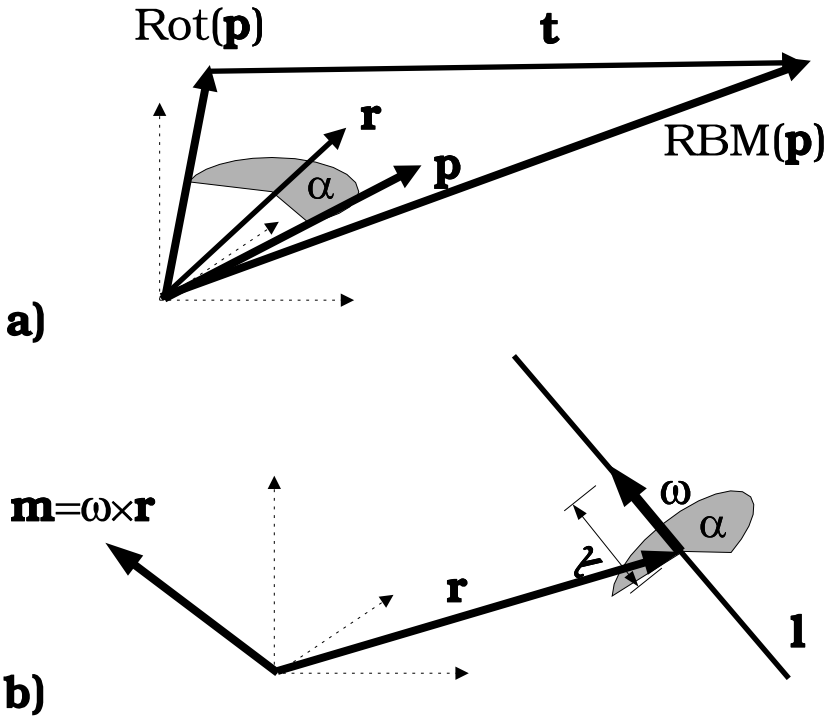

Figure 13: Two Representations of a Rigid Body Motion. a) Combination of rotation and translation. b) Twist representation: A rotation around a line $\mathbf{l}$ in the 3D Space with direction $\mathbf{w}$ and moment $\mathbf{m}$ and a translation along $\mathbf{w}$ with magnitude λ is performed.

(parametrised by $\mathbf{r} = (r_1, r_2, r_3)$). In figure 13a 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 $\mathbf{t}$.[7] 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[8] $\mathbf{e}$ in the first frame to a 3D entity $\mathbf{e}'$ in the second frame[9]

[7]There exist other ways to formalize an RBM, e.g., by Euler angles or dual quaternions (see section 4.7.2). However, it is always a six–dimensional manifold that describes the RBM

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

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

$$RBM^{(\mathbf{t},\mathbf{r})}(\mathbf{e}) = \mathbf{e}'. \tag{2}$$

To apply equation (2) we need to define correspondences between visual entities $\mathbf{e}$ and $\mathbf{e}'$.[10] 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

$$||RBM^{(\mathbf{t},\mathbf{r})}(\mathbf{e}) - \mathbf{e}'|| = 0. \tag{3}$$

Note that the norm $||\ ||$ can vary. This especially holds for different choices of entities $\mathbf{e}$. We discuss this issue in section 4.7.4. If we have a set of constraints (based on a set of 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 4.3.2). 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.[11] Therefore, we may not want to directly apply equation (3) but instead may want to embed this 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 4.1). However, as discussed in section 2 and 4.1.2, 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 4.4.1), we can apply entities of different semantic (see section 4.4.2): In equation (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 4 and 15). These groups can include

[10] 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 (2) would appear only implicitly in these methods (see, e.g., [18, 136, 93, 49]). In our approach, we do not follow this implicit approach. However, we will discuss the implications of the different methods in section 4.3.1.

[11] Note, that there exist sensors that record 3D information directly such as range finders [111]. 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.

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 4.5) that is even more serious than the correspondence problem in the stereo case (see section 4.1.2) since the epipolar constraint is not directly applicable[12]. The correspondence problem becomes even more severe in scenes with multiple independent motions. In section 4.5, 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 4.7. For example,

  - the solution of equation (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 4.7.1).
  - 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 4.7.2).
  - the way we represent mathematical entities such as points and lines (see section 4.7.3) influences the formulation of our constraint equations (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 (3) to ensure stability of computation. This will be discussed in section 4.7.4.

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

Having described basic problems of RBM estimation in section 4.3, 4.4, and 4.5, we will derive four desired requirements of RBM estimation algorithms for real world applications in section 4.6: 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.

[12] However, the epipolar line constraint can be used implicitely (see [131])

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 4.7 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 [116, 115, 117, 38, 114]. 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., [77]). 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.

4.3 Classification of Methods and Situations

4.3.1 Different types of Methods

In RBM estimation entities used to define correspondences can be represented explicitely as features (as done in equation 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 [131]).

- **Feature based methods:** In feature based methods [112, 84], at first features (e.g., junctions [101] or lines [77]) 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., [77, 21, 75]).

- In **optic flow methods** (see, e.g., [18, 49]) the optic flow with all its inherent ambiguities (see section 2) 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 [136, 93, 25]. 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.

4.3.2 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., [85, 116]). In the constraint equations we therefore need correspondences between 3D object and 2D image equations (see figure 14b).[13]

- **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[14]:

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

  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., [77]). 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.

