Published in the proceedings of the workshop Brain, Vision and Artificial Intelligence. First International Symposium 2005, BVAI 2005, Naples, Italy. LNCS3704, Springer. p. 167–176.

Multi-modal Primitives as functional Models of Hyper-columns and their use for contextual Integration

Norbert Krüger[1] and Florentin Wörgötter[2]

[1] Media Lab, Aalborg University Copenhagen, nk@cs.aue.auc.dk
[2] Computational Neuroscience, University of Stirling, worgott@cn.stir.ac.uk

Abstract. In this paper, we describe a biological motivated image representation in terms of local multi–modal primitives. These primitives are functional abstractions of hypercolumns in V1 [13]. The efficient and generic coding of visual information in terms of local symbolic descriptiones allows for a wide range of applications. For example, they have been used to investigate the multi–modal character of Gestalt laws in natural scenes [14], to code a multi–modal stereo matching and to investigate the role of different visual modalities for stereo [11], and to use a combination of stereo and grouping as well as Rigid Body Motion to acquire reliable 3D information as demonstrated in this publication.

1 Introduction

The aim of this work is to compute reliable feature maps from natural scenes. To establish artificial systems that perform reliable actions we need reliable features. These can only be computed through *integration across the spatial and temporal context and across visual modalities* since local feature extraction is necessarily ambigious [1, 15]. In this paper, we describe a new kind of image representation in terms of local multi–modal Primitives (see fig. 1) which can be understood as functional abstractions of hypercolumns in V1. These Primitives can be characterized by three properties:

Multi-modality: Different visual domains describing different structural properties of visual data are well established in human vision and computer vision. For example, a local edge can be analyzed by local feature attributes such as orientation or energy in certain frequency bands. In addition, we can distinguish between line and step–edge like structures (contrast transition). Furthermore, color can be associated to the edge. This image patch also changes in time due to ego-motion or object motion. Therefore time specific features such as a 2D velocity vector (optic flow) can be associated to this image patch. In addition the image patch has a certain source in 3D space and therefore also depth information can be associated. In this work we define local multi–modal Primitives that realize these multi-modal relations. These modalities are also processsed in so called hyper-columns in the first area of visual processing (V1) [7].
Adaptability: Since the interpretation of local image patches in terms of the above mentioned attributes as well as classifications such as ‘edgeness’ or ‘junctionness’ are necessarilly ambigious when based on local processing stable interpretations can only be achieved *through integration* by making use of contextual information [1]. Therefore,

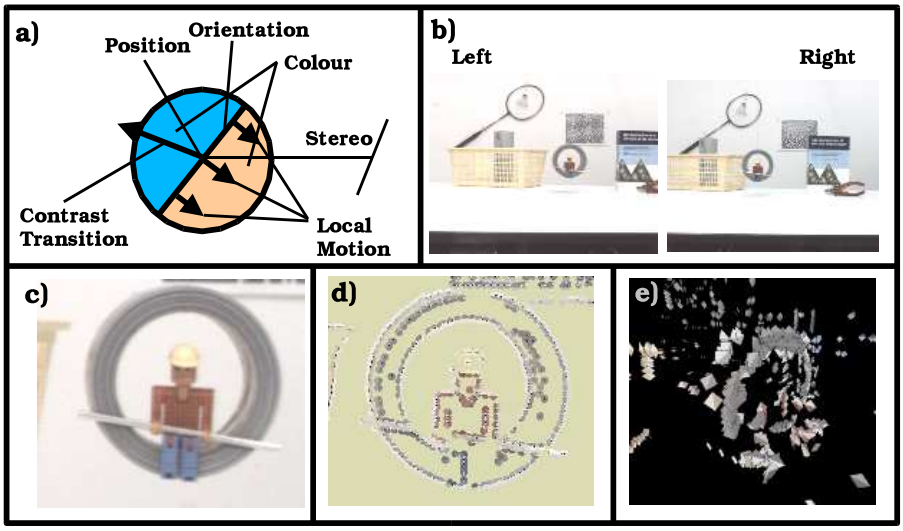

Fig. 1. Multi-modal Primitives **a)** One primitive covers different aspects of visual information in a condensed way. **b)** Stereo Image Pair. **c)** Frame taken from c). **d)** Representation of an image by multi-modal primitives (local motion and stereo information not shown for sake of understandability). **e)** 3D view of extracted stereo representation.

all attributes of our Primitives are equipped with confidences that are essentially *adaptable according to contextual information* expressing the reliability of this attribute. Adaptation occurs by means of recurrent processes (see, e.g., [21]) in which predictions based on statistical and deterministic regularities disambiguate the locally extracted and therefore neceassarily ambigious data.

