AISB Journal **1**(4), © SSAISB, 2003
http://www.aisb.org.uk

Biologically Motivated Multi-modal Processing of Visual Primitives

Norbert Krüger*, Markus Lappe† and Florentin Wörgötter‡

* Department of Psychology, University of Stirling, Scotland, UK, *norbert@cn.stir.ac.uk*
† Psychologisches Institut, Universität Münster, Germany, *mlappe@uni-muenster.de*
‡ Department of Psychology, University of Stirling, Scotland, UK, *worgott@cn.stir.ac.uk*

Abstract

We describe a new kind of image representation in terms of local multi–modal Primitives. These Primitives are motivated by processing of the human visual system as well as by functional considerations. We discuss analogies of our representation to human vision and concentrate specifically on the implications of the necessity of communication of information in a complex multi-modal system.

1 Introduction

In this paper, we describe a new kind of image representation in terms of local multi–modal Primitives (see Figure 1). These Primitives are motivated by processing in the human visual system as well as by functional considerations. The work described here has been evolved from a project started in 1998 which has been focused on the integration of visual information (ModIP, 2003). The image representation described here is now a central pillar of the ongoing European project (ECOVISION, 2003) that focuses on the functional modelling of early visual processes.

In the human visual system beside local orientation also other modalities such as colour and optic flow (that are also part of our multi–modal Primitives) are computed in the hyper-columns of V1 (Hubel and Wiesel, 1969; Gazzaniga, 1995). *All these low level processes face the problem of an extremely high degree of vagueness and uncertainty (Aloimonos and Shulman, 1989).* This arises from a couple of factors. Some of them are associated with image acquisition and interpretation: owing to noise in the acquisition process along with the limited resolution of cameras, only erroneous estimates of semantic information (e.g., orientation) are possible. Furthermore, illumination variation heavily influences the measured grey level values and is hard to be modelled analytically (Ikeuchi and Horn, 1981). Information extracted across image frames, e.g., in stereo and optic flow estimation, faces (in addition to the above mentioned problems) the correspondence and aperture problem which interfere in a fundamental and especially difficult way (Ayache, 1990; Klette et al., 1998).

However, the human visual system acquires visual representations which allow for actions with high precision and certainty within the 3D world under rather uncontrolled conditions. *The human visual system can achieve the needed certainty and completeness*

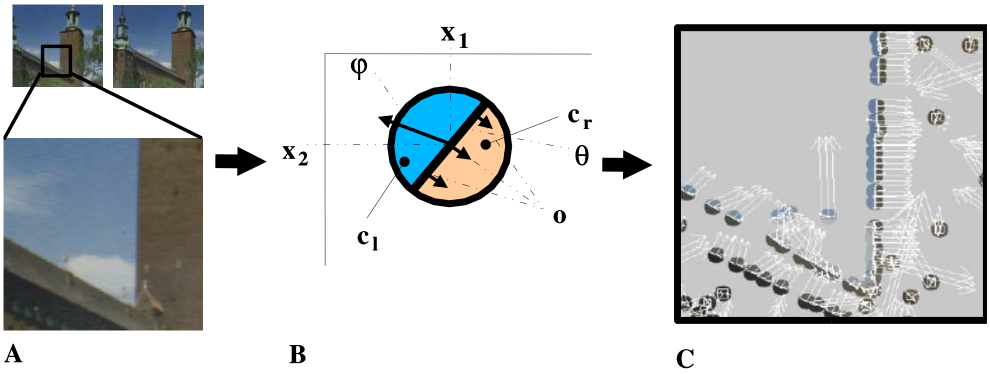

Figure 1: **A:** Image sequence and frame. **B:** Schematic representation of the multi–modal Primitives. **C:** Extracted Primitives at position with high amplitude.

by integrating visual information across modalities (Hibbard et al., 2000) and by utilising spatial and temporal interdependencies (Phillips and Singer, 1997; Hoffman, 1980). This integration is manifested in the huge connectivity between brain areas in which the different visual modalities are processed as well as in the large number of feedback connections from higher to lower cortical areas (Gazzaniga, 1995). The essential need for integrating visual information in addition to optimising single modalities to design efficient artificial visual systems has also been recognised in the computer vision community after a long period of work on improving single modalities (Aloimonos and Shulman, 1989).