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

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

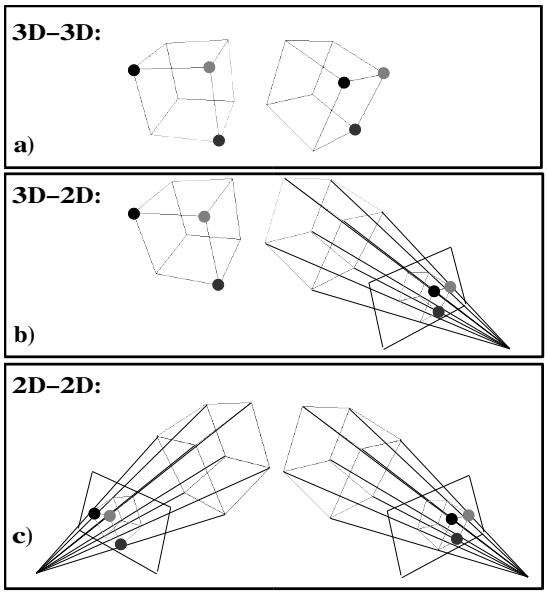

Figure 14: RBM–Estimation from different Correspondences. a) RBM estimation from 3D correspondences (displayed as circles). b) RBM estimation having a 3D model and 2D correspondences in an image. c) RBM estimation having 2D image coordinates in one image and its 2D correspondences in a second image.

4.4 Using Different kinds of Entities

In our constraint equations, we need correspondences between visual entities. These entities can have different spatial dimension (see section 4.4.1) as well as different semantic (see section 4.4.2). We will see that in the context of grouping both aspects are relevant.

4.4.1 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 14a). Second, we can have a model of an object that inherits 3D aspects, either by manual design (see, e.g., [85, 116]) or by some kind of acquisition mechanism that has taken place beforehand (see, e.g., [74]). In this case, 3D aspects of the object can be brought into correspondence with 2D aspects of its projection (see figure 14b). Thirdly, we can deal with 2D projections only (see figure 14c).

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 [111]). Then we can define correspondences in 3D and our constraint equations have the simple form

$$RBM(\mathbf{e}) = \mathbf{e}'. \tag{4}$$

From a mathematical point of view, this is the easiest case since we can avoid any problems resulting from the perspective projection (see section 4.1.1).
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 (see section 2) since wrong 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., [29]). 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 [41, 114].[15] In mathematical terms we have the following kind of constraint equations:

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

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

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

[15] 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, [68, 26]).

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, 132]) formulate the pose estimation problem by making use of the orthographic map.[16]

- **Simplified formulation in image coordinates:** In Lowe's pioneering work [85] 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 [5], who generalise Lowe's algorithm [85] 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 figure 16b) and an image line together with the optical center generates a 3D plane (see figure 16c). 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., [100, 38, 116]. 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 [79]).

[16] 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.

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., [122, 68, 101]). 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 [133, 44].

4.4.2 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., [41, 100, 85]) and lines (see, e.g., [48, 125]) and also for higher entities such as circle-like structures (see, e.g., [67, 114]).
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 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 15 shows some examples of possible groups. A particular property of groups (as will be discussed in section 5) 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., [135]).

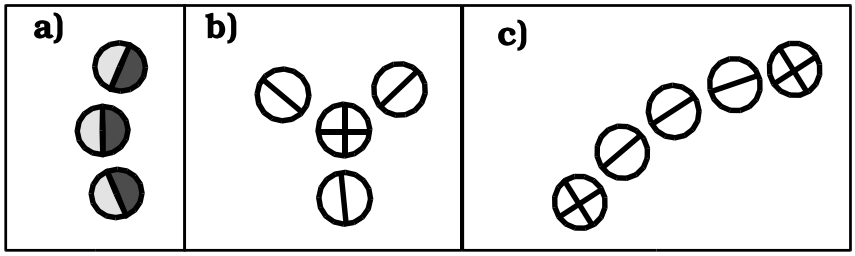

Figure 15: Examples of groups: a) Constellation of collinear line segments. b) A junction as a combination of an intrinsically two–dimensional and 3 intrinsically one–dimensional primitive. c) A collinear group with two defined endpoints.

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 [31] while dual quaternions are also suited to describe the RBM of a line (see, e.g., [125] and 4.7.2). 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., [38, 114]. A specific algebraic formulation in 'conformal algebra' (see, e.g., [46]) that allows for dealing with different kind of entities at the same time was helpful to derive such a formulation.

4.5 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 4.3.2). 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., [101]). However, for 3D–2D pose estimation from a single image (see, e.g., [84]) 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 4.7.4, 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 [53]). 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., [10]).
From the discussion in 4.1.2 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., [115]). 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, [80]). It has been shown that also the human visual system makes use of different modalities to improve matching performance (see, e.g., [47]).

- **Initial Estimate based on few Correspondences:** For RBM estimation we only need a small number of correspondences (see section 4.5). 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 [33].

- **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., [77]). 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., [101, 131]).

4.6 RBM Estimation and Grouping

In section 4.2 we have introduced the RBM estimation problem. For feature beased methods (see section 4.3.1) we have the option to formulate correspondences for entities of different dimension (see 4.4.1) and different complexity (see 4.4.2). As discussed in section 4.5 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 (2) and (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 4.5, 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., [51]).

- **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., [77]).

- **Flexibility:** Since groups may consist of different kinds of entities (e.g., points and line–like features, see figure 15) 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 15b), one correspondence is sufficient.

4.7 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 4.8.

4.7.1 Different kind of Optimisation Algorithms

The constraint equations (2) and (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

$$||RBM(\mathbf{p}) - \mathbf{p}'|| = ||A^{RBM}\mathbf{p} - \mathbf{p}'|| \tag{5}$$

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[17]. Therefore, additional (non–linear) constraints need to be defined to make sure that the matrix represents an RBM (see, e.g., [31]).
Using non–linear methods (see, e.g., [137]) 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, [131]). 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 4.7.4, the pose estimation algorithm [116, 115, 117, 38, 114] combines some of the advantages of linear and non–linear optimization methods.

[17] 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.

