ulm university universität uulm

# Neural mechanisms of feature extraction for the analysis of shape and behavioral patterns

**Dissertation zur Erlangung des Doktorgrades (Dr.rer.nat.)**
der Fakultät für Informatik und Ingenieurwissenschaften
der Universität Ulm,

vorgelegt von

**Ulrich Weidenbacher**

aus Heidenheim a.d. Brenz

2010

Universität Ulm
Institut für Neuroinformatik

| | |
|---|---|
| Amtierender Dekan: | **Prof. Dr. Michael Weber** |
| Erstgutachter: | **Prof. Dr. Heiko Neumann** |
| Zweitgutachter: | **Prof. Dr. Günther Palm** |
| Tag der Promotion: | **7.12.2010** |

# Abstract

The human visual system segments 3D scenes in surfaces and objects which can appear at different depths with respect to the observer. The projection from 3D to 2D leads partially to occlusions of objects depending on their position in depth. There is experimental evidence that surface-based features such as occluding contours or junctions are used as cues for the robust segmentation of surfaces. These features are characterized by their robustness against variations of illumination and small changes in viewpoint. We demonstrate that this feature representation can be used to extract a sketch-like representation of salient features that captures and emphasizes perceptually relevant regions on objects and surfaces. Furthermore, this representation is also suitable for learning more complex form patterns such as faces and bodies in different pose.

In this thesis, we present a biologically inspired, recurrent model which extracts and interprets surface-based features from a 2D grayscale intensity input image. Based on the neurophysiology of the primate brain, the model is based on few basic processing mechanisms which are reused at several model stages with different parameterization. Furthermore, the architecture is characterized by feedforward and feedback connections which lead to temporal dynamics of model activities. The model simulates the two main processing streams of the primate visual system, namely the form (ventral) and the motion (dorsal) pathway. In the model ventral pathway prototypical views of head and body poses (snapshots) as well as their temporal appearances were learned unsupervised in a two-layer network. In the dorsal pathway prototypical velocity patterns are generated by local motion detectors. These learned patterns are combined into typical motion patterns appearing from head and body movements during establishment of visual contact. Activity from both pathways is finally integrated to extract a combined signal from motion and form features. Based on these initial feature representation we demonstrate a multi-layered learning scheme that is capable of learning form and motion features utilized for the detection of specific behaviorally relevant motion patterns (e.g. turn away and turn towards actions of the body). We show that the combined representation of form and motion features is superior compared to single pathway based model approaches.

# Acknowledgements

I would like to thank my supervisor, Prof. Dr. Heiko Neumann, who supported me and my work during all the years I spent in his group. I will miss the daily coffee breaks which were a good opportunity to discuss new ideas or to get feedback on new results. He also gave me the chance to regularly visit international conferences where I could present my work and discuss with other researcher from different countries. I would also like to thank Prof. Dr. Günther Palm, who kindly accepted to write the second expert's report. I also want to thank Dr. Pierre Bayerl who supported me with his experience as a researcher in the first two years of my thesis. I thank all other members of the Vision group for valuable discussions and also for the time in numerous social events that we spend together. Finally, I would like to appreciate my wife Heike who had to bear the long days and evenings alone with our daughter Hannah while I was working on this thesis.

# Contents

# Chapter 1

# Introduction

## 1.1 Motivation

The ability to *see* is one of the most fundamental skills of our species to interact with our environment. The perception and correct interpretation of an ordinary scene which consists, for instance, of some people and some objects requires robust processing mechanisms that must be able to handle image variations caused by different illumination, partly occlusion, size, and different viewpoint. Moreover, our visual system must be able to also recognize very subtle details such as recognizing different faces or correctly interpreting different face expressions. Thus, in many situation when environmental conditions are not clearly predefined human vision still outperforms most of the computational visions system of today.

Over the last 50 years, neuroscientists have tried to unravel what are the underlying basic mechanisms that make the primate brain such a powerful information processing system. With the advent of computers, *Computational Neuroscience,* a new interdisciplinary research field has emerged that links neuroscience, cognitive science and psychology with electrical engineering, computer science, mathematics and physics. Computational models are developed to simulate detailed circuits of the cortex organized across several cortical areas (e.g. of the visual system). These computational models can 1) help to summarize and organize existing data, 2) help planning, coordinating and interpreting new experiments, and 3) help to get inspirations from biology on how to build robust image processing algorithms for technical applications, for example, driver-assistance systems, video surveillance, or face/smile detection in digital cameras.

Apart from that, engineers and computer scientists have developed application-based algorithms for image processing tasks that are not related to a biological background. In this research field, referred to as *machine vision* most algorithms are tuned to be highly efficient for a specific task (e.g., face recognition) and processing time is often fast enough to reach real time performance. However, a major drawback of these approaches is that they often do not generalize to work in different domains. By contrast the human visual system is highly adaptive (i.e. invariant against illumination, contrast polarity, object size and rotation, and robust against image noise). Hence, there are many examples where

humans vision outperforms machine vision.

## 1.2 Biological background

In this section, we present a brief description of the individual processing stages in the primate visual cortex.

The cortex can be divided into different functional regions, denoted as cortical areas. Each area contains layers of neurons building a neural network of feedforward and feedback and lateral connections [Fellemann and van Essen, 1991]. In early visual areas (e.g., V1 through V5) the neurons have a *retinotopic organization* in the sense that they form a 2D representation of the visual image formed on the retina in such a way that neighboring regions of the image are represented by neighboring regions of the visual area. However, the retinotopic representation in cortical areas is distorted. For instance, the foveal area is represented by a larger number of neurons in V1 than the peripheral areas. An overview of the hierarchical structure in the ventral pathway, the signal flow between cortical areas, and corresponding receptive filed sizes is given in Fig. 1.1.

#### From retina to primary visual cortex

Visual processing starts at the retina, where light passes across different layers to hit the photoreceptors. This elicits chemical transformation mediating a propagation of signal to the bipolar and horizontal cells. The signal is then propagated to the amacrine and ganglion cells. These neurons may ultimately produce action potentials on their axons. The signals are further relayed through the optic nerve, the chiasm, and the LGN (lateral genicualte nucleus) to the first cortical stage of visual processing, the *primary visual cortex* (V1) (e.g., [Nolte, 2002]).

### Two visual pathways

Processing of visual information in primates is believed to occur in at least two separate cortical pathways, commonly labeled the "form" and "motion" pathways [Mishkin *et al.*, 1983; Fellemann and van Essen, 1991]. This division lies in marked contrast to our everyday visual experience, in which we have a unified percept of both the form and motion of objects.

In the motion pathway, local motion information is represented by pools of cells that are tuned to different speeds and directions, thus describing spatiotemporal patterns of local image structures [Movshon *et al.*, 1985; Smith and Snowden, 1994]. In the *medial temporal cortex* (MT) retinotopically arranged cells receive projections from V1 and also project back to this area [Albright, 1984]. Importantly, this indicates that processing in the visual cortex is not purely feedforward. The existence of massive cortical feedback projections leads to a bidirectional signal flow between cortical areas [Sillito *et al.*, 2006]. One possible advantage of such an architecture is that feedback from higher motion areas can influence the transfer of ascending input when, or even before, the input arrives.

The *medial superior temporal cortex* (MST) includes cells that are responsive for motion patterns such as global rotation or translation [Duffy and Wurtz, 1991; Graziano *et al.*, 1994] and opponent motion [Saito, 1993]. Cells in area MST show substantial position and scale invariance.

In the form pathway, processing starts with the extraction of local oriented contrasts by simple and complex cells in area V1 [Hubel and Wiesel, 1968]. These cells are linked to neurons in areas V2 and V4 which are selective for more complex features similar to elongated contours, corners, and junctions [Hedgé and van Essen, 2000; Anzai *et al.*, 2007; Peterhans, 1997; Pasupathy and Conner, 1999]. For instance in V2, like-oriented contrasts that lie along smooth contours are grouped by long-range lateral connections. Interestingly, these cells also respond to *illusory contours* bridging the gap between like-oriented contour fragments [von der Heydt *et al.*, 1984].

The *inferotemporal cortex* (IT) receives input from invariant feature detectors of the previous hierarchy level (V2 and V4). They show substantial position- and scale-invariance and are selective for complex shapes [Tanaka, 1996] and can become tuned to complex shapes through learning [Logothetis *et al.*, 1995]. These authors have found that monkey IT neurons represent a series of two-dimensional views of objects and faces which is important for the recognition of a three-dimensional object from different viewpoints.

## Integration of visual pathways

In our visual perception there is no ambiguity about which objects in a scene are moving and which are static. This is surprising given neurophysiological evidence, from both monkeys and humans, suggesting that visual processing is performed by different functional pathways [Mishkin *et al.*, 1983]. This implies that both pathways, the form and the motion pathway, are integrated at a certain cortical processing stage. However, it is still an open question *how* and *where* in the visual cortex information from both pathways is combined. To address this question, neurophysiologists have searched for cortical areas that show selectivity for both types of features. In fact, studies in macaque monkeys have shown that neural populations in the anterior part of the *superior temporal sulcus* (STSa) are sensitive to both, form and motion [Oram and Perrett, 1996]. The STS has been identified as the primary area involved in the perception of biological motion. In particular, neurons in the STSa respond selectively to full-body [Oram and Perrett, 1994b; Perrett *et al.*, 1985a]or hand movements [Perrett *et al.*, 1989]. On the other hand, there is evidence that populations of cells in STSa are tuned to multiple views of the same animate object [Perrett *et al.*, 1985b]. Such response selectivity is most likely obtained through pooling of outputs of cells coding for separate views of distinct stimuli. In summary, these findings suggest that STS cells integrate information about form and motion of animate objects.

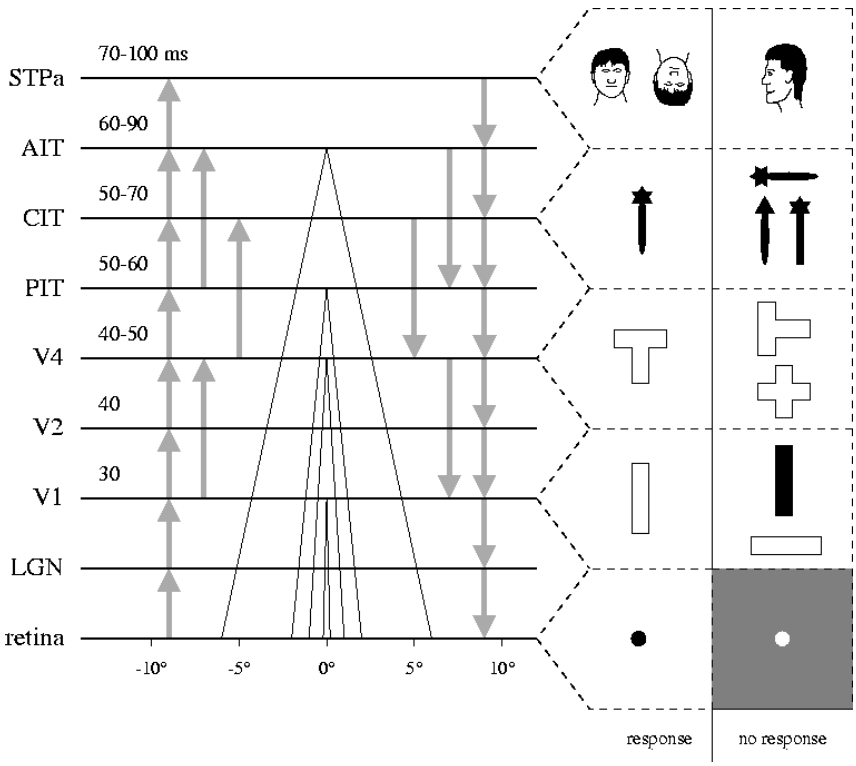


Figure 1.1: Hierarchical structure of the ventral pathway (adapted from [Oram and Perrett, 1994a]). The pathway has a layered structure (from retina to STPa) with layer to layer connections (gray upward arrows). Forward connections are usually mirrored by feedback connections (gray downward arrows). Neurons in different layers respond to features of increasing complexity from bottom to top (sample stimuli to the right). Receptive field sizes and with it the shift invariance also increase from bottom to top (triangles in the center, tip indicating a neuron and base indicating its receptive field size). Response latencies of neurons in individual stages are shown on the left.

