Appeared in the proceedings of the workshop 'On growing up artefacts that live' at

SAB'2002 in Edinburgh

Different Degree of Genetical Prestructuring in the Ontogenesis of Visual Abilities based on Deterministic and Statistical Regularities

Norbert Krüger and Florentin Wörgötter

July 27, 2002

Abstract

In this paper we address the problem of predetermined structural knowledge human and artificial visual systems have to apply to be able to start a bootstrapping process that enables them to interact with the world efficiently. We focus on two different kind of regularities, namely deterministic and statistical, which visual system need to apply to overcome the ambiguity of visual low level processes.

1 Introduction

An important problem the visual system (as well as each system with the ability to *learn*) has to deal with is the *bias variance dilemma* [12]: 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 having many internal degrees of freedom —the "variance" problem. 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. As a conclusion [12] argue "bias needs to be *designed* to each particular problem".

Within a biological system bias can be established by genetical coding. The question of predetermined components is also most essential for the design of any visual artificial 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. 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 knowledge that is too restricted may result in an increase of the search space, leading to unrealistic learning time and bad generalisation.

How can we escape the bias/variance dilemma? The existence of a the human visual system with its ability to deal with its surroundings efficiently *and* with sufficient adaptivity

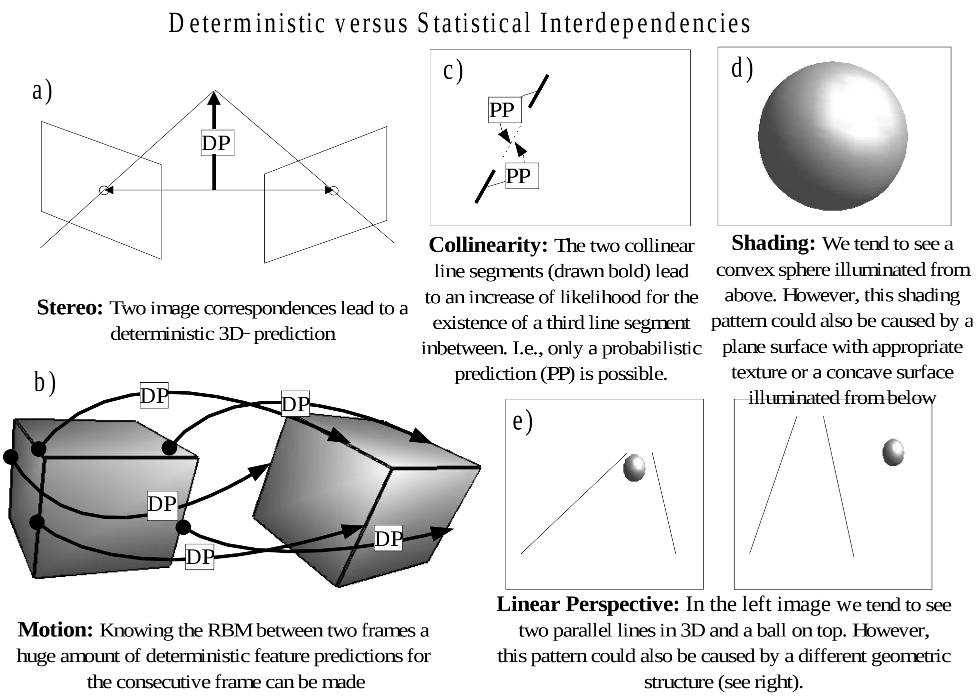

Figure 1: Examples of deterministic predictions based on geometrical regularities (left) and probabilistic predictions based on statistical regularities.

raises hope that this problem can be indeed solved. The predefined structural constraints have evolved during evolution and appear to be well suited to organise visual experience. Therefore they seem to cover essential structure 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.

Another important problem the visual system has to cope with is an extremely high degree of vagueness and uncertainty in its low level processes such as edge detection, optic flow analysis and stereo estimation [1]. However, by *integrating information* across visual modalities (see, e.g., [32, 8]), the human visual systems acquires visual representations which allows for actions with high precision and certainty within the 3D world even under rather uncontrolled conditions .

The power of modality fusion arises from the huge number of intrinsic relations across visual modalities. We suggest to distinguish between two kind of regularities:

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

2) **Statistical interdependencies** between features as applied in, e.g, Gestalt laws such as collinearity (see figure 1c) or in pictorial depth cues (see figure 1d,e).

Deterministic interdependencies are based on *analytically describable geometrical* relations grounded on different perspectives of a scene or an object (see figure 1b). For example, knowing the RBM between two frames *deterministic predictions* between frames can be made (see, e.g., [10, 38]): The occurrence of an event in the first frame makes the occurrence of a certain event in the second frame mandatory (except in the special case of occlusion). 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.