4.7.2 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 [31] and an RBM of a line as dual quaternions [125]. 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., [31]). A $RBM^{(\mathbf{t},\mathbf{r})}$ can be written as

$$RBM^{(\mathbf{t},\mathbf{r})} = \begin{pmatrix} r_{11} & r_{21} & r_{31} & t_1 \\ r_{12} & r_{22} & r_{32} & t_2 \\ r_{13} & r_{23} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} A(\mathbf{r}) & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix} \tag{6}$$

  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 ($||\mathbf{r}|| = \alpha$). 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 4.7.1): 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., [11]). 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 [100]. 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., [11]) that can be used to describe the RBM of lines (see, e.g., [125]). 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 [116, 115, 117, 38, 114] 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 [38]. A formalization of the very same approach using geometric algebra is given in [116, 115, 117, 114].

  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

$$e^{\tilde{w}\alpha} = \sum_{n=0}^{\infty} \frac{1}{n!}(\tilde{w}\alpha)^n \tag{7}$$

  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 (7), i.e.,

$$e^{\tilde{w}\alpha} \approx I_{3\times 3} + \tilde{w}\alpha. \tag{8}$$

On the other hand, having $\tilde{w}$ and α we can compute $\mathbf{A}(\mathbf{r})$ by the formula of Rodriguez (see, e.g., [89]):

$$\mathbf{A}(\mathbf{r}) = I + sin(\alpha)\tilde{w} + (1 - cos(\alpha))\tilde{w}\tilde{w}. \tag{9}$$

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 13b 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 4.7.3). In addition to the rotation a translation with magnitude λ along the line l is performed. According to Chasles' theorem, each RBM can be expressed in this way (see, e.g., [89]).

Then an RBM can be represented as

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

with

$$e^{\tilde{\xi}\alpha} = \sum_{n=0}^{\infty} \frac{1}{n!}(\tilde{\xi}\alpha)^n \tag{10}$$

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 (8) a straight forward linearisation is given by

$$e^{\xi\alpha} \approx (I_{3\times 3} + \tilde{\xi})\alpha. \tag{11}$$

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 (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 (6), since we still deal with a 12 dimensional description. However this representation expresses the motion parameters directly and, as will be shown in 4.7.4, can be used to derive a formulation that is very compact and efficient.

4.7.3 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 [116, 115, 117, 38, 114] that we use in our system [77], 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

$$\mathbf{F}^{\mathbf{p}}(\mathbf{x}) = \begin{pmatrix} p_1 - x_1 \\ p_2 - x_2 \\ p_3 - x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{12}$$

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[18]:

$$\mathbf{F}^{\mathbf{p}}(\mathbf{x}) = \begin{pmatrix} 1 & 0 & 0 & -p_1 \\ 0 & 1 & 0 & -p_2 \\ 0 & 0 & 1 & -p_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{13}$$

Note that the value $||\mathbf{F}^{\mathbf{p}}(\mathbf{x})||$ represents the Euclidian distance between $\mathbf{x}$ and $\mathbf{p}$. This will be important to derive interpretable constraint equations (see section 4.7.4).

- **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[19].

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

[18]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 (13) be identical to (12).

[19]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, $\mathbf{p}_2 \times \mathbf{r} = (\mathbf{p}_1 + \lambda\mathbf{r}) \times \mathbf{r} = \mathbf{p}_1 \times \mathbf{r} + \lambda\mathbf{r} \times \mathbf{r} = \mathbf{p}_1 \times \mathbf{r} + \mathbf{0} = \mathbf{p}_1 \times \mathbf{r}$

$$\mathbf{F^L}(\mathbf{x}) = \begin{pmatrix} 0 & r_x & -r_y & -m_x \\ -r_z & 0 & r_x & -m_y \\ r_y & -r_x & 0 & -m_z \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = 0 \tag{14}$$

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) [60, 114].

- **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:

$$F^{\mathbf{P}}(\mathbf{x}) = \begin{pmatrix} n_1 & n_2 & n_3 & -d_H \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{15}$$

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

In section 4.7.4, we will see that this implicit representation of entities in combination with the twist representation of an RBM (see section 4.7.2) and the formulation of the pose estimation problem in the Euclidian space (see section 4.4.1) allows for defining suitable and geometrically interpretable constraint equations.

4.7.4 Constraint Equations

After having formalized an RBM as a twist transformation in section 4.7.2 and geometric entities in section 4.7.3 we can now define constraint equations for different kind of correspondences.

3D-point/3D-point constraint: One can express the constraint equation (4) for the case that our corresponding entities are 3D points by using the linear approximation (11) of the twist $\tilde{\xi}\alpha$ and the implicit representation of points (12) by

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

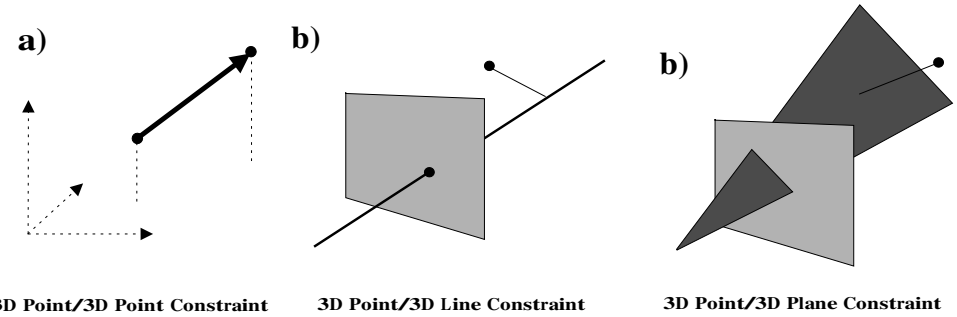

Figure 16: Geometric Interpretation of constraint equations. a) The 3D-3D point constraint realized the Euclidian distance between the two points. b) The 3D point/3D line constraint realizes the shortest Euclidian distance between the 3D Point and the 3D line. c) The 3D Point/3D Line constraint realizes the shortest Euclidian distance between the 3D Point and the 3D Plane.