However, integration of information makes it necessary that local feature extraction is subject to modification by contextual influences. As a consequence *adaptability* must be an essential property of the visual representation. Moreover, the exchange of information between visual events has necessarily to be paid for with a certain cost. This cost can be reduced by limiting the amount of information transferred from one place to the other, i.e. by reducing the bandwidth. This is the reason why we are after a *condensed* description of a local image patch, which however *preserves the relevant information*. Here relevance has to be understood not only in an information theoretical sense, but in a global sense (the system has to be subject to modifications by global interdependencies, in particular local entities have to be connectable to more complex entities) and action oriented sense (the transfered information has to be relevant for the actions the individual has to perform).

Taking the above mentioned considerations into account, the Primitives, which are the basic entities of our image representation, can be characterised by four properties:

> **Multi-modality:** Different domains that describe different kinds of structures in visual data are well established in human vision and computer vision. For example, a local edge can be analysed by local feature attributes such as orientation or energy in certain frequency bands (Krüger and Sommer, 2002). In addition, we can distinguish between line and step–edge like structures (contrast transition). Furthermore, colour can be associated to the edge. This image patch also changes in time due to ego-motion or object motion. Therefore time specific features such as a 2D velocity vector (optic flow) are associated to our Primitives (see Figure 1).

Adaptability: Since the interpretation of local image patches in terms of the above mentioned attributes as well as classifications such as 'edge–ness' or 'junction–ness' are necessarily ambiguous when based on local processing (Krüger and Felsberg, 2003), stable interpretations can only be achieved *through integration* by making use of contextual information (Aloimonos and Shulman, 1989). Therefore, all attributes of our Primitives are equipped with a confidence that is essentially *adaptable according to contextual information* expressing the reliability of the attribute. Furthermore, feature attributes themselves are subject to correction mechanisms that use contextual information.

Condensation: Integration of information requires *communication between Primitives* expressing spatial (Krüger and Wörgötter, 2002; Krüger et al., 2002b) and temporal dependencies (Krüger et al., 2002a). This communication has necessarily to be paid for with a certain cost (as will be made explicit in section 3). This cost can be reduced by limiting the amount of information transferred from one place to the other, i.e., by reducing the bandwidth. Therefore we are after a *condensed* representation. Also for other tasks it is essential to store information in a *condensed way*, e.g., for the learning of objects to reduce memory requirements.

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

We will describe our feature processing in section 2 and will compare it to early human visual processing in Section 3.

2 Feature Processing and Application

In this section we describe the coding of modalities associated to our Primitives. In addition to the position $\mathbf{x}$, we compute the following semantic attributes and associate them to our Primitives (see also Figure 1).

Frequency: We describe the signal on different frequency levels f independently. Often the decision in which frequency band the relevant information does occur is difficult, therefore we leave this decision open to be decided at later stages of processing. It may be even that for the same position on different frequency levels there occur different kinds of semantic information (for example, the top of the toy in Figure 2A on a high frequency level can be described as texture–like while on a lower frequency level it resembles an edge).

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

Contrast transition: The contrast transition is coded in the phase φ of the applied filter (Felsberg and Sommer, 2001). The phase codes the local symmetry, for example a bright line on a dark background has phase 0 while a bright/dark edge has phase $-\pi/2$ (in Figure 3 the line that marks the border of the street is represented as a line or two edges depending