We claim that in order to develop the ability to exploit these regularities a distinguishable amount of genetical pre-structuring in the human visual system is determined: We suggest that the use of *deterministic interdependencies* is to a large degree *genetically determined.* On the other hand, we argue that the use of *statistical interdependencies* presupposes an *extensive interaction with visual real world data.* We are aware that we address a to a high degree arguable issue and there do exist divergent point of views (see, e.g., [13]) that suggest that also the use of deterministic regularities based on geometrical relations are learned.

This work is organised as follows: In section 2 we discuss structural differences between geometrical and statistical regularities in visual data. In section 3 we summarize neurophysiological and psychophysical research indicating a large degree of genetically determination in the ontogenesis of geometrical regularities. In section 4 we discuss analog kind of research for statistical regularities indication a later development and a larger degree of adaptivity in the ontogenetic process. In section 5 we draw conclusions for the design of artificial systems and give examples of the application of both regularities applying different degrees of adaptivity.

This work is in similar spirit to that in [25] in which a priori constraints for object recognition have been motivated by neurophyiological and psychophysical investigations. In [25] we focused on *2D aspects* of vision and *local* feature processing. However, the work introduced here goes beyond [25] by looking at aspects in the *3D–time domain* and discussing *intermediate stages* of visual processing, namely the coding of *constellations of local features.*

2 Structural Differences of Deterministic and Statistical Interdependencies

Deterministic and statistical regularities interdependencies are already widely used in artificial systems to stabilise uncertain and vague image information (for applications of deterministic interdependencies see, e.g., [10, 26, 38]; for applications of statistical interdependencies see, e.g., [16, 36]). As it will be shown below, these two regularities have different properties. As a consequence they have mostly been treated independently ([36, 37]).

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 applied for, e.g., scene reconstruction (see, e.g., [23, 22]) the potential of statistical regularities has only been exploited to a much smaller degree [36, 37]. This holds even more for their combined exploitation [36, 37]. We argue that one reason for the different success in exploitation of deterministic and statistical interdependencies lies in their structural differences: Deterministic regularities are easier to model since they are analytically treatable with a machinery of reasonable complexity.

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 analytically describable by six parameters, three for translation and three for rotation (see, e.g., [10, 29]). The motion of a camera, the motion of a car within a static scene and also the motion of an 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. Once the parameters of motion are known, RBM can be used for feature integration and robust feature extraction (see e.g., [4, 26]) since RBM allows for the *deterministic prediction* of a huge number of feature events in the following frame based on simple mathematical transformations (see also figure 1b).

Collinearity and parallelism are two examples of *statistical regularities* in visual data, which are also associated with the so called 'Gestalt laws' formulated by Gestalt psychologists (see, e.g., [40]). These occur as statistical correlations between events which only allow *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 1c). Another example in which probabilistic predictions are involved is the pictorial depth cue 'shape from shading'. The pattern in figure 1d makes us believe to see a 3D convex sphere illuminated from above. However, the same pattern can be produced by an appropriate textured plane or an concave surface illuminated from below. A further pictorial depth cues is linear perspective. The image in figure 1e (left) makes us see a sphere on a road-like surface. However, the same projective pattern can be also produced by a completely different 3D structure (see figure 1e (right)).

In contrast to RBM, statistical relations between features *cannot normally be described analytically.* A lot of work has focused on the usage of these relations to achieve robust feature extraction in different domains, e.g., edge detection (see, e.g., [16]) or stereo estimation ([7]). In most of these contributions the relation between features, i.e., the applied Gestalt principle, has so far been only *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 ([41])). However, we argue that by relating statistical regularities to statistics in visual data we can overcome such heuristic settings. However, 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 the stereo cue only works for close depth since the basis width of the camera system has to be sufficiently large in relation to the depth. Using ego motion we can increase the basic width and we can extent the range of depth for which reconstruction is possible. However, we always need *different perspective views of the scene.* Statistical regularities are applied in 'pictorial cues' (see figure 1d,e for two examples). Pictorial cues allow to extract 3D information from 2D images without direct geometric experience of the 3D space. In section 4 we will see that the ability to use pictorial cues evolves later in the ontogenetical development of the visual system, indicating the need of a certain amount of adaptivity of the system.