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 16a).
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 16a).[20] 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 (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 4.7.3 (see also figure 16b).
Simple reordering gives:

$$\begin{pmatrix} 0 & -r_3 & r_2 & -p_3r_3-p_2r_2 & p_1r_2 & p_1r_3 \\ r_z & 0 & -r_x & p_2r_1 & -p_1r_1-p_3r_3 & p_2r_3 \\ -r_2 & r_x & 0 & p_3r_1 & p_3r_2 & -p_2r_2-p_1r_1 \end{pmatrix} \begin{pmatrix} \alpha v_x \\ \alpha v_y \\ \alpha v_z \\ \alpha w_x \\ \alpha w_y \\ \alpha w_z \end{pmatrix} =$$

$$\begin{pmatrix} p_3r_2-p_2r_3+m_1 \\ p_1r_3-p_3r_1+m_2 \\ p_2r_1-p_1r_2+m_3 \end{pmatrix}.$$

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 16c). 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

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

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

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 16c shows the geometric interpretation of the 3D point/2D line constraint.

4.8 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 [116, 115, 117, 38, 114] that has been described in section 4.7.3 and 4.7.4 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 4.7.3. Implicit representations of geometric entities had also been used by [60] 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., [141].

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.

5 The Utilisation of Statistical Regularities in Artificial Visual Systems

So called 'Gestalt laws' were defined by Gestalt psychologists such as Wertheimer [138], Koffka [69] and Köhler [70]. The Gestalt approach emphasizes the idea that we perceive the world as a whole rather than a set of isolated entities. The Gestalt laws represent holistic rules that are applied by the human visual system to group local visual entities into more complex groups. However, the exact number of rules remains unclear. Estimates range between 1 and 114 (see, e.g., [102]).
The most prominent Gestalt laws are the following:

> **Law of Proximity:** Visual entities near each other tend to be perceived as a unit (see figure 2b).
>
> **Law of Similarity:** Visual entities that are similar (in e.g., shape, colour, texture) tend to be perceived as a unit (see figure 2a).
>
> **Law of Good Continuation:** Visual entities that are organized in a straight or curved line tend to be perceived as a unit (see figure 2c).
>
> **Law of Closure:** Gaps in the organisation of visual entities are filled to perceive rather a unit than a set of isolated entities (see figure 2e).
>
> **Law of Common Fate:** Visual entities that move in a similar fashion tend to be perceived as a unit (see figure 2f).
>
> **Law of Prägnanz:** Of several possible organisations of visual entities the best, simplest and most stable one is chosen (see figure 2g).

In the context of the ambiguity problem of visual information (see section 2), most Gestalt laws represent statistical regularities in visual data that can be used to disambiguate visual information. For example, there is a broad spectrum of work that utilizes the Gestalt Law 'Good continuation' to improve edge detection (see, e.g., [140, 24, 43]). For a detailed overview of the application of Gestalt laws in artificial visual systems we refer to, e.g., [120, 14].
All the approaches that utilize Gestalt laws rely on the fact that visual data is not a mix of accidental structures but that there occur certain patterns of organisation that are generated by a common cause. Given a certain organisation of visual features, we are interested in the prediction of a such a cause behind the organisation. In terms of probability theory, we are interested in the expression

$$P(\text{Causality}|\text{Organisation}).$$

For example, the occurrence of a specific organisation of points with a common motion in a scene makes it very likely that this common motion is caused by the same object and therefore it is liekly that these points belong to the same object. As another example, the specific occurrence of a set of collinear line segments is likely not to be an accidental event but is likely to have a common cause, e.g., in belonging to a collinear 3D structure (possibly belonging to a 3D object boundary).
However, it is the organization that is detectable in visual scenes and not the cause. But we can refer from the organisation to the underlying cause by using Bayes rule:

$$P(\text{Causality}|\text{Organisation}) = \frac{P(\text{Organisation}|\text{Causality})P(\text{Causality})}{P(\text{Organisation})} \tag{16}$$

The different terms in (16) have different meaning:

> **P(Causality)**: Without any causality in visual data $P(Causality|Organisation)$ would always be low and no kind of perceptual organsisation could occur. For example, in white noise (see figure 8a) this factor is zero. However, as discussed in section 3.1, visual data is dominated by statistical and deterministic regularities and therefore $P(Causality)$ is expected to be high.
>
> **P(Organisation)**: In case that a certain pattern of organisation is likely its value for predictions does decrease since $P(Organisation)$ occurs in the denominator of equation (16). In this sense, the occurrence of line segments that are collinear in 3D are 'more valuable' than collinear line segments in 2D.
>
> **P(Organisation|Causality)**: This term represents a measure for the effect of a cause to a specific organisation. If this term is low the cause, although present, might not lead to the expected pattern of organisation. Therefore, we want the likelihood of the occurrence of a certain organisation given the cause to be high.

In the next subsection, we will discuss two issues concerning perceptual organisation.

> **Entities in grouping:** The question which entities to use for perceptual organization arises as for the formalization of RBM (see section 4.4). As for RBM the choice of entities used is crucial and is discussed in section 5.1. However, in the grouping process new, more complex entities can emerge which gives the problem of visual entities a more dynamic character. Furthermore, we will see that although grouping in the 2D domain has been addressed by many researches, there is little work about grouping in 3D and a void of work that addresses grouping in the spatial-temporal domain.

Grounding and Grouping: The questions of grounding of Gestalt laws is more urgent for statistical regularities than for deterministic regularities for which an analytic description is feasible. Furthermore, in the context of the bias–variance problem (see section 3.3) the question arises whether Gestalt laws are learned or genetically determined. It has been shown in section 3.3.2, that the establishment of some abilities based on Gestalt laws occurs rather late in the development of the human visual system. This makes the involvement of learning likely. In this paper, we suggest a grounding of percetual organisation based on the aquisition of statistical properties of the visual input.

In this section, we will restrict ourselves to issues that are relevant to the problem of combining RBM and Grouping.