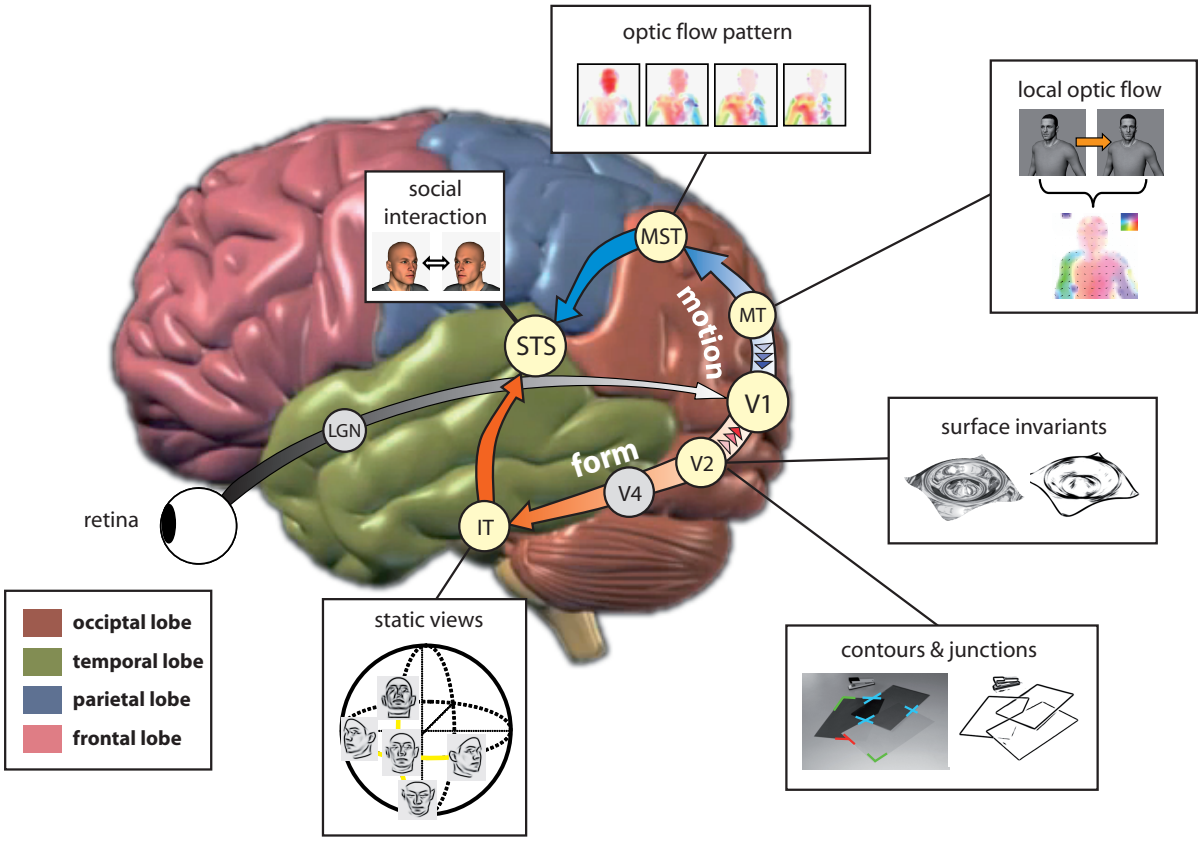


Figure 1.2: Outline of cortical processing pathways that are modeled in this thesis. Visual input enters the eye and is absorbed by the retina. Visual information then travels through the lateral geniculate nucleus (LGN) to the primary visual cortex (V1) which forms the basis of two different processing streams. The form pathways starts in area V1 with the extraction of local oriented contrast and is continued in area V2 with the representation of invariant form features such as contour fragments and junctions. Additional visual areas such as V4 are also involved in form processing but are not considered in detail in this thesis. In the inferotemproal cortex (IT) cells respond selectively to different views of objects and faces. The dorsal stream primarily deals with processing of motion information. Local motion features are extracted in V1 and are further integrated by cells in medial temporal cortex (MT). In the medial superior temporal cortex (MST) cells specifically respond to optic flow patterns. There is physiological evidence that the superior temporal sulcus (STS) receives projections from both form and motion pathways to analyze motion patterns that are relevant in social interaction.

## 1.3 Neural modelling

Modeling and simulation of interactions between pools of neurons and the neurons themselves provides a powerful tool to analyze how information is represented and processed in the human brain and central nervous system. In the field of neuroscience, neurons are often identified as groups of neurons that perform a specific physiological function.

A single neuron may be connected to many other neurons. The connections, called *synapses*, are usually formed from *axons* and *dendrites*. Modeling can be performed at the individual level of neurons, for instance, modeling the spike[1] response curves of neurons to a stimulus (*single compartment model*) or at the a more detailed level of several parts (compartments) of a neuron (*multi compartment model*). A simple model that describes the membrane potential of a single neuron is the integrate-and-fire model [Abbott, 1999]. This model basically assumes that action potentials are simply spikes occurring when the membrane potential reaches a threshold. After firing membrane potential is reset. This simplifies the modeling dramatically as we only deal with sub-threshold membrane potential dynamics. More detailed and complex models of single spiking neurons with non-linear dynamics are the Izhikevich model [Izhikevich, 2003] or Hodgkin-Huxley Model [Hudgkin and Huxley, 1952].

Moreover, we can differentiate between two different types of models:

- *Spiking models* are based on spiking neurons which are utilized to encode information by means of the frequency [Rieke *et al.*, 1996], temporal order [Thorpe, 1990] or synchronicity [Singer, 1999] of generated spikes. These models operate on a fine temporal resolution.

- *Firing-rate models* on the other hand are more abstract and do not encode individual spikes of neurons. Instead for each neuron only the mean spike rate or the spike frequency is represented. In general, this type of model requires less resources for computational simulations while yielding a realistic computational behavior on the level of networks and computational maps [Koch, 1999].

In this thesis we use a firing-rate model as we model thousands of neurons in our simulations. The *model structure* connects the model neuron to other neurons (feedforward, lateral, and feedback connections are modeled) and the *model dynamics* is defined by differential equations that describe the temporal change of neural activity or by steady state solutions of these equations.

[1] An action potential or a spike is a short-lasting electrical potential that is generated if a neuron is excited sufficiently to reach a threshold. The action potential then travels along the axon through synapses to other neurons.

## 1.4 Learning and Plasticity

### 1.4.1 Biological relevance

Activity-dependent synaptic plasticity provides the basis for most models of learning, memory and development in neural circuits. The functional and behavioral role of synaptic plasticity can be investigated by studying how experience and training changes synaptic strength, and how these modifications change neural firing patterns to affect behavior (see [Dayan and Abbott, 2001] for a comprehensive overview).

Experimental investigations have revealed mechanisms of how neural activity can affect synaptic strength. Furthermore, empirically inspired synaptic plasticity rules have been employed in several fields including auto- and heteroassociative memory, pattern recognition, function approximation, and recall of temporal sequences.

One of the most influential synaptic plasticity rules was introduced by Donald Hebb [Hebb, 1949]. It is called *Hebb's rule* and states that if input from neuron A frequently contributes to the firing of neuron B, then the synapse between A and B should be strengthened. The theory is often summarized as "*cells that fire together wire together*".

Experimental work in a number of brain regions have revealed activity-dependent processes that can produce changes in the efficacies of synapses that persist for varying amounts of time.

Changes that persist for tens of minutes or longer are generally called *long-term potentiation* (LTP) and *long-term depression* (LTD). In Fig. 1.3 experimental results with a hippocampal slice of the rat illustrated that demonstrate the long-lasting effect of persistent high-frequency stimulation induced potentiation and low-frequency induced depression.

Bienenstock, Cooper and Munro [Bienenstock *et al.*, 1982] suggested one such mechanism. In the BCM model, correlated pre- and postsynaptic activity evokes LTP when the postsynaptic firing rate is higher than a threshold value and LTD when it is lower. To stabilize the model, the threshold shifts or slides as a function of the average postsynaptic firing rate. For example, the threshold increases if the postsynaptic neuron is highly active, making LTP more difficult and LTD easier to induce. Although this idea is attractive as a computational model, experimental evidence for the sliding threshold is largely indirect [Abraham, 1997].

More recently several studies (e.g., [Markram *et al.*, 1997]) have revealed that the change in synaptic strength also depends on the precise timing of action potentials in connected neurons, referred to as *spike-timing dependent synaptic plasticity* (STDP). In particular, the change in synaptic efficacy is different if A) the presynaptic spike precedes the postsynaptic spike or B) the postsynaptic spike precedes the presynaptic spike. In general, the first case (A) produces LTP and the latter case (B) leads to LTD. Thus, input that contributes to the firing of a cell strengthens the connection of a synapse while input that follows a spike weakens the synaptic connection.

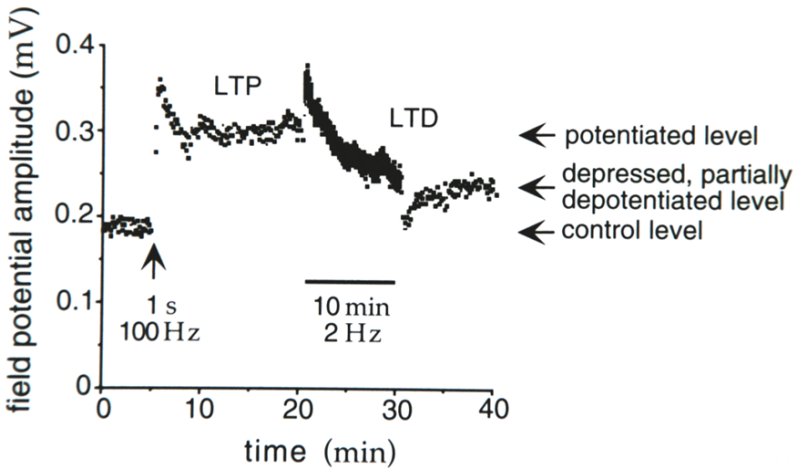


Figure 1.3: LTP and LTD observed in an experiment on a rat hippocampal slice (CA1). The points show the amplitudes of field potentials evoked by constant amplitude stimulation. Stimulation at 100 Hz for 1 s caused significant increase in the response amplitude. Next stimulation at 2 Hz was applied for 10 min. This reduced the amplitude of the response. After a transient dip, the response amplitude remained at a reduced level approximately between the original and post-LTP values, indicating LTD (data of J. Fitzpatrick and J. Lisman published in [Dayan and Abbott, 2001]).

### 1.4.2 Supervised learning

The main characteristic of *supervised learning* is that an external teacher signal is involved in the process of learning. During training, an explicit teacher-signal is imposed to the network including a set of input-output relationships. Supervised learning methods are not biologically plausible as the desired output of the network has to be provided in advance. Therefore, these learning mechanisms are predominantly employed in artificial neural networks in the field of machine learning. The most frequently used supervised learning methods are *radial basis function networks* (RBFs) [Moody and Darken, 1989], *multi-layer perceptrons* (MLPs) [Cybenko, 1989], *support vector machines* (SVMs) [Cristianini and Shawe-Taylor, 2000].

### 1.4.3 Unsupervised learning

In *unsupervised learning*, a network learns to respond to a series of inputs without a given teacher signal. Here, the targets are the same as the inputs. In other words, unsupervised learning usually performs the same task as an *autoassociative network*, compressing the information form the inputs.

*Hebbian learning* is the most common variety of unsupervised learning [Hertz *et al.*, 1991]. Hebbian learning minimizes the same error function as an auto-associative network with a linear hidden layer, trained by least squares, and is therefore a form of dimensionality reduction.

Another form of unsupervised learning is *cluster analysis* which is the assignment of a

set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense.

Among neural network models, the *self-organizing map* (SOM) [Kohonen, 2001] and *adaptive resonance theory* (ART) [Carpenter and Grossberg, 1988]are commonly used unsupervised learning algorithms. The SOM is a topographic organization in which nearby locations in the map represent inputs with similar properties. The ART model allows the number of clusters to vary with problem size and lets the user control the degree of similarity between members of the same clusters by means of a user-defined constant called the vigilance parameter.

## 1.5 Initial form feature extraction and models for object recognition

### 1.5.1 Form feature extraction

In this section we introduce same basic form feature extraction methods commonly used for (pre-)processing in image analysis.

#### Structure Tensor

The *structure tensor* is an efficient mathematical tool for the extraction of local patterns when compared with the directional derivative through its coherence measure. It is typically used to represent and detect gradients, edges, or corners in an image [Harris and Stephens, 1988]. The structure tensor matrix is formed as

$$S = \sum_u \sum_v w(u,v) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \tag{1.1}$$

where $I_x \equiv I(u,v)_x$ and $I_y \equiv I(u,v)_y$ denote the derivatives of the image intensity function $I$ in $x$ and $y$ directions at the spatial positions $u$,$v$ weighted by a (Gaussian) kernel function $w$ that integrates over a spatial neighborhood. The derivatives of a discrete intensity function can be approximated with a Sobel operator [Jähne *et al.*, 1999]. Eigendecomposition is then applied to the structure tensor matrix $S$ to get the eigenvalues $(\lambda_1, \lambda_2)$ and eigenvectors $(\vec{e}_1, \vec{e}_2)$. These values have the following properties:

1. $\vec{e}_1$gives an approximation of the direction of the local gray scale gradient. The local direction of image structure is given by $\vec{e}_2$. Furthermore, $\vec{e}_1$and $\vec{e}_2$ are orthogonal to each other.

2. The eigenvalues $\lambda_1$ and $\lambda_2$represent a confidence measure for the approximation of $\vec{e}_1$and $\vec{e}_2$.

3. Undirected structure leads to a low difference of the eigenvalues while directed structure lead to a high difference of the eigenvalues

4. A measure for the cornerness of a local 2D structure is given by [Förstner, 1986]

$$C = \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \rightarrow max \tag{1.2}$$

In chapter three, we compare a biologically motivated model approach for the extraction of junctions and corners with the structure tensor as a machine vision method.