Condensation: Integration of information requires *communication between Primitives* expressing spatial [14, 11] and temporal dependencies [9]. This communication has necessarily to be paid for with a certain cost. This cost can be reduced by limiting the amount of information transferred from one place to the other, i.e., by reducing the bandwidth. Therefore we are after a *compression* of data. Essentially we only need less than 5% of the amount of the pixel data of a local image patch to code a Primitive that represents such a patch. However, condensation not only means a compression of data since communication and memorization not only require a reduction of information. Moreover, we want to reduce the amount of information within an image patch *while preserving perceptually relevant information*. This leads to *meaningful* descriptors such as our attributes position, orientation, contrast transition, color and optic flow. In [14], we have also shown that these descriptors (in particular when jointly applied) allow for strong mutual prediction that can be related to classical Gestalt laws.

In section 2, we describe the Primitive attributes and their extraction and in section 3 we describe the biological background. In section 4, we refer to applications of our Primitives for the modelling of disambiguation processes in mid-level vision.

2 Multi–modal Primitives

We compute the following semantic attributes and associate them to our Primitives (see also fig. 1).

Intrinsic Dimension: Local patches in natural images can be associated to specific local sub-structures, such as homogeneous patches, edges, corners, or textures. Over the last decades, sub-domains of Computer Vision have extracted and analysed such sub-structures.

The intrinsic dimension (see, e.g., [23]) has proven to be a suitable descriptor that distinguishes such sub-structures. Homogeneous image patches have an intrinsic dimension of zero (i0D); edge-like structures are intrinsically 1-dimensional (i1D) while junctions and most textures have an intrinsic dimension of two (i2D). In [10, 4] it has been shown that the topological structure of intrinsic dimension essentially has the form of a triangle with the corners of the triangle representing 'ideal cases' of homogeneous structures, edges or corners (see figure 2b). This triangular structure can be used to associate 3 confidences $(c_{i0D}, c_{i1D}, c_{i2D})$ to homogenous-ness, edge–ness, or junction–ness according to the positioning of an image patch in the iD–triangle.

This association of confidences to visual attributes is a general design principle in our system. These confidences as well as the attributes themselves are subject to contextual integration via recurrent processes. Aspects with associated low confidences have a minor influence in the recurrent processes or can be disregarded.

Orientation: The local orientation associated to the image patch is described by θ. The computation of the orientation θ is based on a rotation invariant quadrature filter, which is derived from the concept of the *monogenic signal* [5]. Considered in polar coordinates, the monogenic signal performs a *split of identity* [5]: it decomposes an intrinsically one-dimensional signal into intensity information (amplitude), orientation information, and phase information (contrast transition). These features are pointwise mutually orthogonal. The intensity information can be interpreted as an indicator for the likelihood of the presence of a certain structure with a certain orientation and a certain contrast transition (see below).

Contrast transition: The contrast transition is coded in the phase ϕ of the applied filter [5]. The phase codes the local symmetry, for example a bright line on a dark background has phase 0 while a bright/dark edge has phase $-\pi/2$ (see fig. 2a). There exists a whole continuum of i1D structures that can be coded in the phase by one parameter (see also [6, 8]).

Color: Color $(\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r)$ is processed by integrating over image patches in coincidence with their edge structure (i.e., integrating separately over the left and right side of the edge as well as a middle strip in case of a line structure). In case of a boundary edge of a moving object at least the color at one side of the edge is expected to be stable since (in contrast to the phase) it represents a description of the object.