5.1 Basic Entities in perceptual Organisation

To classify different approaches to perceptual organisation, Sarkar and Boyer [120, 14] have suggested a table with two axes. One axis represents the dimension of the input space which can be 2D entities (e.g., a pixel image), 3D entities (e.g., a range image or visual entities extracted by stereo), 2D–temporal entities (e.g., an image stream), or 3D–temporal entities. The other axis represents the entities that evolve in the grouping process. Sarkar and Boyer distinguish between a signal, primitive, structural and assembly level that correspond to different levels of abstraction. In table 5.1, we show a simplified version of the table of Sarkar and Boyer.

An essential property of perceptual organisation is the *change in complexity* of visual entities in the process of perceptual organisation (represented with the vertical axis in table 5.1). In this respect, the utilisation of statistical regularities differs fundamentally from the utilisation of RBM as discussed in section 4: The change of entities in the process of rigid motion only affects parameters such as position and orientation but the entity itself does not change its structure. However, it is a central property of perceptual organisation that a hierarchy of complexity becomes established. A specification of levels in this hierarchy is difficult to achieve and the whole process is rather dynamic: New constellations of visual entities become organized to more complex entities in a process of self-emergence.

This dynamic nature of the grouping process in which new, more complex entities evolve (indicated in the top row of each structural level) leads necessarily to some fuzziness in the vertical axis of table 5.1. The different levels in the hierarchy of entities might not be precisely definable since they are the results of a complex, dynamic process. It may even lead to a process which starts with 2D features that become grouped to 3D or temporal features, i.e., the hierarchy of complexity might not only be vertical (in terms of table 5.1) but might also occur along the horizontal axis.

Assembly Level	arrangement of polygons [108, 119]	surface clusters [34]		
Structural Level	closed regions, Polygons [3]	surface combinations [34]		
Primitive Level	regions, edge chains [12, 3]	parametrised surfaces, 3D curves [34, 13, 19]	flow patterns [143]	
Signal Level	point clusters, features [144, 43]	surface patches, 3D line segments range segmentation, stereo grouping	optic flow optic flow work	6D motion descriptors [61]
	2D points, grey level image	3D range images, stereo	2D + time image sequences	3D + time stereo image sequences

Table 1: Classification of Perceptual Organisation according to Sarkar and Boyer [120, 14]. The horizontal axis of the table represents the input domain while the vertical axis represents different levels of complexity.

What becomes most apparent in table 5.1 is the empty space at the top right corner, representing the lack of work addressing the combination of grouping in the temporal-spatial domain as addressed here in this review. In our view, this lack has two main reasons:

Conceptional and technical complexity: To deal with spatial-temporal patterns in visual scenes a machinery of considerable complexity is needed: 2D features need to be extracted from which 3D entities can be computed (preferably by making use of different visual modalities), motion has to be computed and the motion has to be used to stabilize the ambiguous input data. This requires a system in which the different modules realizing the different sub-aspects are organized in an integrated software structure (see, e.g., [87]). Also perceptual organization in itself requires a machinery of similar complexity. Therefore, it is a conceptional and technical pretentious task to bring these two streams together.

Difference of the formalisation framework: As discussed in section 3.4, the machinery to realize perceptual organization and the machinery that utilizes RBM have to be different. While for RBM an analytical framework can be applied in perceptual grouping a statistical framework is necessarily involved. The combination of perceptual grouping and RBM therefore requires a detailed knowledge of techniques of rather different character.

5.2 Approaches to ground Gestalt Laws

As discussed in 4, there exists a solid analytic framework of reasonable complexity that describes rigid body motion. No such framework has yet been defined for statistical reg-

ularities and it is unlikely that this is possible at all since statistical interdependencies are of much higher complexity. The classical Gestalt laws have been defined by heuristic rules and their exact formalisation allows for many, not necessarily equivalent, alternatives. This leads to the problem of grounding the Gestalt laws to overcome such heuristic assumptions.
Some attempts have been made for such a grounding. They will be briefly discussed now. Among them, the idea to ground Gestalt laws on the statistics of visual scenes (first formulated by Brunswick [17]) has become recently supported by different sources of evidence (see, section 5.3).

Iterative application of rules: For example, in [92] it is suggested that an application of low and high level rules may lead to significant percepts. A problem of such approaches is that the specific formulation of rules involves a considerable amount of arbitrariness. A grounding of such rules in a more general concept is therefore desired.

Coding theory Leeuwenberg [83] argued that perceptual organization is guided by a 'minimal coding principle': The percept is preferred for which a minimal amount of bits for memorizing is needed. However, the amount of information needed to represent a percept depends on, e.g., the choice of features and the process of development of more complex feature assemblies. Furthermore, measuring the amount of information in a complex visual representation might not be as easy as for an array of bits. The problem of heuristics in the definition of rules might then be displaced to the problem of heuristics in the measurement of quantities of information.

- **Group Theory:** Palmer [97] states that a good percept shows stability over a group of (e.g., Euclidean) transformations. For example, relative length of lines or number of lines are stable over a large variety of viewpoints and are therefore good concepts for visual representation. Such approaches take the importance of the spatial-temporal domain in visual perception into account. However, grouping occurs also in still images. Furthermore, it is difficult to incorporate learning in such an approach.

- **Global Coherence through Energy Minimization:** The relation of global coherence through local interaction is addressed in optimization methods that minimize a global energy term (see, e.g., [42, 1]). The global energy is computed by associating energies to local interactions between visual entities. A very convenient property is that a global guidance of the organisation is ensured. In a similar spirit, in [63] local rules are derived from a global principle. However, as for the approaches

that are based on coding theory a significant amount of arbitrariness is involved in the terms that compute the global energy from local interactions.

In the next subsection, we discuss another approach to ground Gestalt laws that takes the aspect of learning into account.