#### Gabor Filter

The Gabor filter is a spatial frequency and orientation selective linear band pass filter which is used to acquire information about periodic properties of image patterns [Daugman, 1988]. A 2D Gabor filter consists of a Gaussian kernel function modulated by a sinusoidal plane wave. Gabor filers are self-similar which means that all filters can be generated from one mother wavelet by scaling and rotation.

$$g(x, y) = exp\left(\frac{x'^2 + \gamma y'^2}{2\sigma^2}\right) cos(2\pi \frac{x'}{\lambda} + \psi) \tag{1.3}$$

given

$$x' = x \cos\theta + y \sin\theta$$

$$y' = -x \sin\theta + y \cos\theta$$

Here, $\lambda$ represents the wavelength of the cosine factor, $\theta$ represents the orientation of the normal to the parallel stripes of a Gabor function, $\psi$ is the phase offset, $\sigma$ is the spatial width of the Gaussian envelope and $\gamma$ is the spatial aspect ratio, and specifies the anisotropy of the elliptical Gabor function.

Several Gabor filters can be combined in a filter bank to represent all orientations of a specific scale. In this thesis, Gabor filters are used to model the properties of simple cells arranged in hypercolumns in primary visual cortex [Hubel and Wiesel, 1968].

### 1.5.2 Models for object recognition

Here, we present a brief overview of different approaches for object recognition.

#### Scale-Invariant Feature Transform (SIFT)

[Lowe, 2004] presents a method for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene. An image is transformed into a large collection of local feature vectors, each of which is invariant to image translation, scale, and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. The scale-invariant features are efficiently identified by using a staged filtering approach. The first stage identifies key locations in

scale space by looking for locations that are maxima or minima of a difference-of-Gaussian function. Each point is used to generate a feature vector that describes the local image region sampled relative to its scale-space coordinate frame. The features achieve partial invariance to local variations, such as affine or 3D projections, by blurring image gradient locations. The resulting feature vectors are called SIFT keys. The SIFT keys derived from an image are used in a nearest-neighbor approach to indexing and to identify candidate object models. Collections of keys that agree on a potential model pose are first identified through a Hough transform hash table, and then through a least-squares fit to a final estimate of model parameters. When at least 3 keys agree on the model parameters with low residual, there is strong evidence for the presence of the object. It is possible to have substantial levels of occlusion since there may be dozens of SIFT keys in the image of a typical object.

### Geon-Model

According to Marr [Marr, 1982] human object recognition can be best understood by algorithms that hierarchically decompose objects into their parts and relations in order to access an object-centered 3D model. Based on the concept of non-accidental properties, Biederman proposed in his recognition by components (RBC) theory [Biederman, 1987], that the human visual system derives a line-drawing-like representation from the visual input, which is parsed into basic geometric primitives (geons) that are orientation invariant. Object recognition would be achieved by matching the geons and their spatial relations to a geon structural description in memory. This theory has been implemented in a connectionist network that is capable of reliably recognizing line drawings of objects made of two geons [Hummel and Biederman, 1992].

### HMAX-Model

The model proposed by [Riesenhuber and Poggio, 1999] combines and extends several recent models (e.g., [Poggio and Edelman, 1990; Perrett and Oram, 1993; Fukushima, 1980]) and effectively summarizes many experimental findings. The view-based module shown in the inset of Fig. 1.4 is a hierarchical extension of Hubel and Wiesel's classical paradigm of building complex cells from simple cells. The circuitry consists of a hierarchy of layers leading to greater specificity and greater invariance by using two different types of mechanisms (a MAX pooling mechanism (dashed lines), to increase invariance, and a template match operation (solid lines), to increase feature specificity, see text).

The output of the view-based module is represented by view-tuned model units Vn that exhibit tight tuning to rotation in depth (and illumination, and other object-dependent transformations such as facial expression, etc.) but are tolerant to scaling and translation of their preferred object view. Invariance to rotation in depth (or other object-specific transformations) is obtained by combining in a learning module several view-tuned units Vn tuned to different views (or differently transformed versions) of the same object [Poggio and Edelman, 1990], creating view-invariant (object-tuned) units On. These, as well as

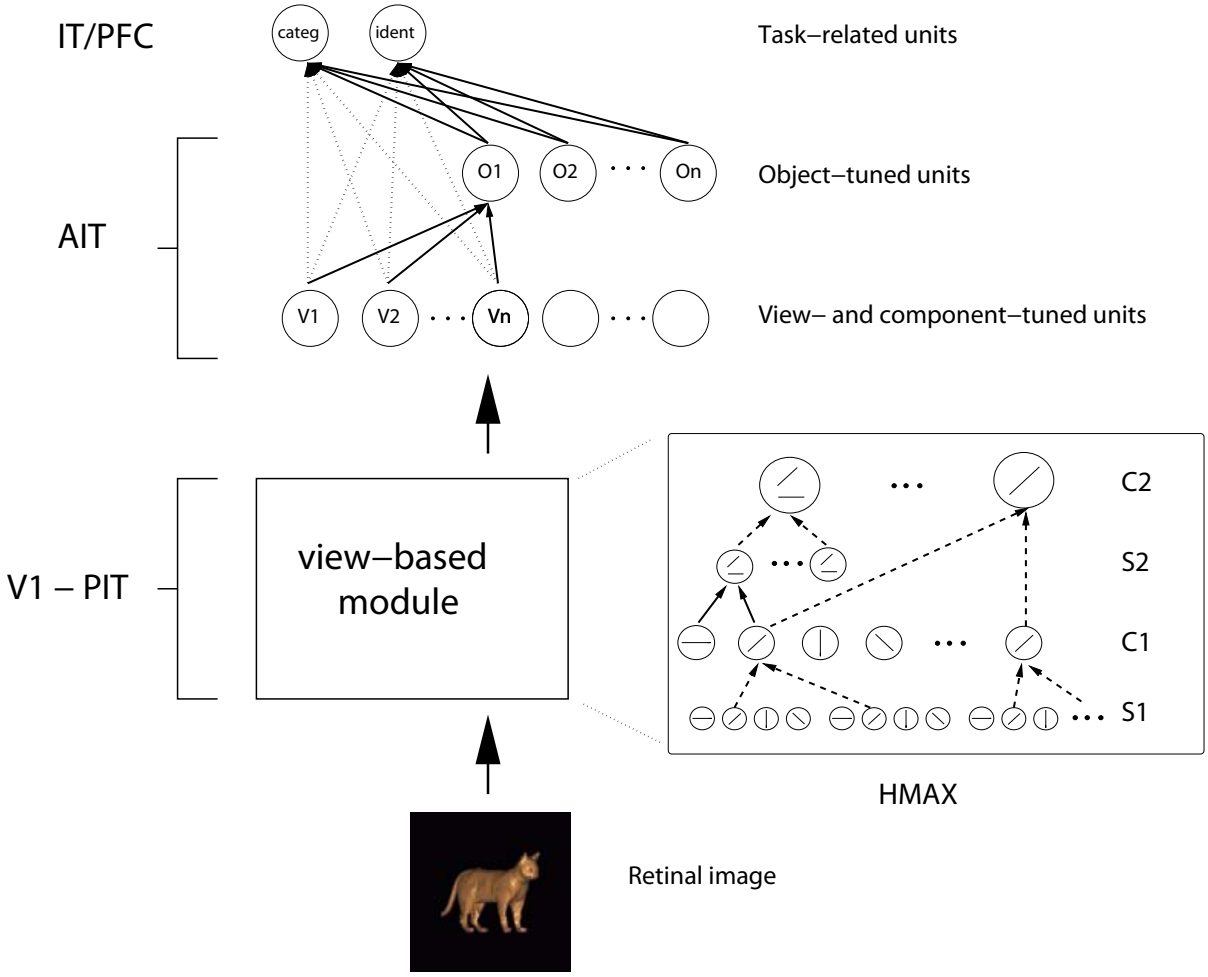


Figure 1.4: Model of the ventral pathway proposed by [Riesenhuber and Poggio, 2002]. The model consists of a view-based module where local features are processed in hierarchical layers of alternating SUM (solid lines) and MAX (dashed lines) operations. This corresponds roughly to simple (Sn) and complex (Cn) cells which is related to findings of Hubel and Wiesel. Outputs of the view-tuned module units (Vn) can be used to learn identification/discrimination tasks or object categorization. Invariance to rotation in depth (or other object-specific transformations) is obtained by combining several view-tuned units of the same object, creating view-invariant (object-tuned) units On. Object- and view-tuned units then serve as input to task modules that learn to perform different visual tasks such as identification/discrimination or object categorization.

the view-tuned units, can then serve as input to task modules that learn to perform different visual tasks such as identification/discrimination or object categorization. They consist of the same generic learning circuitry but are trained with appropriate sets of examples to perform specific tasks. The stages up to the object-centered units probably encompass V1 to anterior IT (AIT). The last stage of task dependent modules may be localized in AIT or prefrontal cortex (PFC) (see Fig. 1.4).

## 1.6 Outline of the thesis

Throughout this thesis we describe neurobiologically inspired models that simulate processing between cortical areas of the human visual system. Each chapter can be read independently from the rest of the thesis as they all contain the necessary information including an introduction, the particular model mechanisms, the results as wells as the corresponding discussion.

In *chapter 2* we introduce a biologically motivated recurrent model for the extraction of visual features relevant for the perception of 3D shape information from images of mirrored objects. We analyze qualitatively and quantitatively the results of computational model simulations and show that bidirectional recurrent information processing leads to better results then pure feedforward processing. Furthermore we utilize the model output to create a rough non-photorealistic sketch representation of a mirrored object, which emphasizes image features that are mandatory for 3D shape perception (e.g. occluding contour, regions of high curvature). Moreover, this sketch illustrates that the model generates a representation of object features independent of the surrounding scene reflected in the mirrored object.

In *chapter 3* we propose an extension of the neural model that suggests how surface-related cues for occlusion can be extracted from a 2D luminance image. The model employs feedforward and feedback mechanisms to combine contextually relevant features in order to generate consistent boundary groupings of surfaces. Moreover, contour junctions are localized and read out from the distributed representation of boundary groupings. Then, surface-related junctions are made explicit such that they are evaluated to interact as to generate surface-segmentations in static images. In addition, we compare our extracted junction signals with a standard computer vision approach for junction detection to demonstrate that our approach outperforms simple feedforward computation-based approaches.

In *chapter 4* we present a neural architecture that simulates hierarchical processing of human action patterns in the context of social interaction. The model simulates several cortical stages of the human form and motion pathways. Low-level feature processing is performed using mechanisms of recurrent interactions between early visual areas (V1-V2 in the form pathway and V1-MT in the motion pathway). Synaptic plasticity between mid- and higher-level visual areas (MT-MST and V2-IT) is simulated by incorporating Hebbian learning mechanisms. Finally, a mechanism is proposed of how form and motion signals could interact in order to produce more robust and reliable recognition of motion

patterns than single-pathway based models.

In *chapter 5* we summarize the main findinds of this thesis and present a list with the most relevant publications that originate from this thesis.

---

# Chapter 2

# Depicting the 3D Shape of Objects and Surfaces

## 2.1 Introduction

Computer vision systems for recovering 3D shape from single static images typically impose stringent restrictions on the lighting conditions or reflectance properties of the object under scrutiny. For example, it is common for shape-from-shading algorithms to require orthographic projection, a single, infinitely distant point light source, or Lambertian reflectance (e.g. Bruckstein [1988]; Horn and Brooks [1985]; Samaras and Metaxas [1999]; Zheng and Chellapa [1991]) ; for a review see [Horn and Brooks, 1989; Zhang *et al.*, 1999]. By contrast, the human visual system is extremely flexible. Although the appearance of a surface can change dramatically depending on its material composition, we rarely experience any difficulty in recovering a detailed and accurate estimate of an object's shape, irrespective of its reflectance properties.

One of the most striking examples of this is our ability to recover the shape of a perfectly specular (i.e. mirrored) surface, such as a chrome bumper or polished kettle. Perfectly specular surfaces are particularly problematic for the visual system because the images that they project onto the retina consist of nothing more than a distorted reflection of the surrounding scene. Consequently, as a mirrored surface is moved from scene to scene, the image changes dramatically. Indeed, depending on the context in which it is placed, a mirrored object can be made to take on any arbitrary appearance. For example, by carefully modifying the reflected scene, it is possible to make a surface appear to contain bumps or dents. The visual system would have no way of knowing that it was the environment and not the object's geometry that was responsible, and thus the problem of recovering the 3D shape of a mirrored surface is fundamentally ill-posed [Hadamard, 1902].

Given the inherent ambiguity of the problem, it is not possible to completely recover 3D shape without imposing additional assumptions or constraints. One solution is to

assume that the positions of features in the surrounding environment are known in advance [Savarese and Perona, 2001, 2002], so that their reflection in the surface can be identified and interpreted. However, as a model of human shape perception this is not very satisfying, as it seems quite unlikely that the visual system constructs a complete representation of the environment surrounding the object prior to recovering its shape.