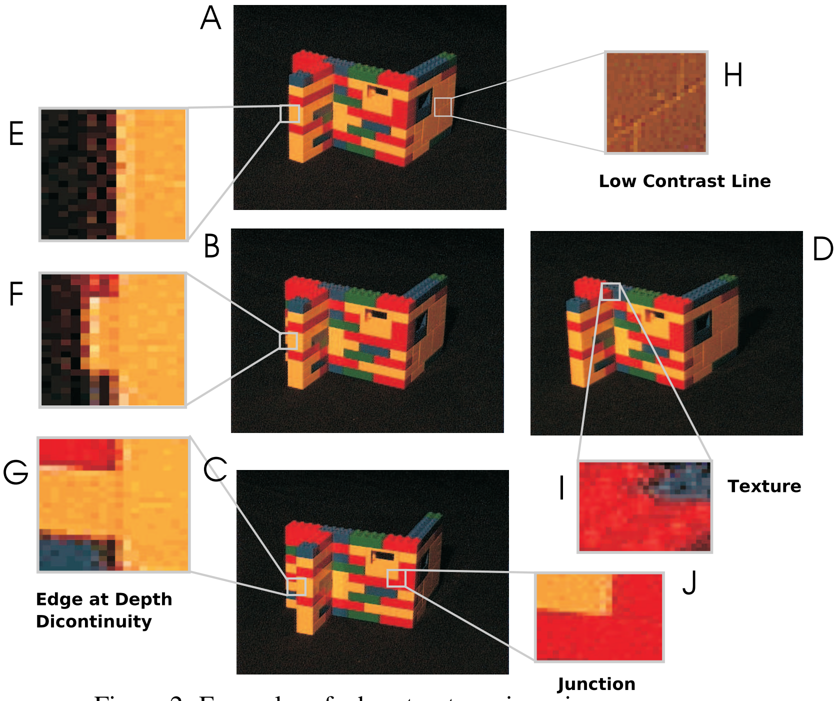

Figure 2: Examples of edge structures in an image sequence.

on the distance from the camera). In case of boundaries of objects, the phase represents a description of the transition between object and background (Kovesi, 1999; Krüger and Wörgötter, 2002).

Colour: Colour $(\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r)$ is processed by integrating over image patches in coincidence with their edge structure (i.e., integrating separately over the left $(\mathbf{c}^l)$ and right $(\mathbf{c}^r)$ side of the edge as well as a middle strip $(\mathbf{c}^m)$ in case of a line structure). In case of a boundary edge of a moving object at least the colour at one side of the edge is expected to be stable (see Figure 2E–G) since it represents a description of the object.

Optic Flow: Local displacements $\mathbf{o}$ is computed by the well known optic flow technique (Nagel, 1987).

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

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

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

We have applied our image representation to different contexts. First, an image patch also describes a certain region of the 3D space and therefore 3D attributes can be associated such as a 3D-position and a 3D-direction. In (Krüger et al., 2002b; Pugeault and Krüger, 2003), we have defined a stereo similarity function that makes use of multiple-modalities to enhance matching performance. Second, the Primitives can be subject to spatial contextual modification. We define groups of Primitives based on a purely statistical criterion in (Krüger and Wörgötter, 2002). Once these groups are defined, we

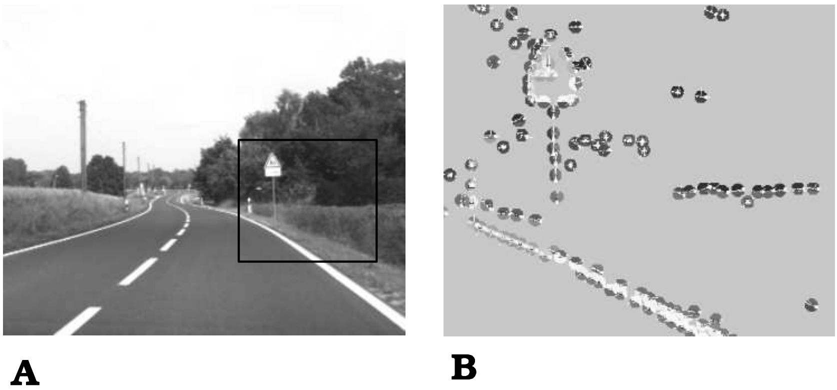

Figure 3: A: Original Image. B: Extracted Primitives with high amplitude.

modulate the confidences of our Primitives: confidences are increased if the Primitives are part of a bigger group, otherwise the confidences are decreased. Thirdly, we have stabilised features according to the temporal context. In (Krüger et al., 2002a; Krüger et al., 2002c), we make use of the motion of an object to predict feature occurrences and showed that we can stabilise stereo processing by modifying the confidences according to the temporal context.

3 Hyper-columns of Basic Processing Units in early Vision

In this section, we discuss aspects of the processing of visual information in the human visual system and draw analogies to our image representation.