5.3 Relating Gestalt principles to the statistics of natural images

In section 3.3, we discussed evidence for a large influence of learning in the development of abilities based on statistical regularities. This is not reflected in the attempts to ground Gestalt laws discussed in section 5.2. However, decades ago, Brunswick and Kamiya [17] first had stated that Gestalt principles should be related to the statistics of the natural world. This offers the perspective to establish a bootstrapping process in which perceptual rules are learned by statistical measurements in visual scenes. This could lead to rule-based systems that are grounded *in specific statistical properties of the world.*
Unfortunately, the limited computational power at Brunswick's time made it difficult to quantitatively support his ideas quantitatively. Only recently, the strong prevalence of Gestalt laws such as collinearity and parallelism in natural images have been investigated by [72] and [27] (see figure 9). There results have been confirmed and extended by [127, 35]. These investigations suggest that Gestalt laws are reflected in the statistics of low level filter operation well established in human and artificial vision. In addition, it has been shown that these interdependencies become much stronger (in order of magnitudes) when we look at multi-modal statistics (taking also color, optic flow and contrast transition into account). The diagram in figure 9c shows that the probability for two segments being collinear rises if the segments show also similarities in other modalities (for details see [80]). Therefore, there is conceptional evidence for the *possibility* to learn such interdependencies from statistical measurements in visual data. Elder and Goldberg [28] demonstrated that not only collinearity and parallelism can be related to statistical properties of visual scenes but that also other Gestalt laws such as the 'Law of Proximity' and the 'Law of Similarity' are reflected in such statistics. They could also demonstrate a high correlation between the strength in the statistical occurrence of Proximity in images as well as the role of Proximity in human perception.
Evidence from computational neuroscience has been given by, e.g., [106] who have implemented a neural network model of primary visual cortex that leads to the *emergence* of collinearity when exposed to visual real world data. This emergence was accelerated by the additional use of motion that supports a segmentation of the object. Moreover, in [104] it is indicated that even more complex feature constellations (i.e., vertices) become significant in the statistics of natural images when segmentation by motion is applied as a pre–processing step. Here the idea is to use common fate as an initial cue to organize visual perception that initiates a bootstrapping algorithm in which other, more

complex Gestalt laws develop. This approach can be motivated by the specific order in the development of Gestalt laws as discussed in section 3.

6 Combination of Statistical and Deterministic Regularities

In this review, we have discussed two main regularities in visual data. There exists a solid analytic framework for RBM as discussed in section 4 and the utilisation of RBM in artificial visual systems has lead to impressive work in scene analysis (see, e.g., [44, 68, 66]). However, many problems remain open. The estimation of motion in complex scenes is not yet fully solved. This is not so much grounded in a lack of analytical understanding of motion but more in the problem of finding corresponding features across frames. Therefore, all the aspects of ambiguity in visual information (see section 2) enter the motion estimation problem. This holds even more in case of multiple motions where we have a segmentation problem on top of the correspondence problem.
In contrast to deterministic regularities, there is lack of grounding of statistical interdependencies for which an analytic framework can not be defined (see section 5.2). This is probably also the reason why the full potential of the statistical regularities is by far not used yet. The combined exploitation of statistical and deterministic regularities has only been addressed by very few scientists (see, e.g., [34, 19]). However, we hope that this review will help to address such a combined exploitation.
Sarkar and Boyer [14] have recently described five open problems in formalization of perceptual organisation from which we have addressed three in this paper.

Learning in perceptual organisation: In this review, we have argued that learning is a necessary component in the formalization of perceptual organisation. In the human visual system abilities based on perceptual organisation develop much later than abilities based on deterministic regularities leaving time for incorporating visual experience into the system. We think that the incorporating of statistical measurements may also lead to a *grounding* of Gestalt laws (see section 5.3).

Perceptual Organisation in 3D: Sarkar and Boyer discuss the understanding or perceptual organisation in range images as one important field of future research. Another example is the combination of grouping and stereo. In perceptual organisation new complex entities emerge from combination of less complex entities. This dynamic process of feature emergence can be used also to improve stereo matching: To give an example, assume a local line segment in the left image is part of a group (defined initially by collinearity of local line segments) in the left image (see figure 12b). Furthermore, assume that this line segment has a correspondence in the right image which in a similar way is part of another group, then all local entities of the group in the left image must have a correspondence in one of the local entities of

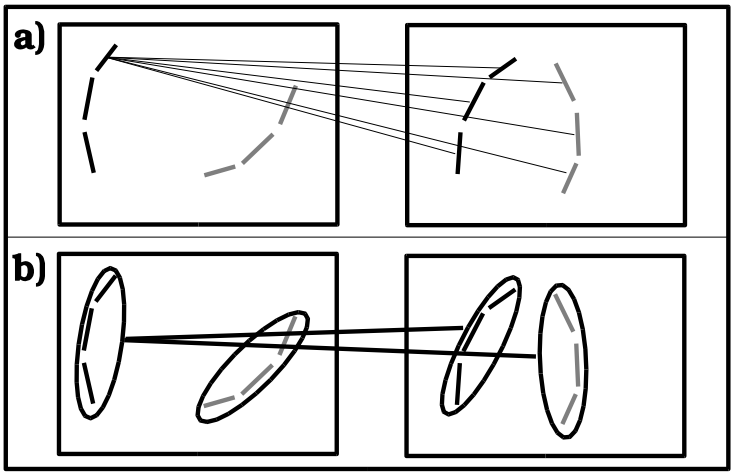

Figure 17: Motion Estimation and Grouping: a) For each entity in the top row there are 6 correspondences possibly leading to $6^6 = 46656$ possible patterns of correspondences. b) Grouping leads to a reduction to 2 correspondences only.

the group in the right image. By using this stereo constraint the existence of individual groups (first characterised by collinearity only) gets reconfirmed now also by the stereo correspondences of the local line segments the group consists of. As a consequence, groups of collinear local line segments emerge in this simple recurrent prediction process. This type of emergence is embedded, supported and gets support in the stereo correspondence process.