Optic Flow: There exist a large variety of algorithms that compute the local displacement in image sequences. [2] have them devided into 4 classes: differential techniques, region-based matching, energy based methods and phase-based techniques. After some comparison we decided to use the well-known optic flow technique [16]. This allgorithm is a differential technique in which however (in addition to the standard gradient

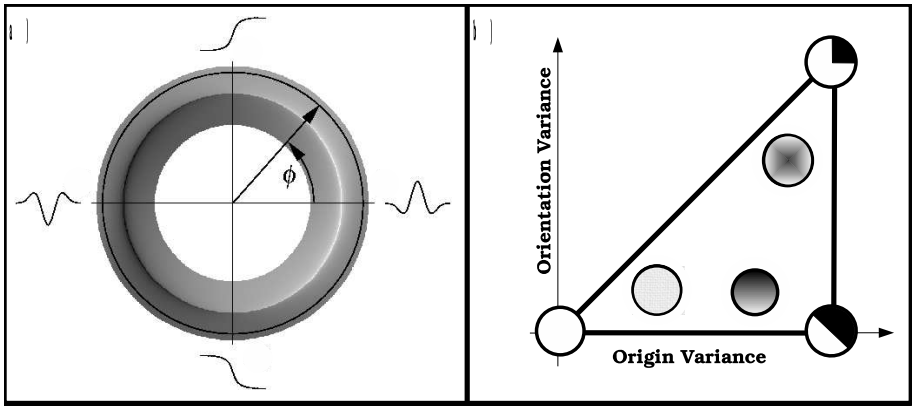

Fig. 2. a) The continuum of phases (indicated by ϕ) taking values between $-\pi$ and π correspond to a continuum of oriented grey-level structures as expressed in the changing circular manifold (sub–figure a) is based on a figure in [3]). **b)** The likelihood of a local image patch to be a homogenous image patch, an edge or a junction can be visualised as a triangle with corners representing ideal patterns. Points inside the triangle represent structures that are only with a certain likelihood categorizable as ideal homogenuous image patches, edges, or junctions. For example, there is a slight texture on the patch close to the lower left corner which produces a filter response with low but measureable magnitude and orientation variance or the structure close to the upper corner has some resemblance to a junction. In this triangular representation distances from the corners represent the likelihood of the structures being of the ideal type. This is used for the formulation of confidences indicating such likelihoods in [10]. Note that figure 2b is thought to be a schematic description. The exact positioning of patches in the triangle depends on two parameters (for details see [10]).

constraint equation) an anisotropic smoothing term leads to better flow estimation at edges (for details see [16]). The optic flow is coded in a vector $\mathbf{o}$.

Stereo: By performing a matching between primitives in the left and right image and finding correspondences we can compute a 3D-primitive (see figure 1e). We code the correspondence by a link l to a primitive in the right image.

To determine the position $\mathbf{x}$ of the primitives we look for locations in the image where the magnitude of the response of a set of edge-detection filters [5] has local maxima. To avoid the occurrence of very close line–segments produced by the same image structure we also model a competition process between the primitives. Basically, for each primitive position it is checked whether another primitive exists with a position closer than a given threshold distance. If that is the case, the position with lower magnitude is dropped (for details see [12, 13]). Finding of suitable positions is a sophisticated task and is also part of a cruicial transformation process from a signal–based to a symbol–based representation. Once the positions of the primitives are determined, the other attributes computed from the filter response at the found position is associated to the primitive.

Usually an image patch that is represented by our Primitives has a dimension of $3 \times 12 \times 12 = 432$ values (3 color values for each pixel in a 12×12 patch). However, the