Here, we take an alternative approach. Rather than attempting to fully reconstruct 3D shape, we develop a biologically motivated image processing model that is designed to extract a restricted but highly informative class of shape measurements from the image. Importantly, the model requires only weak assumptions about the statistical properties of the reflected scene, and thus operates across a wide range of real and artificial illumination conditions.

### 2.1.1 Goals

We apply the image-processing architecture to achieve two distinct goals. The first goal is to provide a model of the front-end of a 3D shape estimation system, inspired by the physiology of the early visual system. We aim to provide a plausible model of how the human visual system could use simple image measurements to achieve shape constancy across variations in the illumination. Additionally, by applying further constraints to the output of the model, the image processing architecture we present here could also form the basis of a computer vision system for fully recovering 3D shape under complex, unknown illumination.

The second goal is a concrete application of the model to computer graphics and visualization, specifically, facilitating the visualization of 3D surface geometry. The image of a mirrored surface under natural illumination is riddled with complex, high contrast patterns, which can be distracting if the aim of the user is to quickly visualize the most important properties of a shape (see Fig. 2.1). The image processing system presented here produces as output a modified 'sketch-like' representation of the input image, in which salient shape features that are invariant across illuminations are emphasized, while distracting, illumination-specific image features are suppressed. This non-photorealistic sketch could be used to enhance shape apprehension in industrial or graphic design, somewhat like a technical illustration. On the other hand, it could also be useful for aesthetic applications, to create, bold, charcoal-like renditions of objects. Finally, the model could also be used to guide the design of novel shape visualization systems, as it provides a principled explanation of which shape properties should be emphasized to confer an illumination-invariant impression of shape.

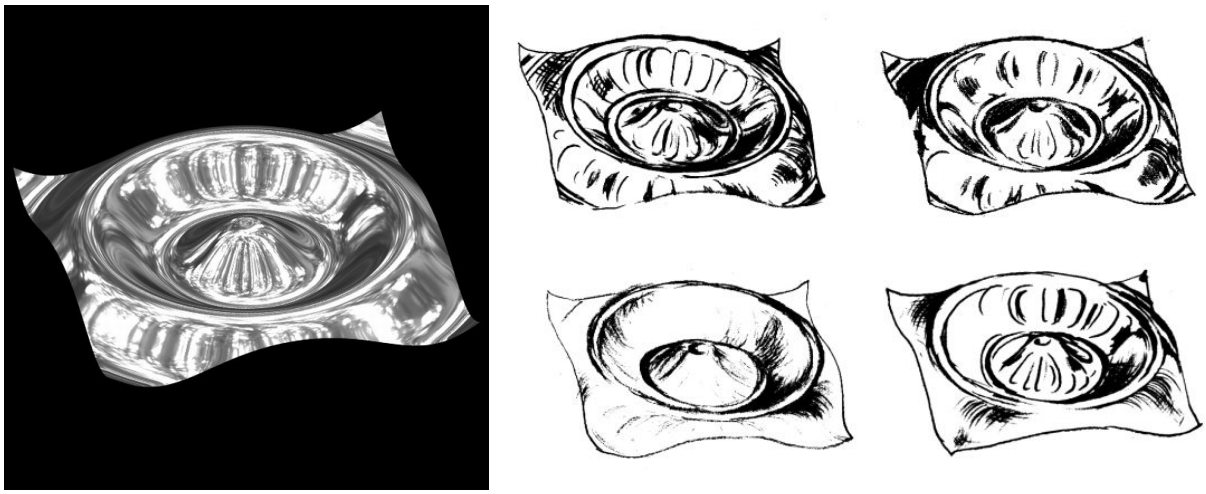

Figure 2.1: (a) Perfectly specular object illuminated by a forest scene (left) and different versions of sketch drawings created by a skilled artist (right). Note that the artist seems to emphasize those features that seem to be important for the perception of 3d shape.

### 2.1.2 Previous work

#### Human Perception

It is well established that specular reflections facilitate human shape perception. Psychophysical studies have shown that specular reflections contribute to shape estimation in the presence of other cues, such as shading, binocular stereopsis and texture [Blake and Bülthoff, 1990, 1991; Norman *et al.*, 2004; Todd and Mingolla, 1983; Todd *et al.*, 1997].

[Savarese *et al.*, 2004] showed subjects patches that were cropped out of photographs of mirrored surfaces reflecting a standard checkerboard pattern. The subjects' task was to identify which of three categorically different shapes the patch belonged to (sphere, cylinder or hyperbolic paraboloid). They found that subjects performed barely above chance levels. By contrast, using the popular 'gauge figure' task, [Fleming *et al.*, 2004] found that humans are good at estimating 3D shapes from mirrored surfaces, even when the objects are shown in isolation (i.e. so that there is no information about the surrounding scene). The inconsistency between these findings likely results from differences in the stimuli. Fleming et al. used complex, irregular object shapes; the entire object was visible simultaneously; and the patterns reflected in the surface were richly structured real-world scenes. Under these conditions, humans seem to be excellent at inferring 3D shape from specular reflections.

#### Computational work

Compared to the immense body of work on shape from shading, specular reflections have received relatively little attention. A number of authors have used active light techniques to overcome the inherent ambiguity of specular reflections. For example, [Ikeuchi, 1981]

analyzed photometric stereo for the special case of specular surfaces. [Sanderson *et al.*, 1988] developed SHINY, a structured light system to recover surface depth and orientation for industrial applications, using both single and multiple cameras. More recently, [Zheng and Murata, 2000] developed a system in which a rotating specular object was illuminated by extended circular light sources. The shape for each rotation plane was computed from motion stereo, or by tracing the specularities' motion across the surface.

Other authors have used multiple views or camera motion. [Koenderink and van Doorn, 1980] described the qualitative behavior of specular highlights as they move across curved surfaces in response to viewer motion. Blake and colleagues [Blake, 1985; Blake and Brelstaff, 1988; Blake and Bülthoff, 1990, 1991] analyzed the problem of specular stereo, showing how the position in depth of a specular highlight is related to the curvature of the surface. [Zisserman *et al.*, 1989] provided a quantitative analysis of the information available to a camera undergoing known motion. One key result was that the convex/concave ambiguity can be resolved under unknown illumination. [Oren and Nayar, 1996] developed an algorithm for discriminating between real and virtual features based on their motion across the surface. The authors use their analysis to uniquely recover 3D surface profiles by tracking a single virtual feature across the surface.

In elegant computational work [Savarese and Perona, 2001, 2002] were the first to provide a general solution for recovering shape from mirror reflections in single static images. However, the solution requires that a calibrated scene be reflected in the surface. Where three intersecting lines are visible in the reflected pattern, first order local information can be recovered.

Several authors have observed that the shapes of specular highlights are influenced by surface geometry [Beck and Prazdny, 1981; Hartung and Kersten, 2003; Longuet-Higgins, 1960; Todd *et al.*, 2004], noting that this could be important in discriminating highlights from other effects such as texture markings. Extending this observation, [Fleming *et al.*, 2004] showed how populations of simple filters tuned to different orientations could be used to extract information related to 3D surface curvatures directly from single static images of mirrored surfaces under unknown illumination. Surface curvature distorts the reflected environment into complex patterns of image orientation that vary continuously across the surface. Fleming et al. showed how these *'orientation fields'* are systematically related to the underlying geometry.

The current work builds on this observation to create a complete neurally-inspired architecture for extracting clean, reliable orientation fields from noisy images. The most important contribution is the addition of an iterative grouping circuit, that refines the local orientation estimates depending on the neighborhood. This substantially improves the accuracy with which shape properties can be estimated. Furthermore, it is this feedback circuit that enables the model to produce the non-photorealistic sketch representation for emphasizing the illumination-invariant features of the image.

### Non-photorealistic rendering and shape visualization

The technical illustrator's art is to depict the essential structural and functional components of a device without overpopulating the image with confusing or distracting details. It is widely believed that simplified illustrations of objects actually improve perception (and/or comprehension), although the empirical evidence for this (e.g., [Biederman, 1987; Dwyer, 1967; Fraisse and Elkin, 1963; Ryan and Schwartz, 1956]) is rather mixed, and likely depends on the task to be performed (e.g. recognition of familiar objects *vs.* assembling a complex object from instructive illustrations). However, there are certainly cases in which exaggerated or caricatured stimuli are preferred to their realistic counterparts in a biological context [Tinbergen and Perdeck, 1950; Tinbergen, 1951] or yield superior task performance in humans [Benson and Perrett, 1991; Rhodes *et al.*, 1987], suggesting that non-photorealism might be exploited to facilitate perception. Here we attempt to create sketch-like representations of the input image (see Fig. 2.1) to aid shape apprehension, and for aesthetic applications.

There is a large body of previous research on image-based non-photorealistic rendering (NPR), in which arbitrary images or videos are fully or semi-automatically modified to create the impression of a particular medium or artistic style, including paint [Curtis *et al.*, 1997; Hays and Essa, 2004; Hertzmann, 1998; Shiraishi and Yamaguchi, 2000], pencil [Jin *et al.*, 2002; Yamamoto *et al.*, 2004], stipple drawings [Deussen and Strothotte, 2000], mosaics [Hausner, 2001], cubism [Collomosse and Hall, 2003], impressionism [Litwinowicz, 1997], or simply stylized [DeCarlo and Santella, 2002].

A number of researchers have involved some degree of user interaction, to improve the quality of results. For example, [Durand *et al.*, 2001] developed a system for creating artistic renditions of photographs, in which the user determines stroke density and important structural features to interactively create the drawing. Recently, [Kang *et al.*, 2005] developed an interactive technique for generating cartoon-like sketches from photographs. The system uses *wavelet frames* to allow multi-resolution control of B-splines, which represent the depictive strokes. Other NPR sketch systems are designed more to give an overall 'gist' of the depicted person or object [Chen *et al.*, 2004; Gooch *et al.*, 2004] rather than faithfully showing 3D shape properties of arbitrary objects.

There has also been considerable amount of work on the optimal rendering parameters for visualizing 3D shape from geometric models. [Gooch *et al.*, 1998] developed a shading technique for automatically generating technical illustrations from 3D models. The resulting sketches combine edges for depicting boundaries and a highly stylized 'cool-to-warm' shading to produce the impression of curvature in 3D, although no psychophysical motivation or validation was offered. Interrante and colleagues [Interrante and Kim, 2001; Interrante *et al.*, 2002; Kim *et al.*, 2003] have systematically explored the influences of texture on 3D shape visualization, and developed methods for depicting transparent surfaces so that two superimposed surface shapes can be visualized simultaneously. Recently, [Bair *et al.*, 2005] used a combination of psychophysics and machine learning techniques to find perceptually optimal textures for visualizing two superimposed surfaces.

By contrast to most previous NPR research, our sketch algorithm is intended to take

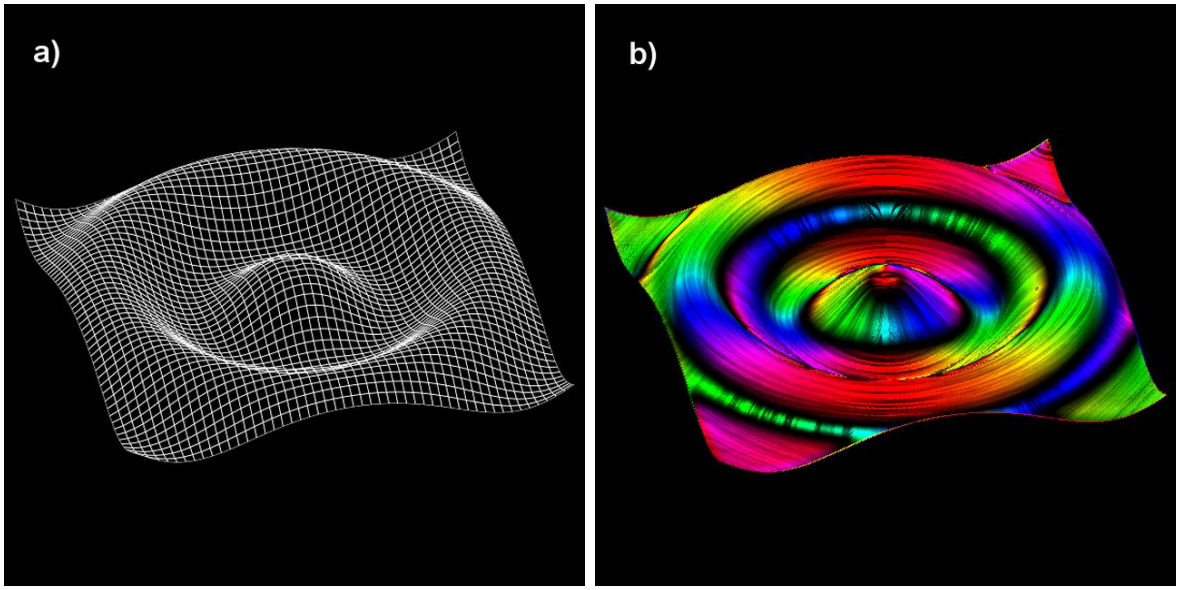


Figure 2.2: 3D model and corresponding curvature information. Minimal curvature orientations are color coded. The anisotropy of curvature is displayed as the intensity of the color. We use the LIC method described in [Cabral and Leedom, 1993; Stalling and Hege, 1995] in addition to the color coding to illustrate local orientations. Note that there are distinct lines of isotropic curvature (dark) which belong to inflection points of the surface. Here one curvature component changes the sign, resulting in an abrupt change of the orientation by 90 degrees.

a single grayscale image as input and to automatically produce a modified version of the image as output. In the resulting 'sketch', image regions containing reliable illumination-invariant shape features are emphasized, while regions containing spurious orientations that are due to reflections of the environment are suppressed.