Perceptual Organisation in Motion Sequences: The joint power of statistical and deterministic regularities can also be applied in the temporal domain. As discussed in section 4.5, there is a serious correspondence problem in the motion estimation task that becomes even more significant in case of multiple motions. The perceptual organisation of local entities to more complex entities can support RBM estimation in respect of 3 aspects:

- The number of possible correspondences decreases with the complexity of entities (see figure 17).
- For perceptually organised entities more powerful constraint equations can be defined as combination of the constraint equations corresponding to the single entities (see section 4). Actually, for one sufficient complex group one correspondence might be sufficient to estimate an RBM.

– The grouping allows for a more precise estimate of the parameters of the local entities and therefore allows for more precise RBM estimates.

7 Conclusion

In this review, we have characterised two different kind of regularities in visual data. Their role in human vision and their application in artificial visual systems has been discussed. We have shown in section 2 and 3, that the use of regularities is necessary to disambiguate visual information. Both regularities, statistical and deterministc ones, have different properties and are realized on a different time scale in human vision (see section 3). They have been exploited also in artificial visual systems, however rarely in a combined way. We have discussed the potential of such an combined usage as well as the underlying problems. We hope that this review will help to overcome some of these problems.

Remark: This review addresses a broad field. Actually, it addresses a couple of broad fields. Although we have tried to get a good literature overview of all these fields, we found out that this is probably an impossible task. Therefore, it is likely that some researchers may feel that their work should be mentioned within the scope of this paper. Some may feel that their has not been described correctly or they may have some remarks to the paper as such and the mentioning of their work in particular. We would like to ask all readers who want to comment on the issues addressed here to give us feedback. Our feeling and hope is that the combination of perceptual organisation and motion will become a 'hot topic' for future research and we are happy to address all remarks in future publications.

Acknowledgment: We thank Bodo Rosenhahn for his help during the writing of this work. His patient support was essential for section 4. We also thank Eckhard Steinbach and Oliver Granert for their feedback. We would like to thank Michael Felsberg, Gösta Granlund, Bill Phillips, Jan Koenderink and Andrea van Dorn for the fruitful discussions about issues addressed in section 3.3. We especially thank all students that have been involved in the Modality Integration project ([87]): Markus Ackermann, Kord Ehmcke, Christian Gebken, Oliver Granert, Danial Grest, Marco Hahn, Thomas Jäger, Nicolas Pugeault, Martin Pörksen, Torge Rabsch, Daniel Wendorff and Jan Woetzel.

References

[1] N. Ahuja and M. Tuceryan. Extraction of early perceptual structure in dot patterns: Integrating region, boundary, and component estalt. *Computer Vision, Graphics,*

and Image Processing, 48:304–356, 1989.

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

[3] A. Amir and M. Lindenbaum. A generic grouping algorithm and its quantitative analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:168–185, 1998.

[4] R.A. Andersen, L.H. Snyder, D.C. Bradley, and J. Xing. Multimodal representation of space in the posterior parietel cortex and its use in planning movements. *Annu. Rev. Neuroscience*, 20:303–30, 1997.

[5] 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.

[6] C. Baillard and A. Zisserman. A plane-sweep strategy for the 3D reconstruction of buildings from multiple images. In *19th ISPRS Congress and Exhibition*, Amsterdam, July 2000.

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

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

[9] 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.

[10] 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.

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

[12] M. Boldt, R. Weiss, and E. Riseman. Token-based extraction of straight lines. *IEEE Transactions on Systems, Man, and Cybernetics*, 1989.

[13] K.L. Boyer, M.J. Mirza, and G. Ganguly. The robust sequential estimator: A general approach and its application to surface organisation in range data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:987–1001, 1994.

[14] 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.

[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] R.A. Brooks. Intelligence without reason. *International Joint Conference on Artificial Intelligence*, pages 569–595, 1991.

[17] E. Brunswik and J. Kamiya. Ecological cue–validity of 'proximity' and of other Gestalt factors. *American Journal of Psychologie*, LXVI:20–32, 1953.

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

[19] 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.

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

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

[22] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W.v. Seelen. Walking pedestrian recognition. *IEEE Trans. Patt. An. Transp. Sys.*, 1(3):155–163, 2000.

[23] W.D. Daley, S. Grullon, and D.F. Britton. Machine-vision-based quality control decision making for naturally varying product. *SPIE*, 1999.

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

[25] 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.

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

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

[28] H. Elder and R.M. Goldberg. Ecological statistics of gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2002.

[29] 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.

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

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

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

[33] 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.

[34] R.B. Fisher. *From Surfaces to Objects: Computer Vision and and THree Dimsnional Scene Analysis*. New York: Wiley, 1989.

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

[36] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1995.

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

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

[39] C.E. Granrud and A. Yonas. Infants' perception of pictorially specified interpositions. *Journal of Experimental Child Psychology*, 37:500–511, 1984.

[40] C.E. Granrud, A. Yonas, and Pettersen. Infants' sensitivity to the depth cue of shading. *Perception and Psychophyiscs*, 37(5):415–419, 1985.

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

[42] S. Grossberg and E. Mingolla. Neural dynamics of perceptual grouping: Texture, boundaries and emergent segmentation. *Perception and Psychophysics*, 38(2):141–171, 1992.

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

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

[45] R. Held, E. Birch, and J. Gwiazda. Stereo acuity in human infants. *Proceedimgs of the Natianal Acadamy of Sciences, USA*, 77, 1980.

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

[47] 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.

[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] B.K.P. Horn and Weldon E, J. Direct methods for recovering motion. *International Journal of Computer Vision*, 2:51–76, 1988.

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

[52] http://robosapiens.mit.edu/p3.htm.

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

[54] J. Huang, A.B. Lee, and D. Mumford. Statistics of range images. *CVPR*, pages 1324–1331, 2000.

[55] 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.

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

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

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

[59] W. James. *The principles of psychology (Vol. 2)*. New York: Holt, 1890.

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

[61] C. Kanbhamettu and D.B. Goldgof. Point correspondence recovery in non-rigid motion. *Proceedings of CVPR*, pages 222–227, 1992.