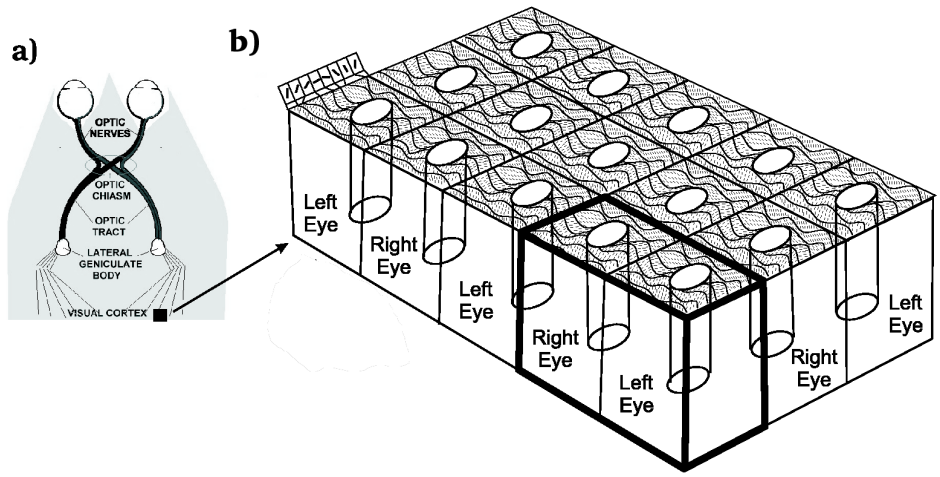

Fig. 3. a) Primary visual pathway and schematic location of a hyper-column (black box), which corresponds in reality to about 1 mm^2 of cortical surface. **b)** Schematic diagram of a hyper-column (thick lines) embedded in the visual cortex. Each hyper-column represents a small location in visual space. Vertically to the surface neurons share similar response properties, whereas their responses differ when moving horizontally on the surface. Information from both eyes is represented in adjacent slabs of the cortex. Each slab contains neurons that encode different orientations (depicted by tiny lines on the surface) but also all other important visual features such as local motion and stereo. In the cylinder-shaped part mainly color is processed. Note, the actual cortical structure is less crystalline than suggested by this diagram.

output of our Primitives has less than 20 parameters. Therefore, the Primitives condense the image information by more than 95%. This condensation is a crucial property of our Primitives that allows to represent meaningful information in a directly accessible and compressed way.

We end up with a parametric description of a Primitive as

$$\pi = (\mathbf{x}, \theta, \phi, (\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r), \mathbf{o}, (c_{i0D}, c_{i1D}, c_{i2D}), l).$$

In addition, there exist confidences $c_i, i \in \{\phi, \mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r, \mathbf{o}\}$ that code the reliabilty of the specific sub–aspects that is also subject to contextual adaptation.

3 Multi-modal Primitives as functional abstractions of hyper-columns

The above–mentioned visual modalities are processed at early stages of visual processing. Hubel and Wiesel [7] investigated the structure of the first stage of cortical processing that is located in an area called ‘striate cortex’ or V1 (see figure 3a). The striate cortex is organized like a continuous, but distorted map of the visual field (retinotopic

map). This map contains a specific repetitively occurring pattern of substructures called hyper-columns. Thus, a hyper-column represents a small location of visual space and the neurons in such a hyper-column represent all important aspects of this spatial location; ideally all orientations, all colors, the complete distance-information (disparity), etc. To be able to achieve this in an orderly manner, hyper-columns themselves are subdivided into "columns" and "blobs". The blobs contain color sensitive cells, while the columns represent the continuum of orientations (see figure 3b). Here one observes that the orientation columns are organized in an ordered way such that neurons representing similar orientations tend to be adjacent to each other. However, it is not only orientation that is processed in an orientation column but the cells are sensitive to additional attributes such as disparity, contrast transition and the direction of local motion (see [22]). Even specific responses to junction–like structures have been measured [19]. Therefore, it is believed that in the striate cortex basic local feature descriptions are processed similar to the feature attributes coded in our primitives.

However, it is not only local image processing that is going on in early visual processing. As mentioned above, there occurs an extensive communication within visual brain areas as well as across these areas. The communication process leads to the binding of groups of local entities (see, e.g., [20]). In [14] we described a self–emergence process in which groups organize themselves based on statistical regularities. Here we use grouping inthe context of improving stereo information.

4 Disambiguation in recurrent process making use of the spatial–temporal context

The processing of primitives is still based on local processes. Therefore, ambiguity can not be resolved at this level. However, using the richness of the image descriptors we can already decrease the amount of ambiguity by interaction of modalities on a local level (section 4.1). Global interdependencies realized in cross–modal recurrent processes based on perceptual organisation and rigid body motion can then further reduce the ambiguity and are described in section 4.2 and 4.3.