The main stream of visual information in the human visual system goes from the two eyes to the LGN (Lateral Geniculate Nucleus) and then to area V1 in the cortex (see Figure 4 and (Wurtz and Kandel, 2000a)). There are two kinds of cell types involved (M (magnocellular) and P (parvocellular) cells) that have different response characteristics: M cells have a low spatial but high temporal resolution and are not colour sensitive. In contrast to M cells, P cells have a low temporal and high spatial resolution and are colour sensitive. Both kinds of cells project into two cortical pathways, the dorsal and ventral pathway (see Figure 4). The ventral pathway goes from the cortical area V1 to V2 to the Inferior Temporal Area (IT) and is believed to be mainly responsible for object recognition (Tanaka, 1993). In the dorsal stream information is transferred from V1 to MT (Middle Temporal Area) to MST (Medial Superior Temporal Area) and is believed to be involved in the analysis of motion and spatial information.

V1 (or Visual Area 1) is the main input of both pathways. The structure of V1 has been investigated by Hubel and Wiesel in their ground-breaking work (Hubel and Wiesel, 1962; Hubel and Wiesel, 1969). V1 is organised in a retinotopic map that has a specific repetitively occurring pattern of substructures called hyper-columns. Hyper-columns themselves contain so called orientation columns and blobs (see Figure 5). The main input of V1 comes from the LGN and targets to layer 4 to which information of both eyes projects (see Figure 5Aiii).

The orientation columns are organised in an ordered way such that columns representing similar orientations tend to be adjacent to each other (see Figure 5Ai). However, it is not only orientation that is processed in an orientation column but the cells are sensitive

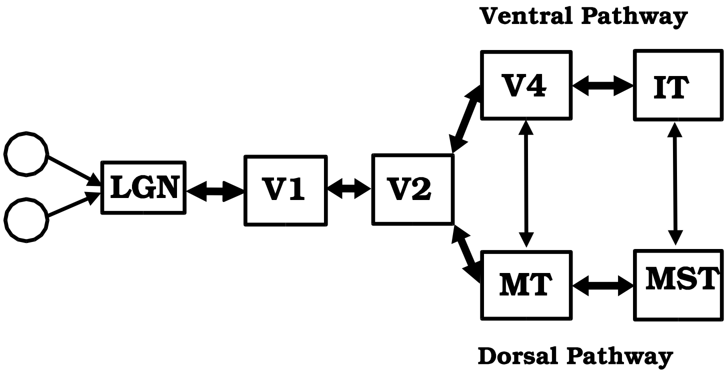

Figure 4: Flow of visual information in the human visual system (schematic).

to additional attributes (see Figure 5D) such as disparity (Barlow et al., 1967; Parker and Cumming, 2001), local motion (Wurtz and Kandel, 2000b), colour (Hubel and Wiesel, 1969) and phase (Jones and Palmer, 1987). Also specific responses to junction–like structures could be measured (Shevelev et al., 1995). Therefore, it is believed that in V1 basic local feature descriptions are processed similar to the feature attributes coded in our Primitives. However, since the processing is local,[1] the ambiguities of visual information is not resolved at this level. For example, response properties of neurons in V1 reflect the aperture problem (Stumpf, 1911). This holds also for our Primitives since the flow is also computed by a local operation.

It is believed that mainly form is processed in the ventral pathway. Neurophysiological equivalents of illusionary contours can be detected in V2 but not in V1 (von der Heydt et al., 1984). This is not surprising since illusionary contours like in the Kanizsa triangle (Kanizsa, 1976) presuppose an integration of information across a large spatial domain as well as across different feature types (e.g., edges and junctions) and can therefore only be processed at a later stage.

The different visual modalities are not computed independently but are combined. For example in V1 the processing of motion is necessarily intertwined with the processing of orientation because of the aperture problem. In V4, colour and orientation is combined (Wurtz and Kandel, 2000b). Accordingly, in our image representation the coding of colour is deeply intertwined with the coding of orientation. Colour is a feature that describes homogeneous surfaces. However, orientation describes discontinuities and can be used to separate the surfaces. In our image representation we therefore first compute orientation and then compute a left and a right colour according to this orientation.