[62] I. Kant. *Kritik der reinen Vernunft*. 1781.

[63] J. Kay, D. Floreano, and W.A. Phillps. Contextually guided unsupervised learning using local multivariate binary processors. *Neural Network*, 11(1):117–140, 1998.

[64] P.J. Kellman and M.E. Arterberry, editors. *The Cradle of Knowledge*. MIT-Press, 1998.

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

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

[67] 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.

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

[69] K. Koffka. *Principles of Gestalt Psychology*. Lund Humphries, London, 1935.

[70] K. Köhler. *Gestalt Psychology: An introduction to new concepts in psychology*. New York: Liveright, 1947.

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

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

[73] N. Krüger. Learning object representations using a priori constraints within orassyll. *Neural Computation*, 13(2):389–410, 2001.

[74] 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.

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

[76] 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.

[77] 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.

[78] 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.

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

[80] 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.

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

[82] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamik link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[83] E.L.J. Leeuwenberg. Quantification of certain visual pattern properties: Salience, transperency and similarity. In E.L.J. Leeuwenberg and H.F.J.M. Buffart, editors, *Formal Theories of Visual Perception*, pages 217–298. John Wiley and Sons, 1978.

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

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

[86] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information.* Feeman, 1977.

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

[88] R. Mohan and R. Nevatia. Perceptual organization for scene segmentation and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992.

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

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

[91] J.E. Nanez and A. Jonas. Effects of luminance and texture motion on infant defensive reaction to optical collision. *Infant Behaviour and Development*, 17:165–174, 1994.

[92] A.M. Nazif and M.D. Levine. Low level segmentation: An expert system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(5):555–577, 1984.

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

[94] R.C. Nelson and A. Selinger. A cubist approach to object recognition. *CVPR'95*, 1995.

[95] M.W. Oram and D.I. Perrett. Modeling visual recognition from neurobiological constraints. *Neural Networks*, 7:945–972, 1994.

[96] S. Oross, E. Francis, D. Mauk, and R. Fox. The ames window illusion: Perception of illusionary motion by human infants. *Journal of Experimantal Psychology*, 13(4):609–613, 1987.

[97] S.E. Palmer. The psychology of perceptual organisation: A transformational approach. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and Machine Vision*, pages 269–339. John Wiley and Sons, 1983.

[98] D.E. Pearson. The extraction and use of facial features in low–bit rate visual communication. *Philosphical Transactions of the Royal Society London B.*, 335:79–85, 1992.

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

[100] 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.

[101] 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.

[102] J.R. Pomerantz. Visual form perception: An overview. In E. Schwab and H. Nusbaum, editors, *Pattern Recognition by humans and machines:Visual Perception*, pages 1–30. Orlando, FL: Academic Press, 1986.

[103] A.P. Pope and D.G. Lowe. Learning object recognition models from images. In T. Poggio and S. Nayar, editors, *Early Visual Learning*. 1995.

[104] M. Pötzsch. *Context specific Statistics of Real Image Sequences leads to Corners (PhD Thesis)*. 1999.

[105] C. Prodöhl. *Beiträge zur visuellen Figur-Hintergrund-Trennung*. Diplomarbeit, Institut für Neuroinformatik, Bochum, 1998.

[106] C. Prodöhl, R. Würtz, and C. von der Malsburg. Learning the gestalt rule collinearity from object motion. *Neural Computation*, 2003.

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

[108] K. Rao and R. Nevatia. Space description from imperfect and incomplete data. *International Conference on Pattern Recognition*, pages 414–426, 1990.

[109] R.P.N. Rao and D.H. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence Journal*, 78:461–505, 1995.

[110] P. Remagnino, Graeme A.J., N. Paragios, and C.S. Regazzoni, editors. *Video-Based Surveillance Systems*. Kluwer, 2001.

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

[112] 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.

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

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

[115] 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.

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

[117] 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.

[118] B. Rosenhahn, Y. Zhang, and G. Sommer. Pose estimation in the language of kinematics. *Second international workshop, Algebraic Frames for the Perception-Action Cycle, AFPAC 2000, LNCS 1888*, 2000.

[119] S. Sarkar. Learning to form large groupings of salient image features. *CVPR*, pages 780–786, 1998.

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

[121] B. Schiele and J.L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. *Advances in Neural Information Processing Systems*, 8:865–871, 1996.

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

[123] O. Schwartz and E. Simioncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825, 2001.

[124] 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.

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

[126] Deep Blue Web Side. *http://www.research.ibm.com/deepblue/.*

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

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

[129] E.S. Spelke. Principles of object perception. *Cognitive Science*, 14:29–56, 1993.

[130] M.E. Spetsakis and J. Aloimonos. Structure from motion using line correspondences. *International Journal of Computer Vision*, 4:171–183, 1990.

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

[132] 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.

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

[134] E. Schulze-Krüger W. Konen. Zn-face: A system for access control using automated face recognition. *In: M. Bichsel (ed.), Proc. Int. Workshop on Automated Face and Gesture-Recognition, Zurich*, 1995.

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

[136] 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.

[137] 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.

[138] M. Wertheimer, editor. *Laws of Organsation in Perceptual Forms.* Harcourt & Brace & Javanowitch, London, 1935.

[139] L. Wiskott, J.M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–780, 1997.

[140] D.M. Wuescher and K.L. Boyer. Robust contour decomposition using constant curvature criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(1):41–51, 1991.

[141] 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.

[142] A. Yonas, L. Petterson, and C.E. Granrund. Infants' sensitivity to familiar size as information for distance. *Child Development*, 53(5):1285–1290, 1984.

[143] J. Zhong, T.S. Huang, and R.J. Adrian. Salient structure analysis of fluid flow. *CVPR*, pages 310–315, 1994.

[144] S.W. Zucker. Computational and psychophysical experiments in grouping: Early orientation selection. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and Machine Vision*. San Diego: Academic Press, 1987.