4.1 Multi-modal Stereo

To be able to reconstruct 3D primitives we require correspondences between image primitives π^l, π^r in the left and right image of a stereo system. For this we make use of a multi-modal similarity

$$sim(\pi^l, \pi^r) = \sum_{i \in \{o,p,c,f\}} \alpha_i d_i(\pi^l, \pi^r) \tag{1}$$

in which distance measures in the different modalities $d_i()$ are combined by a weighted average (see [11, 17] for details). In table 1, we show the performance of the system on a sequence of images with known ground truth (see figure 4). The results for a stereo with only one modality (orientation), two modalities (orientation and phase) and three modalities (orientation, phase and colour) respectively are displayed in the first column of the left , middle and right block in table 1.

Fig. 4. Left and right image of one frame of the stereo image sequence (left) with 3D-ground truth (right).

	Uni-modal (ori)			Two-modal (ori, pha)			Tri-modal (ori, pha, col)		
Trues	Stereo	Group.	Accum	Stereo	Group.	Accum	Stereo	Group.	Accum
100	1479	1064	8	77	60	6	4	5	2
500	2126	1600	32	346	262	11	19	24	16
1000	2878	n.a.	102	832	586	25	85	78	19
2000	n.a	n.a.	1372	n.a.	n.a.	153	328	278	42

Table 1. The number of false positives depending on four fixed numbers of trues is shown for stereo, grouping and accumulation. The results for uni-modal, two-modal and multi-modal representations are kept separately in the three blocks. n.a. stands for 'not applicable' which means that the number of trues as indicated in the left most column was not achieved.

4.2 Stereo and Grouping

We formalized the spatial constraint indicated in figure 4.2a. Basically the constraint states that stereo correspondences must be consistent under collinear line structures. In [18], we have defined a multi-modal grouping process in which the likelihood of two primitives to be originated from a collinear image structure is coded in two link confidences $g(\pi_1^l, \pi_2^l)$ for the left and $g(\pi_1^r, \pi_2^r)$ for the right image. In combination with (1) we have defined an external similarity that is not based on a direct comparison of image patches but on the consistenny of the stereo with the grouping process only based on the two link confidences $g(\pi_1^l, \pi_2^l), g(\pi_1^r, \pi_2^r)$ and the stereo matching similarity $c(\pi_2^l, \pi_2^r)$. We can use this external similarity to enhance stereo processing. Table 1 (second columns for each block) shows quantitative results. Performance usually incresaes by approximately 20–30 percent.

4.3 Accumulation using a Spatial-Temporal Context based on Rigid Body Motion

A spatial–temporal constrained is based on rigid motion. Assuming the egomotion or the motion of objects between frames is known we can predict the occurence of spatial primitives $\pi(t+1)$ in the next frame. This is possible since knowing the 3D structure

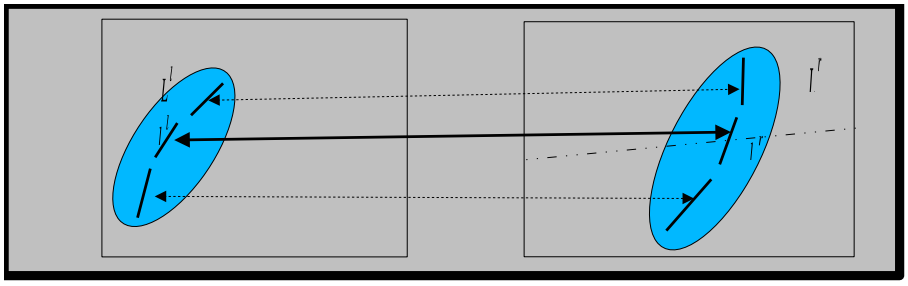

Fig. 5. Top: Stereo–Grouping Constraint.

underling the primitive (as coded in the link l) the spatial-temporal transformation of this primitive can be computed explicitely. The validation of such a correspondences is an indicator for a higher likelihood for the spatial primitive to be a correct one and the associated confidence becomes increased (see also [9]). Table 1 (third column in each block) gives quantitative results. As can be seen from the results even for quite unreliable stereo based on one modality only after only few iterations the number of false positives can be decreased significantly. Note that the scheme also allows for the integration of new hypothese generated in in new frames. In figure 6 the effect for an example sequence is shown.