## 2.2 Methods

In this section, we explain how a particular class of curvature-related information can be extracted directly from a 3D model of an object. This information (the '*ground-truth*') will be used later to assess the accuracy with which our model estimates these values from a rendered image of the object. Then we continue to give an overview of our proposed model for the extraction of curvature information followed by a detailed description of the model and its different components.

### 2.2.1 Extracting ground-truth curvature information from the 3d model

The intrinsic properties of surface geometry can be described by means of differential geometry. For example, a regular surface in $R^3$ is locally defined by its orientation and curvature properties depicted by mutually orthogonal tangent vectors and mutually orthogonal normal sections that define the curves of minimal and maximal normal curva-

tures. The product of the normal curvatures defines the Gaussian curvature of the surface at a selected point on the surface. In perception, we are concerned with the extraction of surface properties from *images* of illuminated surfaces taken from a certain viewpoint. Due to the projection of the visible surface part of an object its shape can be described by its height, that is, as a function $z = f(x, y)$ (known as Monge patch, [do Carmo, 1976])[1]. In this coordinate system (see Fig. 2.3), the first derivatives of $f$ (i.e. the gradient of $f$) describe the slant of the surface which is the angle between the viewer's line of sight and the surface normal. The second derivatives of the surface (i.e. the Hessian matrix) describe the rate at which the surface normal changes with respect to the viewer. This can be described as the *view-centered* curvature of the surface.[2] Note that this has to be distinguished from the *intrinsic* curvature, which is defined in local coordinates. For example, the *intrinsic* curvature is constant in all directions at all locations on a sphere, while the *view-centered* curvature is equal in all directions only in the middle of the projected sphere; close to the boundary, the second derivatives are increasingly large in the direction perpendicular to the circumference, and zero parallel to the circumference.

Each point on the surface has a minimum and a maximum curvature direction which are always perpendicular to each other. In the case of view-centered curvature, they are always perpendicular to one another *in the image plane*. If the maximal and the minimal curvatures have the same magnitude then we speak of an *isotropic* surface. When minimal and maximal curvatures are different magnitudes, then the surface is *anisotropic*. Here, the ratio between maximal and minimal curvature magnitudes describes the strength of the anisotropy of curvature. In other words the anisotropy of curvature describes how spherical or cylindrical a surface patch is at any point. The anisotropy is very important because we'll see later that distortions in the mirrored scene are directly related to this parameter. To extract the discussed parameters, we need to compute the Hessian matrix (Eq. 2.1).:

$$H = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{pmatrix} \tag{2.1}$$

where $f$ is the surface function as mentioned above. The eigenvalues $(\lambda_1, \lambda_2)$ and the eigenvectors $(v_1, v_2)$ of $H$ have the following meaning:

1. the first eigenvector $v_1, v_2$ of the Hessian matrix describe the orientation of maximal curvature and minimal curvature, respectively

2. the ratio of the eigenvalues describes the anisotropy of curvature where the term

[1] We are neglecting influences of perspective projection in image acquisition by assuming that the object size is small in comparison to viewing distance such that the mapping can be approximated through an orthographic projection.

[2] For convenience, from here on we refer to the Hessian matrix and related concepts loosely as 'curvature'. The reader should be careful to keep in mind that when we use the term 'curvature' we mean the *second derivatives*, and when we speak of 'principal curvature directions' and 'principal curvatures', we mean the *eigenvectors* and corresponding *eigenvalues* of the Hessian matrix. For Monge patches the Gaussian curvature can be calculated from the Hessian by taking its determinant scaled by a measure of the slant [do Carmo, 1976]).

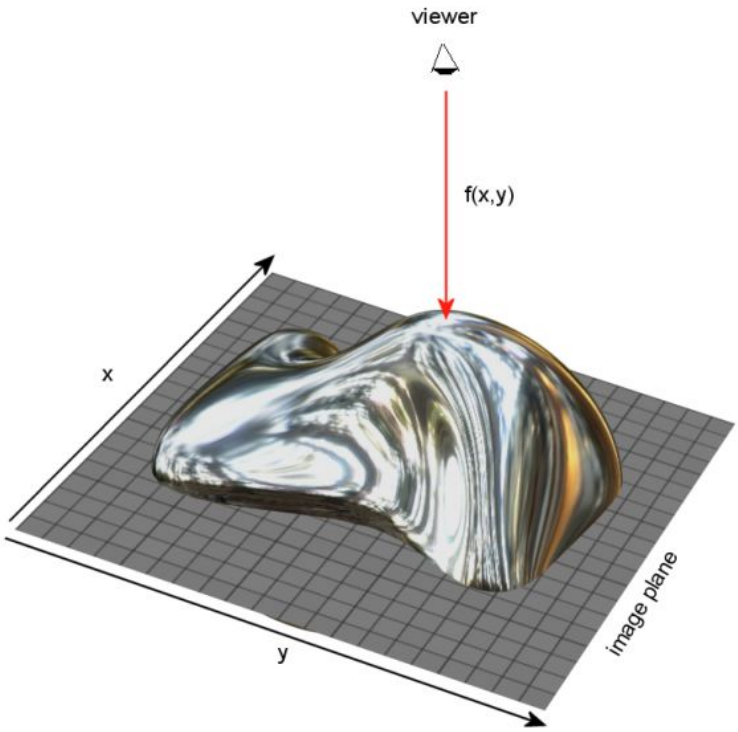


Figure 2.3: View-centered coordinate system. The visible surface part of an object is described by its height, i.e., as a function $z = f(x, y)$ where $z$ represents the depth and $(x, y)$ are the coordinates in the image plane.

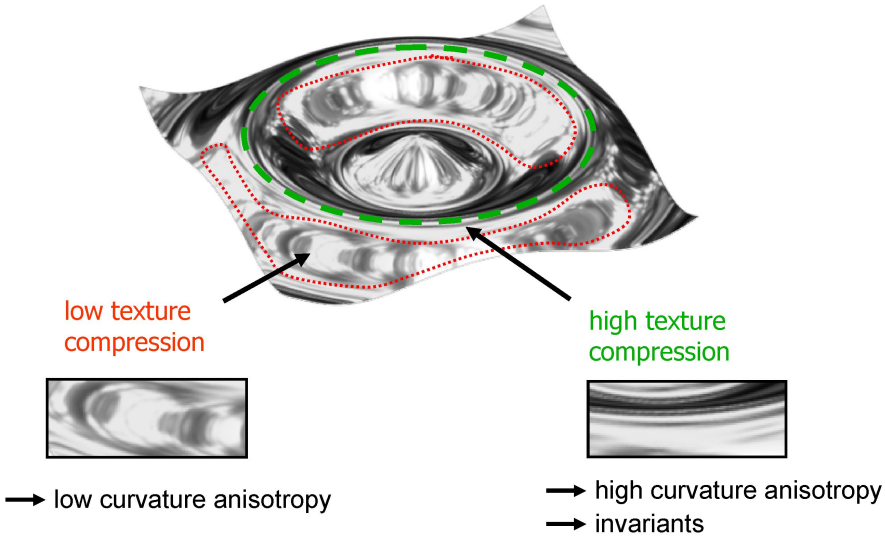


Figure 2.4: Image of a mirrored object. We can see that on areas with high curvature anisotropy such as the curvature ridge marked in green, the surrounding scene is compressed into thin long streaks aligned along the direction of the ridge (which is the direction of minimal curvature). Note that these areas turn out to be the locations where the local image statistics are invariant when changing the surrounding scene. On other areas where the surface curvature is nearly isotropic (marked red) the reflected scene shows much weaker compressions resulting in a flat mirror like reflection. Here, the local image statistics are very dependent of the surrounding scene.

$1 - \frac{\lambda_2}{\lambda_1}$ yields values between zero (isotropic) and one (anisotropic).

The directions of minimal curvature and the anisotropy computed from the 3D model represent the ground-truth information which is illustrated in Fig. 2.2. To understand the ground truth images there are a few things worth mentioning: first, we can see that there are distinct lines of low curvature anisotropy which belong to inflection points of curvature. Here, the surface curvature is equal in all directions, leading to no distortions of the reflected world on the surface. Note that as we cross these lines, the orientation of minimal surface curvature changes abruptly by 90 degrees. The reason for this effect is that the surface changes from a concave or convex condition to a saddle condition where one principal curvature component changes sign. Second, we can see singular points of low curvature anisotropy which belong to locally spherical patches (concave or convex), facing the viewer. These points are usually surrounded by a radial field of minimal curvature orientations. For more information about the surface geometry see [do Carmo, 1976].

### 2.2.2 Evidence for curvature orientation in image space

When a scene is reflected in a curved mirror, the reflection is distorted in a way that depends systematically on the 3D surface geometry [Beck and Prazdny, 1981; Fleming *et al.*, 2003, 2004; Hartung and Kersten, 2003; Longuet-Higgins, 1960]. Intuitively, highly curved surfaces 'see' a large angle of the surrounding scenery, and thus compress many features into a small proportion of the image. By contrast, for slightly curved surfaces the compression is weaker (see Fig. 2.4). When the surface has different curvatures in different directions, the reflections will be differentially compressed in the two directions, leading to a locally affine distortion of the reflection. The strength of the distortion depends on the ratio between minimal and maximal curvatures, which we call the surface anisotropy. For example, a spherical surface patch is curved equally in all directions (i.e. isotropic), and thus the surrounding scene is simply miniaturized in the reflection, it is not subjected to any anisotropic distortion. By contrast, a cylindrical surface patch is somewhat curved in one direction, but completely flat (i.e. curvature equals zero) in perpendicular direction. This leads to a strong distortion of the surrounding scene caused by a high anisotropy of curvature. In this case, the reflections tend to be distorted into parallel streaks that are aligned with the direction of minimum curvature[3] (see also Fig. 2.2). We can summarize that:

1. the orientation of structures in the mirrored scene tends to be aligned with the minimal surface curvature

2. the strength of the distortions in the mirrored scene indicates the ratio of minimum to maximum curvature (anisotropy of curvature)

### 2.2.3 A biological model for the extraction of curvature information

Our model receives as input a grayscale image of a specular object. The model is able to estimate the minimal curvature orientations and the anisotropy of curvature from the 2D input image. To measure the success of this estimation, we can compare the results to the ground truth information derived from the 3D model of the object. The model also produces a sketch-like representation of the input image. Our proposed model consists of three main components: (i) extraction of oriented contrasts, (ii) orientation selective grouping and (iii) a recurrent feedback cycle. An overview of the model architecture is given in Fig. 2.5.

[3]Note that the orientation of maximal curvature is always perpendicular to the orientation of minimal curvature as long as we consider the *view-centered* curvature depending on the second derivatives of the surface (Hessian matrix). As a consequence of this, it is sufficient to recover the direction of minimal curvature.

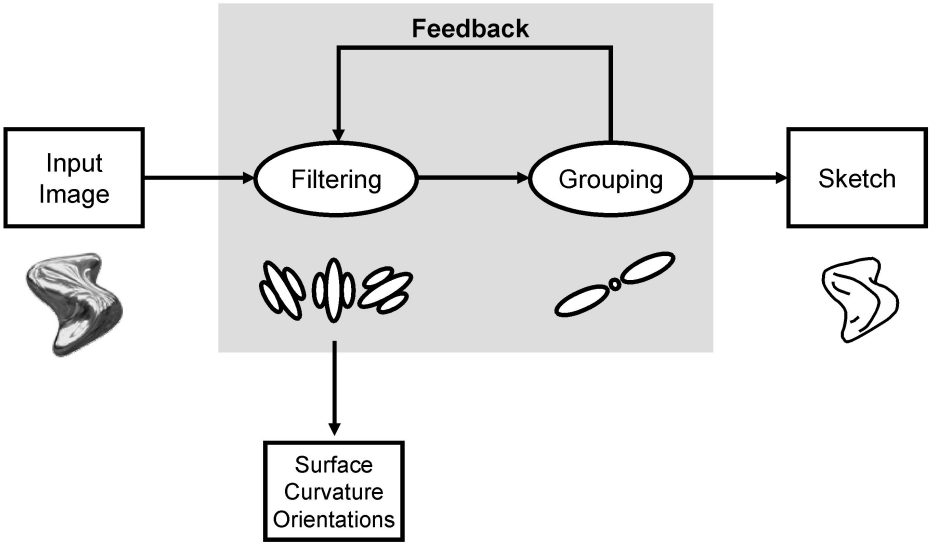


Figure 2.5: Overview of the raw model architecture. The input image is filtered by differently oriented Gabor filters resulting in different feature maps. Following, each feature map is processed by a grouping stage, that groupes together like-oriented contrasts. The resulting feature maps are then use as a feedback signal for the inital filter stage. The feedback cycle is iterated several times until the signals have reached a stable state. From this representation, information about surface curvature can be read out. Also, a rough sketch of the object can be extracted that emphasizes perceptually relevant surface regions.