In the dorsal pathway mainly motion is analysed. Like the occurrence of illusionary contours presuppose global interactions, the aperture problem can only be solved by taking the global context into account. This does not happen (and can not happen because of the local processing) in V1. However, in MT and MST many cell responses indicate a solution to aperture problem (Pack and Born, 2001; Wurtz and Kandel, 2000b). Similar to the cells in V1, our Primitives also reflect the aperture problem. However, we can use the output of our Primitives to apply global mechanisms that disambiguate the local flow.

As in the ventral pathway, cells in the dorsal pathway show multi-modal response

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

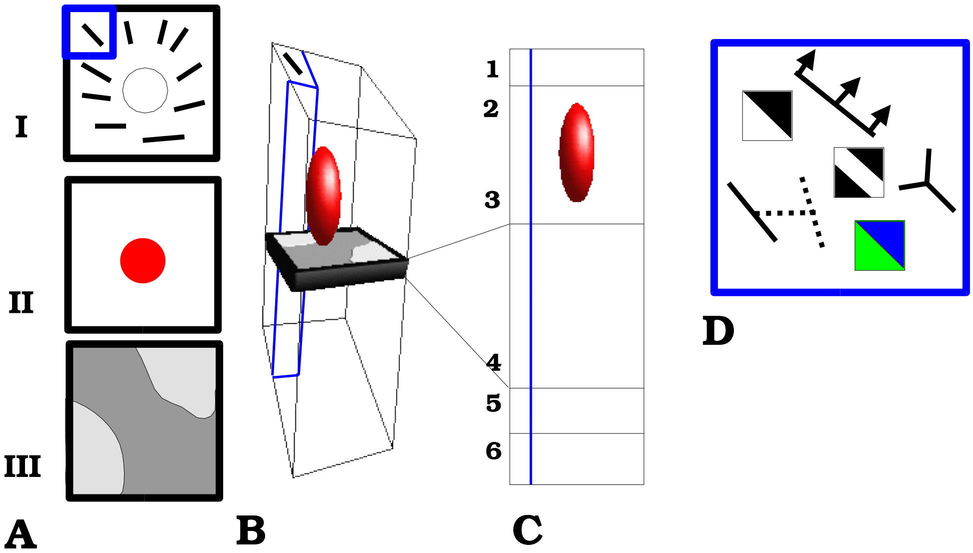

Figure 5: Hyper-columns in V1. A: There exist three physiological distinguishable substructure in a hyper-column: (i) in orientation columns information about oriented edge structure is represented in a topological way. (ii) Colour information is coded in so called 'blobs'. (iii) Information of both eyes are input to the fourth cortical layer (see also B). B: three–dimensional structure of a hyper-column. C: organisation in cortical layers. D: feature attributes that are coded in a hyper-column.

patterns. For example, a moving edge may not be visible as a luminance edge but can be constituted by colour or texture. MT cells respond to these kinds of structures although they are not sensitive to colour alone (Thiele et al., 1999; Wurtz and Kandel, 2000b).

Let us summarise. In V1 visual information is mainly locally processed. However, some semi–local interactions exist. The ambiguities of visual information can not be resolved at this stage of processing. A specialisation to form processing (along the ventral pathway V1–V2–V4–IT) and motion processing (along the dorsal pathway V1–V2–MT–MST) does occur.

As mentioned above, stable and reliable information can only be achieved by disambiguation through integration. However, this integration process makes the exchange of information within and across visual areas mandatory. As discussed before, intra–areal connections are very limited. However, inter–areal connection project to a much wider field of the next layer.

Regarding communication between visual areas we have to address two issues:

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

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

The first question leads to a reflection about costs of communication. In any communication system transfer of information is associated to a cost which normally increases with the amount of information to be transferred and with the distance to be covered. This could concern the costs of "cables" but also the cost of the energy used for the transfer (Attwell and Laughlin, 2001).