To sum up this section, we have shown that geometrical and statistical regularities in visual data have distinguishable properties. 1) While the first regularity leads to deterministic predictions the second only allows probabilistic predictions. 2) Deterministic regularities are reasonably analytical describable while statistical regularities need more sophisticated descriptions. 3) The application in artificial visual systems has been so far more successful realized for deterministic than statistical regularities. 4) Geometric regularities allow depth estimation based on two or more views of a scene only while the statistical regularities enable depth estimation from one view. In the next two sections we will show that there also exist differences in the ontogenetic development of abilities based on geometrical and statistical regularities.

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

At the end of the 19^{th} century William James [19] characterized 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, psychopohysical research indicates that the idea of geometric relations in 3D is very likely not a *learned* one but is to a large degree *genetically determined*.

3D information can be acquired by different cues. Cues based on deterministic dependencies are stereo, shape from motion and convergence of two eyes during fixation. Statistical regularities are used by pictorial cues such as occlusion, shading and linear perspective are applied for static depth extraction. It is interesting in which order these cues develop and whether there is a percept of 3D already working from birth on.

Kellman and Arterberry [20] 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'. There 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 than showing steady increase of performance [17] probably caused by 'maturational change in cortical disparity-sensitive units' [20].

The starting of utilizing motion information (as well based on deterministic regularities) for extracting 3D information is not fully clear. Some work indicates that one month old babies already can use motion information to extract depth [30]. In general, it is assumed that 'motion carried information about space appears to operate at the beginning ...' [20]. But is is also stated that 'more study is needed here to uncover the origins ....' [20].

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 genetically some kind of areas reserved in which our geometrical representation of the world is realized in the brain and this representation is multi–sensorial ([2]), i.e., in these maps information of multiple sensors (e.g., vision, sound, touch, ...) is coded.

The most likely conclusion we can draw from these findings is that the concept of depth (realized in reserved genetically determined maps is existing from birth) on but that this idea is first (coarsely) realized by convergence, then in addition by stereo and motion cues. The use of pictorial cues evolves later as discussed in the next section.

4 Evidence for an Adaptive Component in the Ontogenesis of Abilities connected with Statistical Regularities

In contrast to the early use of deterministic interdependencies in depth perception, the use of pictorial cues more likely involves visual experience since these cues are used by 7 months old babies but still not in the 5 months [20]. This has been independently shown for several pictorial depth cues (linear perspective [31], familiar size [42], occlusion [14] and shading [15]).

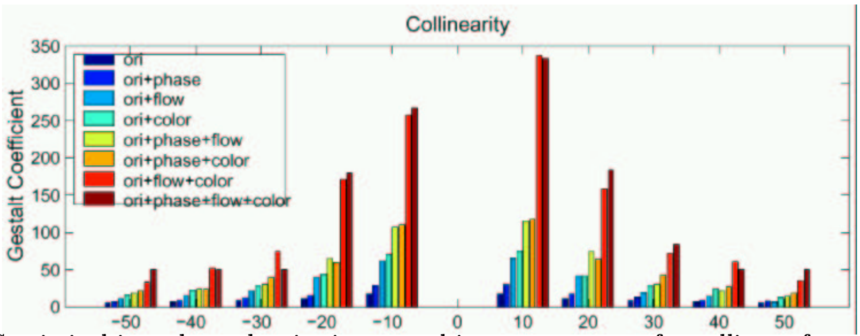

Figure 2: Statistical interdependencies in natural image sequences for collinear features measured by the so called 'Gestalt coefficient' for the modalities phase, optic flow and color and their combinations. The x-axis represents the distance of collinear line segments in pixels. The Gestalt coefficient is a measure for the predictive power of a visual entity. A Gestalt coefficient of 1 means no predictive power, a value larger than 1 expresses the degree of predictive power. As it can be seen from the figure, the predictive power increases tremendously be taking multiple modalities into account (for details, see [27]).

Interestingly, the ability to estimate depth by pictorial cues evolved at the same time as the baby is able to make use of the so called *edge sensitive process* (ESP) [21] for unit–perception, i.e., a process in which oriented line segments are involved (such as, e.g., in the Gestalt law collinearity). At the very same time the occurrence of illusionary contour processing develops [5] (in which the Gestalt law 'collinearity' is also tightly involved).

The late occurrence of the above mentioned mechanisms makes it more likely that those depend on visual experience about the underlying structures in visual data. An interesting question (which to our knowledge has not been addressed so far) is whether the use of pictorial cues starts with a similar fast onset than the use of stereo information. A more gradually development would makes an adaptive mechanism more likely than a sudden maturation process.