### Basic assumptions of the model

Our model assumes that any anisotropy measured in the image is due solely to distortions of the reflected scene caused by the geometry of the mirrored surface. This tacitly assumes that the texture of the environmental scene is isotropic (i.e., contains a uniform distribution of orientations) [Fleming *et al.*, 2004]. For many artificial and natural scenes this is approximately true globally, although it is clearly infringed locally when the scene contains extended oriented structures such as trees or buildings. Despite this, the distortions introduced by surface anisotropy can be very powerful, and—unlike naturally-occurring oriented structures—affect all spatial scales equally. Indeed, even environments with unnaturally anisotropic scene statistics can nevertheless yield orientation fields that are predominantly biased in the correct directions, although this depends on the shape properties of the reflecting object. The effects of this '*isotropy assumption*' are discussed in greater detail in section Section 2.4.

### Extracting oriented contrasts using a population of linear Gabor filters

The initial stage of our model applies a family of orientation selective Gabor filters [Daugman, 1988] resembling the response properties of cortical simple cells [Hubel and Wiesel, 1968]. The result is interpreted as a population code describing local contrast information. In our simulations we employ 18 filters rotated from 0 to 170 degrees to extract local contrast information (Eq. 2.2).

$$R_\phi = I * G_\phi \tag{2.2}$$

where $I$ is the input image of the mirrored object, $G_\phi$ is the oriented Gabor filter rotated by angle $\phi$ (see Fig. 2.6), and $*$ is the convolution operator. In order to get a proper scaling of the computed responses we apply a normalization to the Gabor filter output (Eq. 2.3). Parameter $\mu$ affects the strength of the normalization curve. To compensate for global contrast effects $\mu$ is multiplied with the average contrast over all responses for a given orientation where $x$ and $y$ are the dimensions of the image (see Eq. 2.4).

$$S_\phi = \frac{R_\phi}{\mu k_\phi + R_\phi} \tag{2.3}$$

where

$$k_\phi = \frac{\sum R_\phi}{x \cdot y} \tag{2.4}$$

In Fig. 2.6, the population responses for two different locations in the input image are shown. The first population response is extracted from a location in the image where the texture is strongly distorted in one direction leading to strong responses in one preferred orientation. The other population response is extracted from a location in the image where the texture is only weakly distorted. Here the population responses are nearly equally distributed over all orientations. Fig. 2.6 illustrates how the direction is extracted from the orientation of the Gabor filter producing the maximum response at one specific

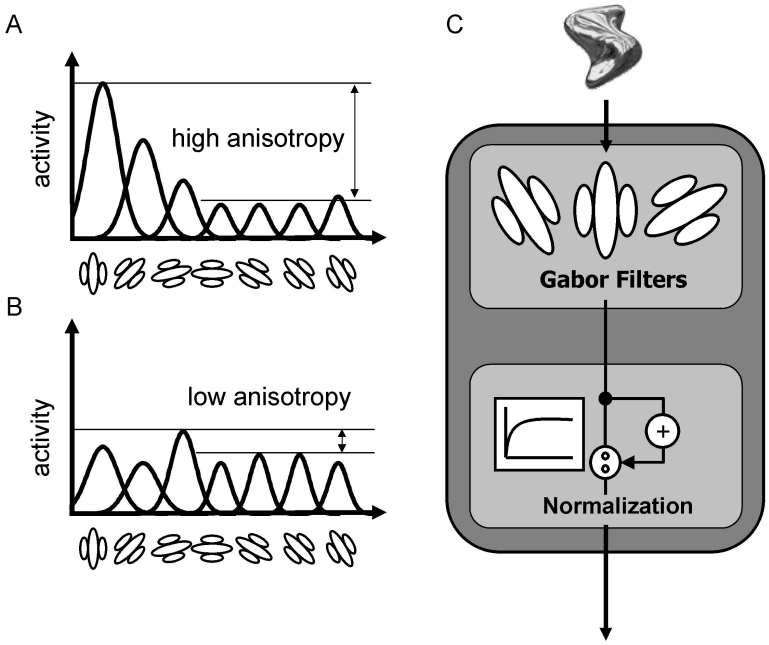


Figure 2.6: Initial filter stage of the model (C). A family of orientation selective Gabor filters is applied to the input image followed by a normalization step in order to get a proper scaling of the responses. The distribution of the filter responses is shown for an anisotropic texture condition (A) and a nearly isotropic texture condition (B). Note that the filter which yields maximal response determines the prevailing orientation of texture distortion. Note also that the more the distribution differs from an equal distribution the stronger the anisotropy of the texture is.

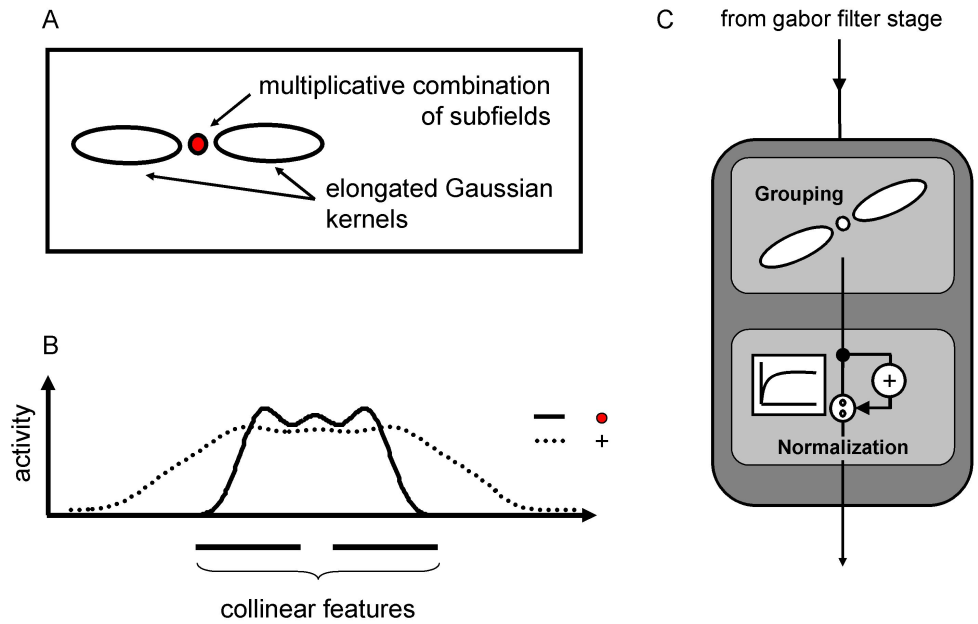


Figure 2.7: Grouping stage of the model consisting of the actual grouping step and an additional normalization that is necessary to achieve bounded output and to get a proper scaling of the filter responses (C). The grouping filter consists of two elongated Gaussian kernels which are combined by a multiplicative connection (A). (B) shows a comparison between filter responses in case of an additive (dotted) and a multiplicative (solid) connection when the input is a short line segment. Note that an additive connection would lead to smearing effect at line ends.

location (Eq. 2.5).[4]

$$\phi_{max} = argmax(S_\phi) \tag{2.5}$$

We employ the ratio between minimal and maximal filter response to compute the anisotropy of the distortion, specifically we use the term in Eq. 2.6.

$$A = 1 - \frac{min(S_\phi)}{max(S_\phi)} \tag{2.6}$$

which yields values between zero (isotropic) and one (anisotropic). Note that this data interpretation has been adopted from [Fleming *et al.*, 2004].

[4]Note that in this case the texture distortion is always unimodal in the direction of minimal surface curvature. Moreover, our filters are relatively small (just a few pixels per diameter) so that it is very unlikely that the distribution of orientation responses has more than one peak. Thus, we assume that the *argmax* operation always yields a unique result.

### Orientation selective grouping

The second component of our model consists of orientation selective grouping filters. The grouping filter is constructed of two displaced elongated Gaussian functions $G^{\leftarrow}_{\sigma_1,\sigma_2}, G^{\rightarrow}_{\sigma_1,\sigma_2}$ (elongation ratio $\sigma_1 : \sigma_2$) that are combined in a multiplicative manner (Fig. 2.7). These filters are applied to locally enhance coherent filter responses and to enforce the initial estimated anisotropy signal. Note that the grouping filters are rotated in the appropriate direction $\phi$ when they are applied for each orientation layer.

$$\begin{aligned} F^{\leftarrow}_{\phi} &= S_{\phi} * G^{\leftarrow}_{\sigma_1,\sigma_2,\phi} \\ F^{\rightarrow}_{\phi} &= S_{\phi} * G^{\rightarrow}_{\sigma_1,\sigma_2,\phi} \\ L_{\phi} &= F^{\leftarrow}_{\phi} \cdot F^{\rightarrow}_{\phi} \end{aligned} \tag{2.7}$$

In particular, this kind of filter has the effect that collinear Gabor responses (from the initial stage) are enforced, while potentially erroneous responses typically occur without context and thus are weakened by this operation. The multiplicative connection in Eq. 2.7 ensures that input from both sides of the center is needed to generate activation. Using an additive connection would lead to smearing effects at line ends (see Fig. 2.7). In other words, this stage produces strong responses if the underlying input signals a continuous orientation pattern. This is consistent with physiological findings about non-linearities in the response of V2 contrast cells.

Finally the grouping filter responses are passed through a second normalization step (Eq. 2.8) similar to the operation in Eq 2.3). This operation is necessary to keep the filter responses bounded while strong filter responses are enforced and weak filter responses are diminished.

$$M_{\phi} = \frac{L_{\phi}}{1 + L_{\phi}} \tag{2.8}$$

### Full recurrent model

So far we have described feedforward connections of the model. Now we use the locally grouped information (Eq. 2.8) to iteratively refine the initial estimates (Eq. 2.3). This operation can be realized by using the grouped filter responses as a recurrent feedback signal. Physiological evidence supports the view that top-down projections serve primarily as a modulation mechanism to control the responsiveness of cells in the primary visual cortex [Bullier *et al.*, 1988]. Accordingly, we use the grouped Gabor responses ($M_{\phi}$) as a feedback signal. Here, the feedback signal acts as a prediction of an error-free input pattern. The feedback modulation is realized by the following equation:

$$S^{new}_{\phi} = S^{init}_{\phi} \cdot (1 + \alpha M_{\phi}) \tag{2.9}$$

where $\alpha$ is a parameter to adjust the influence of the feedback signal $M_{\phi}$ and $S^{init}_{\phi}$

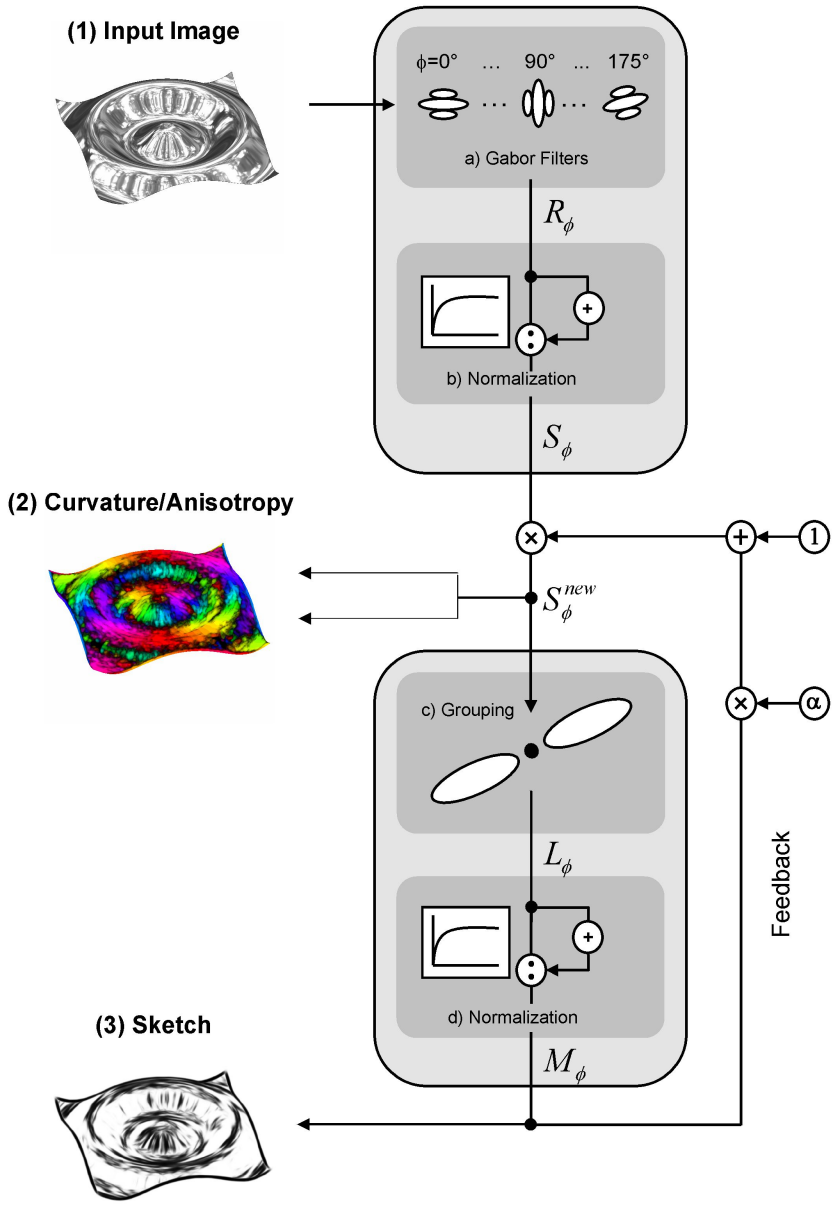


Figure 2.8: The full recurrent model in detail. The model receives a grayscale image of a specular object as input (1). First, a family of orientation selective Gabor filters is applied to the image (a) followed by a normalization step (b) to get a proper scaling. Second, a grouping filter is applied for each orientation to enhance collinear features (c), followed by another normalization step (d). Then the signal is fed back to iteratively refine the initial Gabor responses. Model outputs are the orientation of minimal curvature and the anisotropy of curvature (2). Furthermore, a sketch representation of the input image is extracted in (3).

is the initial signal from Eq. 2.3. Note that Eq. 2.3 has to be replaced by Eq. 2.9 for the full recurrent model (see also Fig. 2.8). The logic of the feedback modulation can be interpreted as follows:

The input signal is enhanced at locations where the initial signal matches the feedback signal. Thus, the feedback signal can be regarded as an expectation signal biasing local input activations. In cases where feedback is zero, the 'one' in Eq. 2.9 ensures that the initial signal remains unchanged. Iterative feedback processing thus strengthens collinear features over time and helps to reduce the influence of noise and the presence of errors.