5 Summary and Conclusion

We have introduced a functional model of hyper-columns in terms of multi-modal primitives representing local image information in a condensed way. This condensation leads to symbol-like descriptors of image information which allows the formalization of cross–modal processes and spatial-temporal integration.

Acknowledgement: We thank Nicolas Pugeault and Sinan Kalkan for their help. Furthermore, we gratefully acknowledge the support of Riegl Ltd. which provided the image data with 3D ground truth shown in figure 4 on which our quantitative evaluation is based.

References

1. Y. Aloimonos and D. Shulman. *Integration of Visual Modules — An extension of the Marr Paradigm*. Academic Press, London, 1989.
2. J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.
3. M. Felsberg. Optical flow estimation from monogenic phase. In B. Jähne, E. Barth, R. Mester, and H. Scharr, editors, *Complex Motion, Proceedings 1st Int. Workshop*, Günzburg, 12.-14.10. 2004.
4. M. Felsberg and N. Krüger. A probablistic definition of intrinsic dimensionality for images. *Pattern Recognition, 24th DAGM Symposium*, 2003.

Fig. 6. Top row: Left: Confidences of different hypothese are displayed by grey level values (white for high confidences and dark for low confidences) projected on the image. Right: Top view of the stereo of the first frame. Bottom row: Left: Image view of all hypothese with high confidence after 5 iterations of the accumulation. Right: Top view of all hypothese with high confidence after five iterations.

5. M. Felsberg and G. Sommer. The monogenic signal. *IEEE Transactions on Signal Processing*, 49(12):3136–3144, December 2001.
6. G. H. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer Academic Publishers, Dordrecht, 1995.
7. D.H. Hubel and T.N. Wiesel. Anatomical demonstration of columns in the monkey striate cortex. *Nature*, 221:747–750, 1969.
8. P. Kovesi. Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3):1–26, 1999.
9. 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.
10. N. Krüger and M. Felsberg. A continuous formulation of intrinsic dimension. *Proceedings of the British Machine Vision Conference*, pages 261–270, 2003.
11. N. Krüger and M. Felsberg. An explicit and compact coding of geometric and structural information applied to stereo matching. *Pattern Recognition Letters*, 25(8):849–863, 2004.

12. N. Krüger, M. Felsberg, and F. Wörgötter. Processing multi-modal primitives from image sequences. *Fourth International ICSC Symposium on ENGINEERING OF INTELLIGENT SYSTEMS*, 2004.
13. N. Krüger, M. Lappe, and F. Wörgötter. Biologically motivated multi-modal processing of visual primitives. *The Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour*, 1(5):417–428, 2004.
14. 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.
15. N. Krüger and F. Wörgötter. Statistical and deterministic regularities: Utilisation of motion and grouping in biological and artificial visual systems. *Advances in Imaging and Electron Physics*, 131, 2004.
16. H.-H. Nagel and W. Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:565–593, 1986.
17. N. Pugeault and N. Krüger. Multi–modal matching applied to stereo. *Proceedings of the BMVC 2003*, pages 271–280, 2003.
18. N. Pugeault, F. Wörgötter, , and N. Krüger. Stereo matching with a contextual confidence based on collinear groups. In *In U. Ilg et al. (eds.), Dynamic Perception, Infix Verlag, St. Augustin.*, 2004.
19. 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.
20. R.J. Watt and W.A. Phillips. The function of dynamic grouping in vision. *Trends in Cognitive Sciences*, 4(12):447–154, 2000.
21. F. Wörgötter, N. Krüger, N. Pugeault, D. Calow, M. Lappe, K. Pauwels, M. Van Hulle, S. Tan, and A. Johnston. Early cognitive vision: Using gestalt-laws for task-dependent, active image-processing. *Natural Computing*, 3(3):293–321, 2004.
22. R.H. Wurtz and E.R. Kandel. Perception of motion, depth and form. In E.R. Kandell, J.H. Schwartz, and T.M. Messel, editors, *Principles of Neural Science (4th edition)*, pages 548–571. 2000.
23. C. Zetzsche and E. Barth. Fundamental limits of linear filters in the visual processing of two dimensional signals. *Vision Research*, 30, 1990.