In the brain, the communication between two neurons is realized by an axon docking to the soma or the dendrites of other neurons. Accordingly, the complexity and, thus, the "cost" of communication increases with the number of connections. This holds in a very general sense and may have been one driving force for the bandwidth reduction that is actually observed in neuronal visual processing. This bandwidth reduction most clearly manifests itself in mechanisms of visual attention and visual awareness. Focused attention is often taken as one central mechanisms used to reduce the bandwidth of computation as well as of information transfer in the brain to a manageable degree. Anatomically the bandwidth limitation requirement may be reflected by the density of fibres which connect different areas which is smaller than that which connects cells within a hyper-column.

A similar mechanism is also used in our image representation were we arrive at a significant reduction of information following the first processing stages. Compared to an average sized image patch of 15×15 pixels represented by a Primitive the output of a Primitive has less than 20 values, i.e., we have a compression rate of more than 96%. This rate becomes even higher when we compare the output of a Primitive to intermediate local stages of processing where feature attributes for all modalities are derived for each pixel.

The second question above concerns the quality of information which needs to be transferred between the different stages of visual processing. Here we refer back to what we have said above noting that pre-processed visual information is exceedingly ambiguous as the consequence of fundamental problems in image data acquisition as well as resulting from the intrinsic structure of the detectors (receptive fields). This leads to the situation that redundant information must be transferred because only through redundancy it can be assured that erroneous information can be disambiguated. For this it is required that a visual event which is represented by the firing of neuron A has a relevance for the event represented by B. Since event A is supposed to be used to correct event B both events need to be highly correlated. This can be quantified by the following measure of statistical interdependencies:

$$\frac{P(B|A)}{P(B)}. \tag{1}$$

If this term takes a high value then there is a high likelihood of the occurrence of event B when we know event A has occurred compared to the likelihood of the occurrence of the event B without prior knowledge. In this case, events A and B can be used to mutually correct each other because they are carrying shared (i.e., redundant) information. The expression (1) has been called 'Gestalt coefficient' in (Krüger, 1998) where it was shown that applying binarised Gabor wavelets to natural images, a high Gestalt coefficient corresponds to the Gestalt laws Collinearity and Parallelism. As an extension of (Krüger, 1998), it has been shown in (Krüger and Wörgötter, 2002) that by using our multi–modal Primitives we can increase the statistical interdependencies measured by (1) significantly compared to using orientation only (Krüger, 1998). That means that by using our Primitives we can increase interdependencies of visual events. In this way in our Primitives not only information is condensed but transferred to *more meaningful descriptors*.

4 Conclusion

We have introduced a novel way to compute visual Primitives which are motivated by early processing in the human visual system in analogy to the output of V1 hyper-columns. These Primitives are multi-modal and give a dense and meaningful description of a scene. Our Primitives can adapt according to the spatial and temporal context that is realized in the human visual system through a high synaptic connectivity. In this way the locally unreliable feature extraction can be disambiguated and stable feature representations can be achieved.

References

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

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

Ayache, N. (1990). *Stereovision and Sensor Fusion*. MIT Press.

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

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

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

Gazzaniga, M. (1995). *The Cognitive Neuroscience*. MIT Press.

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

Hoffman, D., editor (1980). *Visual Intelligence: How we create what we see*. W.W. Norton and Company.

Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Phiyology*, 160:106–154.

Hubel, D. and Wiesel, T. (1969). Anatomical demonstration of columns in the monkey striate cortex. *Nature*, 221:747–750.

Ikeuchi, K. and Horn, B. (1981). Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, 17:141–184.

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

Kanizsa, G. (1976). Subjective contours. *Scientific American*.

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

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

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

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

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

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

Krüger, N., Jäger, T., and Perwass, C. (2002c). Extraction of object representations from stereo imagesequences utilizing statistical and deterministic regularities in visual data. *DAGM Workshop on Cognitive Vision*, pages 92–100.

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

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

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

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

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

Parker, A. and Cumming, B. (2001). Cortical mechanisms of binocular stereoscopic vision. *Prog Brain Res*, 134:205–16.

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

Pugeault, N. and Krüger, N. (2003). Multi–modal matching applied to stereo. *Proceedings of the BMVC 2003*.

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

Stumpf, P. (1911). über die abhängigkeit der visuellen bewegungsrichtung und negativen nachbildes von den reizvorgangen auf der netzhaut. *Zeitschrift fur Psychologie*, 59:321–330.

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science*, 262:685–688.

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

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

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

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