*Output Signals.* There are three output signals of the model. The first and second output signal of the model yield the estimated minimal surface curvature orientation[5] (from Eq. 2.5) and the anisotropy signal (Eq. 2.6), respectively. Both output signals are illustrated in one single output image where the orientations are color coded and the anisotropy is coded as the intensity of the colors (see Fig. 2.8, Part 2). The third output signal (Fig. 2.8, Part 3) is an image of a non-photorealistic sketch of the input image. The sketch output is computed by the difference between maximal and minimal activity of the filter responses (Eq. 2.10). Note that the sketch signal is similar to the anisotropy signal (Eq. 2.6). However, it is extracted after the grouping stage of the model and with a slightly different computation rule.

$$A_{sketch} = max(S_\phi) - min(S_\phi) \tag{2.10}$$

## 2.3 Simulations

In this section, we show the competencies of our proposed model for three different specular objects depicted in Fig. 2.9. In particular, we show the estimated minimal curvature directions in combination with the anisotropy of curvature and a sketch representation of the object for a variety of model parameters.

### 2.3.1 Model input

As input, we use grayscale images of synthetically generated specular objects. All objects have a smoothly curved surface containing concavities, convexities as well as self-occlusions. To give the objects a specular surface we used reflection maps from [Debevec *et al.*, 2000]. In Fig. 2.9, the objects are depicted in a mesh style (to give a clear impression of the shape) and with a perfectly specular surface (used as model input). For computational simulations we utilize input images with a resolution of 600x600 pixels. Fig. 2.9 also shows a reflection map from [Debevec *et al.*, 2000], where the world around the mirrored objects is compressed into a 2D image[6].

[5]Recall that features on the mirrored surface are aligned with the direction of minimal surface curvature. Thus the filter direction with maximal response determines the *minimal* estimated curvature.

[6]Each pixel in the reflection map belongs to a specific direction in space. The spherical surrounding can be divided into degrees of latitude and longitude where each pixel in the reflection map belongs to

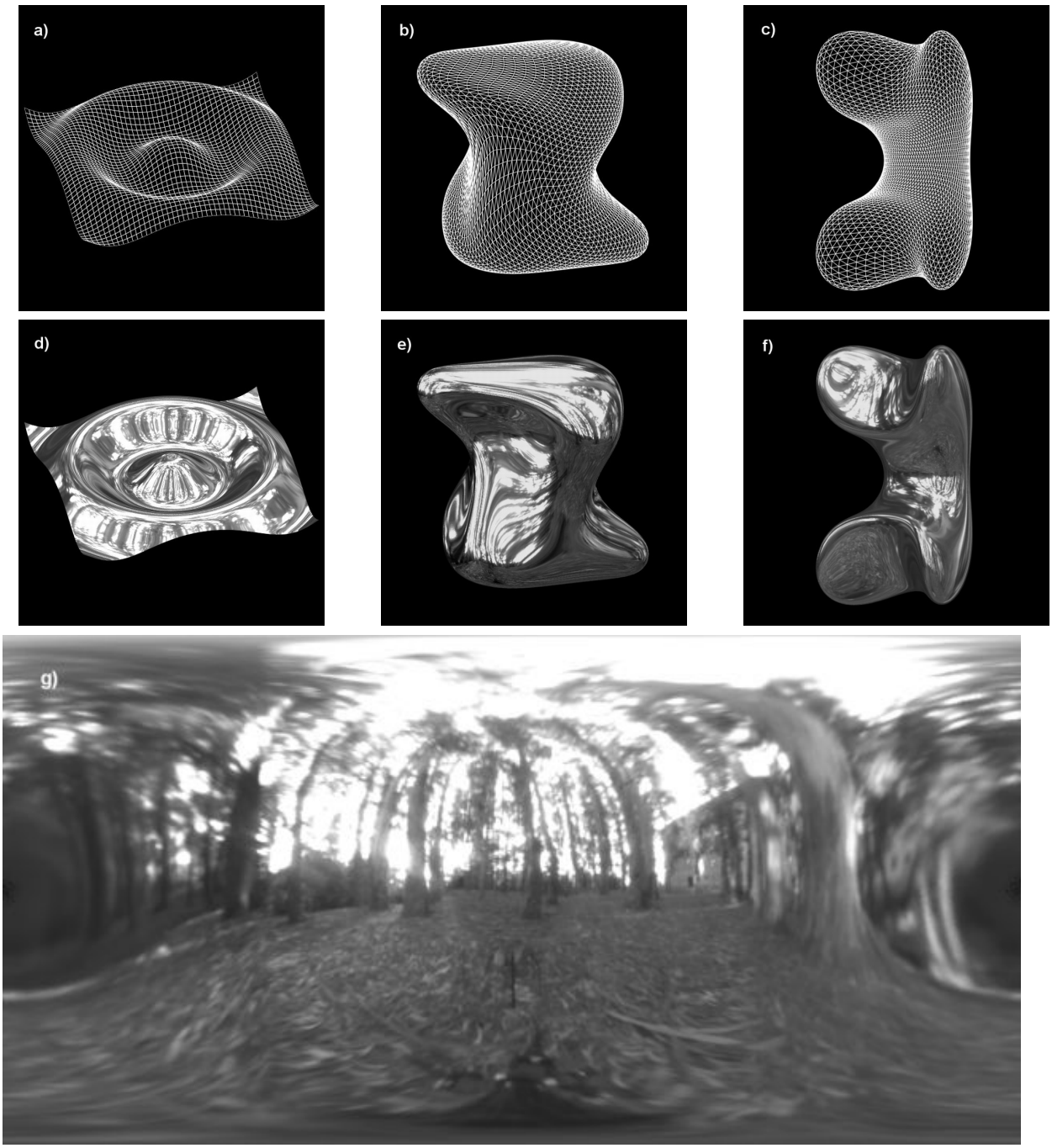


Figure 2.9: Three synthetically created specular input objects (d)-(f) and their corresponding 3D-Models (a)-(c). The objects are mirrored with the 'Eucalyptus Grove' reflection map from [Debevec *et al.*, 2000] (g). Object (a) was inspired by [Todd, 2004] and shows a plane perturbed by a circular wave function. Object (b) shows a sphere which was also squeezed and stretched to achieve a smoothly varying curved surface. Object (c) was created by bending and squeezing an ellipsoid which leads to a surface shape where nearly all possible curvature conditions are visible.

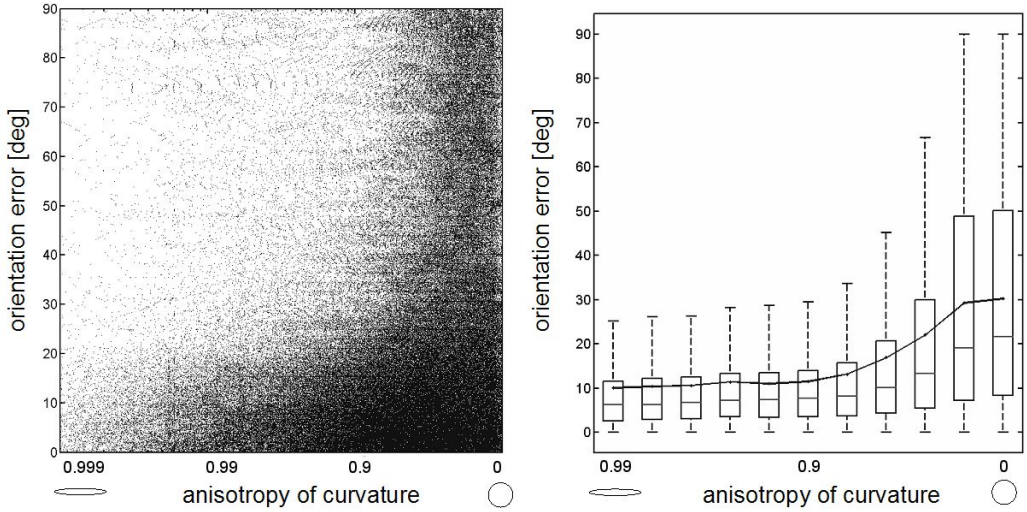


Figure 2.10: The figure illustrates orientation error as a function of surface anisotropy. Left plot shows individual observations, where each dot represents the measurement from a single pixel (several objects and environmental scenes were used). Right plot shows box plots computed from (a), where we divided the anisotropy axis into 11 bins (box represents lower quartile, median and upper quartile). For low curvature anisotropies the distribution of orientation errors is spread from 0-90 degrees while for high curvature anisotropies the distribution is only spread between 0 and 25 degrees. Mean orientation error drops from 30 degrees (isotropic condition) to about 10 degrees (anisotropic condition). Note that in both plots we use a logarithmic scale for the x-axis. Note also that as we employ only 18 different filter orientation in steps of 10 degrees the resulting mean orientation error cannot be lower than 10 degrees.

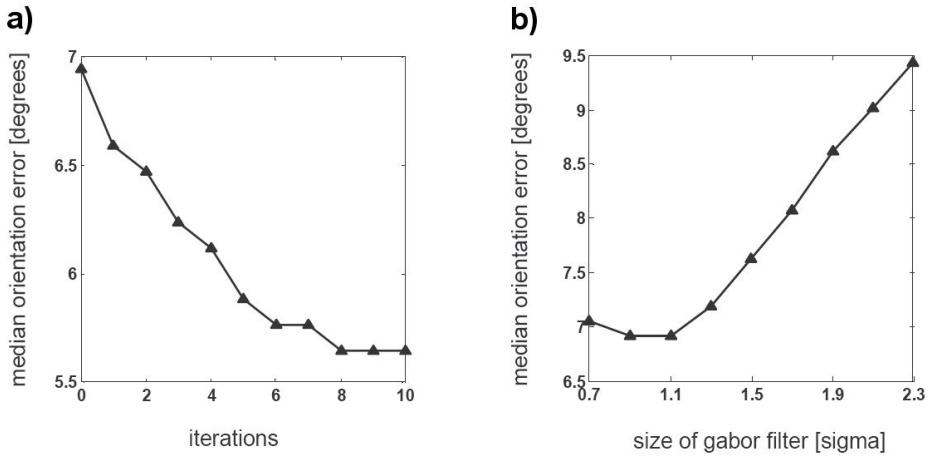


Figure 2.11: The graph in (a) shows the orientation error computed from one object/scene configuration over subsequent time steps of the feedback cycle. We can see that the error is decreasing with each time step and nearly converging after 10 time steps to a value of 5.6 degrees. The graph in (b) shows the median orientation error across different Gabor sizes of the initial filter stage. We can find a minimum error of 6.9 degrees for a Gabor filter size of $\sigma = 0.9$. Increasing the size of the Gabor filters leads to a monotonic increase of the orientation error. We can also observe that too small Gabor filters have a negative effect by producing slightly higher error measures.