Another question is whether the development of pictorial cues depends on our ability to self–navigate, i.e., whether an *active component* is involved. There are two psychophysical results which give contradicting evidence to this assumption. Firstly, the visual and motoric system develop on different time scales, the human visual abilities evolve much faster (some abilities working already at birth and most others maturating during the first 5 months) but motor skills are very limited during the first 5 months. E.g., the ability to self-navigate appears at around 6 months of age. Secondly, the use of pictorial depth information does not depend on crawling competence [3]. However, in other species a correlation between locomotion and space perception has been found [18]. How depth perception does depend on the motion experienced passively has (to our knowledge) not yet been addressed.

Beside the relative late occurrence of the ability to use pictorial cues there is conceptional evidence [24, 9, 11] and evidence from computational neuroscience [34] for an adaptive component in the ontogenesis of the ability to use statistical regularities.

Decades ago, Brunswick and Kamiya [6] first had stated that Gestalt principles should be related to the statistics of the natural world. Unfortunately the limited computational power at this time made it difficult to quantitatively support this statement. Only recently, the strong prevalence of collinearity has been investigated in natural images by [24] and [9] and have been confirmed and extended by [39, 11]. These investigations suggest that Gestalt laws

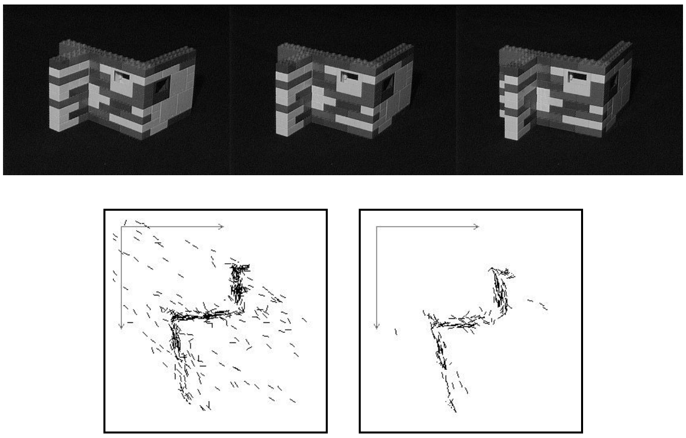

Figure 3: Top: Samples from an image sequence. Bottom: Projected 3D–line segments after first (left) and after 6^{th} iteration (right).

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 2 shows that the probability for two segments being collinear rises if the segments show also similarities in other modalities (for details see [27]). Therefore, there is conceptional evidence for the *possibility* to learn those interdependencies.

Evidence from computational neuroscience has been given by [34] 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, [33] indicates that even more complex feature constellation (i.e., vertices) become significant in the statistics of natural images when segmentation by motion is applied as a pre–processing step.

To sum up this section, there exists evidence from developmental psychology as well as conceptional evidence that the usage of statistical regularities depend on visual experience. Therefore, indirect prediction of 3D from static monocular images is very likely based on the experience of the 3D world. Interestingly pictorial cues are used at the same time when visual control of grasping and crawling evolves but evidence contradicting the necessity of interaction between ego–motion and using pictorial cues in human infants does exist [3].

5 Consequences for the Design of Artificial Visual Systems

As discussed before, the human visual system faces two serious problems: Firstly, it faces the *bias–variance* dilemma and as a consequence thereof in order to be able to learn it already has to know something about its environment in form of *predefined structural constraints.* As a designer of an artificial system we have the awkward task to decide about the structural knowledge we want to use to configure the system. Secondly, it has to deal with a *huge amount of uncertainty* in its low–level modalities but the human has to perform actions with high certainty. It is widely agreed that this precision is achieved by *modality integration.*

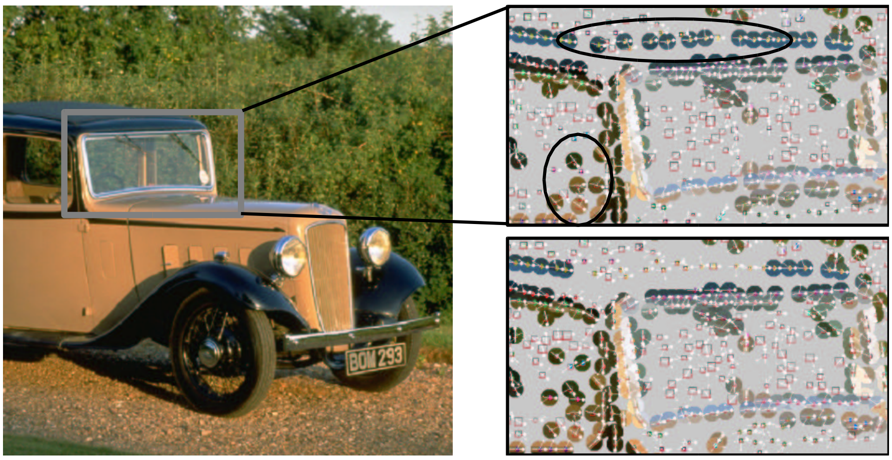

Figure 4: Left: Image of a car. Right top: Extraction of features with grouping based on the Gestalt coefficient. Right bottom: Feature extraction without grouping.

Taking the results described in section 3 into account we think it is also justified to equip an artificial human like system with basic mechanisms for depth extraction from stereo based on geometrical regularities (see, e.g., [22, 26]). Furthermore, we find it justified to equip the system with a basic mechanism to estimate the RBM between frames [35] and an accumulation scheme based om RBM which accumulates stable information from the unreliable stereo information over different time frames [23, 26]. Figure 3 (top) shows object representations extracted from unreliable stereo representations. The representations in Figure 3 (bottom) have been extracted by a mechanisms which stabilises these representations over time making use of RBM (for details see [26]).

Using deterministic relations based on RBM as hardwired components we want to make use of the statistical interdependencies by a mechanisms which relies on *visual experience with real world data*. In [27] we further suggest a mechanism which uses these measured interdependencies to stabilze unreliable visual low-level processes. That means, our grouping mechanism adapts depending on experience based on statistical measurements in real world data. Based on the measured Gestalt coefficient we were able to formulate a grouping procedure in that more complex feature constellations do emerge (see figure 4) and as can be seen in figure 4, the grouping can be used, e.g., for improving feature extraction. Note the detection of the low contrast edge (horizontal ellipse) and the reduction of non meaningful features (more vertical ellipse). More details can be found in [28].

6 Conclusion

It is widely agreed that *there exists no general learning machine for a complex task such as vision*. As a conclusion we have to ask for each visual sub-modality what the necessary *genetical determined structural knowledge* is that we use to be able to learn. However, we may get help *from psychophysical and neurophysiological data* to determine what these structural constraints are that are applied in the human visual system. We claim that this knowledge can be used to establish constraints also for artificial visual systems.

For the sensorial modality vision we can conclude from the psychophysical data that it is likely that in the human visual system *deterministic interdependencies* are to a wide degree genetically coded but that the use of *statistical interdependencies* depend to a certain degree on visual experience. We have further discussed mechanisms that use deterministic and statistical interdependencies within an artificial visual system to reduce the unreliability of visual low–level modalities. The mechanisms based on deterministic regularities were basically hard coded but the mechanism that utilizes statistical interdependencies makes use of visual experience.

Acknowledgement: We would like to thank Michael Felsberg and Gösta Granlund for their critical remarks on the issue this paper deals with.

References

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

[2] 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.

[3] M.E. Arterberry, A. Yonas, and A.S. Bensen. Self–produced locomotion and the development of responsiveness to linear perspective and texture gradients. *Developmental Psychology*, 25:976–982, 1977.

[4] 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.

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

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

[7] R.C.K. Chung and R. Nevatia. Use of monucular groupings and occlusion analysis in a hierarchical stereo system. *CVPR*, 1991.

[8] A. Cozzi and F. Wörgötter. Comvis: A communication framework for computer vision. 2001.

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

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

[11] 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.

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

[13] G.H. Granlund. The complexity of vision. *Signal Processing*, 74, 1999.

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

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

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

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

[18] R. Held and A. Hein. Movement produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, 56:872–876, 1963.

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

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

[21] P.J. Kellman and T.F. Shipley. A theorie of visual interpolation in object perception. *Cognitive Psychology*, 23(2), 1991.

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

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

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

[25] N. Krüger. Learning object representations using a priori constraints within orassyll. *Neural Computation*, 13(2), 2001.

[26] N. Krüger, M. Ackermann, and G. Sommer. Accumulation of object representations utilizing interaction of robot action and perception. *Knowledge Based Systems*, 13(2), 2002.

[27] N. Krüger and F. Wörgötter. Multi modal estimation of collinearity and parallelism in natural image sequences. *Submitted.*

[28] N. Krüger and F. Wörgötter. Self-emergent feature constellations based on statistical interdependencies in visual data. *submitted to DAGM 2002*, 2002.

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

[30] 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.

[31] 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.

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

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

[34] C. Prodöhl, R. Würtz, and C. von der Malsburg. Learning the gestalt rule collinearity from object motion. submitted.

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

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

[37] S. Sarkar and K.L. Boyer. *Special Issue on Perceptual Organization in Computer Vision, October.* CVIU, 1999.

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

[39] 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.

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

[41] D.M. Wuescher and K.L. Boyer. Robust contour decomposition using constant curvature criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991.

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