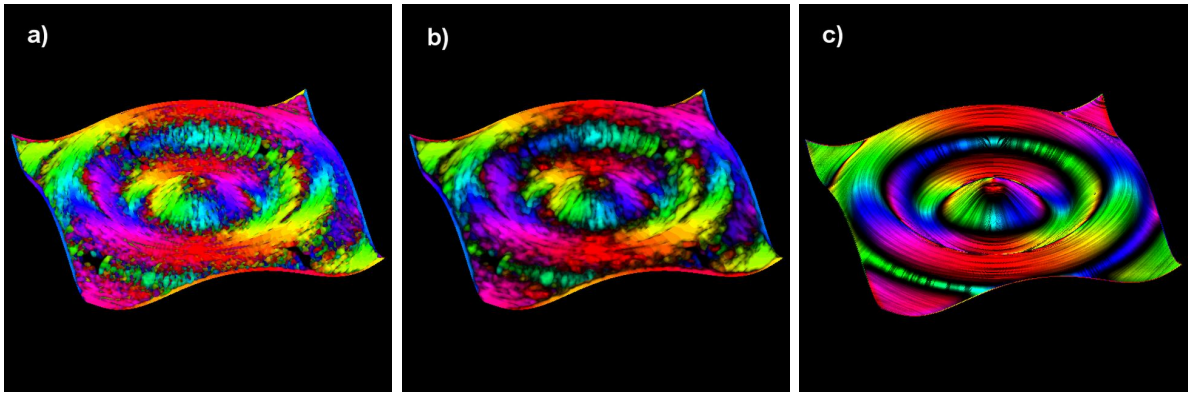


Figure 2.12: The figure shows measures of the minimal curvature orientations and the anisotropy of curvature. Curvature orientations are coded in color whereas the anisotropy is displayed as the intensity of the color (dark means isotropic curvature). The ground truth information (c) is compared to the initial detected curvature (a) (no grouping/feedback) and to the refined model output (b) after 10 cycles of grouping and feedback. We can see that especially in areas where the ground truth image shows isotropic curvatures the feedback helps to reduce orientation errors and to stress the anisotropy signal.

### 2.3.2 Evaluation of principal curvature orientations and anisotropy

In Fig. 2.10 we show the orientation error depending on the actual level of curvature anisotropy. The results demonstrate that for isotropic curvature conditions the orientation error is significantly higher than for the anisotropic curvature conditions. We further illustrate the initial estimates of curvature directions and surface anisotropy (without feedback and grouping) for different Gabor sizes in Fig. 2.13. It is clearly visible that small Gabor filters, tuned to high frequency components lead to better results than Gabor filters of larger scale. We also demonstrate qualitatively (Fig. 2.13) and quantitatively (Fig. 2.11b) that the orientation error increases as we enlarge the size of the Gabor filters. As a consequence of this, we use small sized Gabor filters ($\sigma = 0.9$) for all subsequent simulations.

In Fig. 2.12, we show the curvature directions / anisotropy of curvature initially detected by the model and after 10 time steps of grouping and feedback. The initial output qualitatively matches the ground truth except for some noise in areas where the surface curvature is isotropic. Fig. 2.12 shows that the process of directional grouping and recurrent feedback removes noise in these areas and enhances the anisotropy signal. To corroborate this, in Fig. 2.11a we show quantitatively that the orientation error decreases significantly over several steps of iterative feedback and converges after 10 iterations.

In Fig. 2.14, we show the extracted sketch from the input image for different grouping filter ratios. As the elongation ratio of the Gabor filters (grouping stage) increases, more and more collinear features are enhanced and grouped together. By subsequently applying these grouping filters within the recurrent feedback cycle, smoothly connected object

a specific coordinate in latitude/longitude space. In other words, the reflection map is simply used as a lookup table in the rendering process.

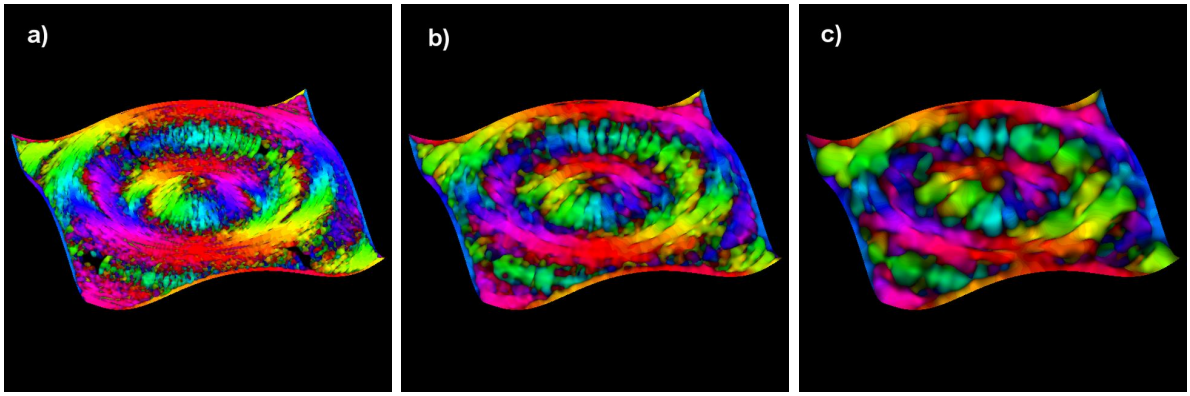


Figure 2.13: Initial detected minimal curvature orientation and anisotropy of curvature for different sizes of the initial Gabor filter stage. In (a) we employed small Gabor filters (4 pixels /cycle), (b) shows results for medium sized Gabor filters (8 pixels/cycle), and (c) shows results for large sized Gabor filters (16 pixels/cycle). We can see clearly, that as the size of the Gabor filter increases we loose more and more detail information. For an explanation of the color code see Fig. 2.12.

structures, such as curvature ridge lines and self-occlusions, are enhanced and completed over time (principle of good continuation).

Fig. 2.15 shows that the quality of the sketch depends on the combination of the two parameters $\alpha$ and $\mu$. The strength of the initial normalization step is controlled by parameter $\mu$ and the strength of the feedback signal is controlled by parameter $\alpha$. In Fig. 2.16 and Fig. 2.17 we employed two different objects rendered under four different surrounding scenes as input for the model. Model simulations show that the extracted sketch images produced from a given shape under different environmental scenes exhibit only marginal differences. Fig. 2.16 therefore gives a very powerful impression that the produced sketch is independent of the surrounding scene. Additional results are also shown in Fig. 2.18.

To demonstrate the performance of the model with images of real world objects, we have also produced sketches from photographs of a kettle and a tap (see Fig, 2.19).

## 2.4 Discussion

We have shown that by making some simple measurements on the image of a mirrored object it is possible to estimate surface curvature properties accurately and reliably. We have shown that these measurements can be performed by extracting and interpreting population codes using simple orientation-selective linear filters. We improved these initial measurements by collinear grouping of perceptually relevant features, which mainly occurs at locations where the surface anisotropy is high. We have shown that recurrent combination of contextual features substantially enhances the quality of the estimates. In addition, it is possible to extract a sketch-like representation from the input image, which is nearly invariant to the surrounding scene. The iterative grouping stage has the effect that isolated errors and noise are reduced in the curvature estimates and that the

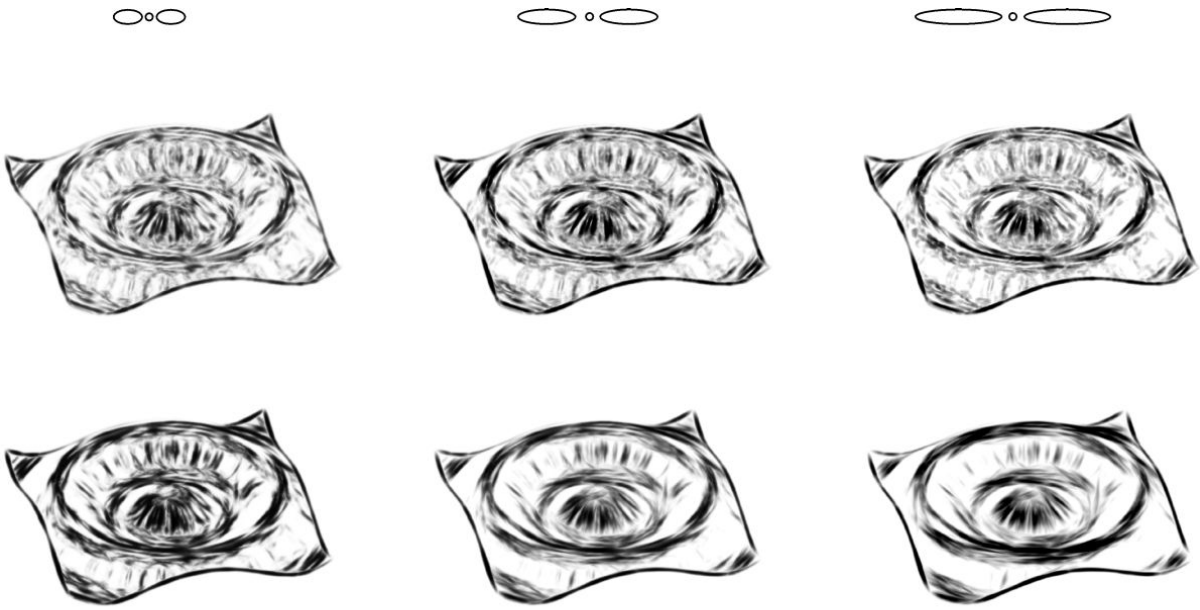

Figure 2.14: Sketch output of the model for different grouping filter elongation ratios. We used elongated Gabor filters with a ratio $\sigma_1 : \sigma_2 = 2{,}4$ and 6 (filter shapes are illustrated scaled uniformly by a factor 3). We show the initial model output (without feedback) (first row) and the output after 10 iterations (second row) for each filter configuration. A low elongation ratio (resulting in a short-range grouping filter) produces a rather noisy initial output whereas a medium or high filter ratio produces more clear and smooth results. Note that even for short-range grouping filters the feedback helps to improve the signal by closing discontinuities between collinear features.

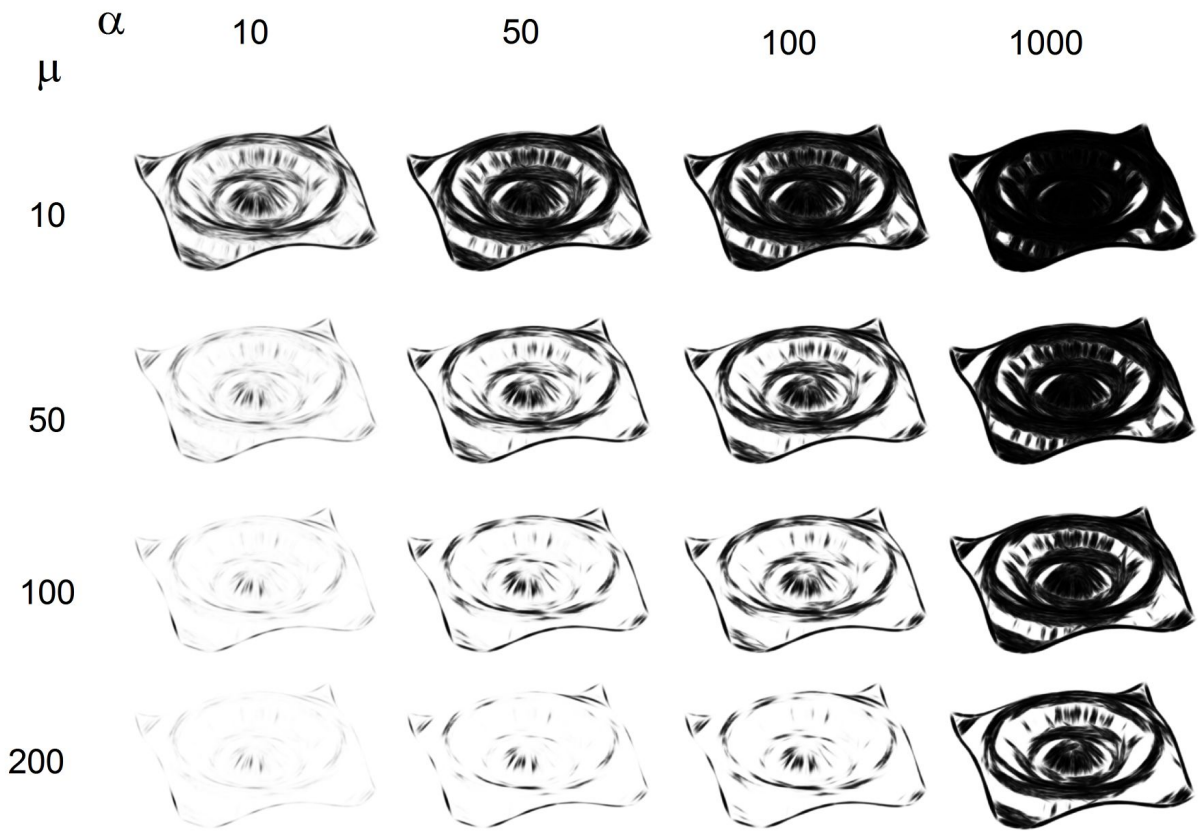


Figure 2.15: This figure shows sketch results after 10 iterations of feedback for different parameterizations of $\alpha$ and $\mu$ where $\alpha$ adjusts the influence of the feedback signal and $\mu$ adjusts the strength of the initial normalization step. The figure illustrates that both parameters have to be chosen carefully in order to receive good looking sketch results. In cases where $\alpha$ is very low the feedback signal has less influence, thus the resulting sketch shows insufficient detail. In cases where $\alpha$ (feedback contribution) is high in combination with a low $\mu$ (normalization) this leads to an overemphasized sketch result. If not mentioned otherwise we used $\alpha = 100$ and $\mu = 100$ for model simulations.

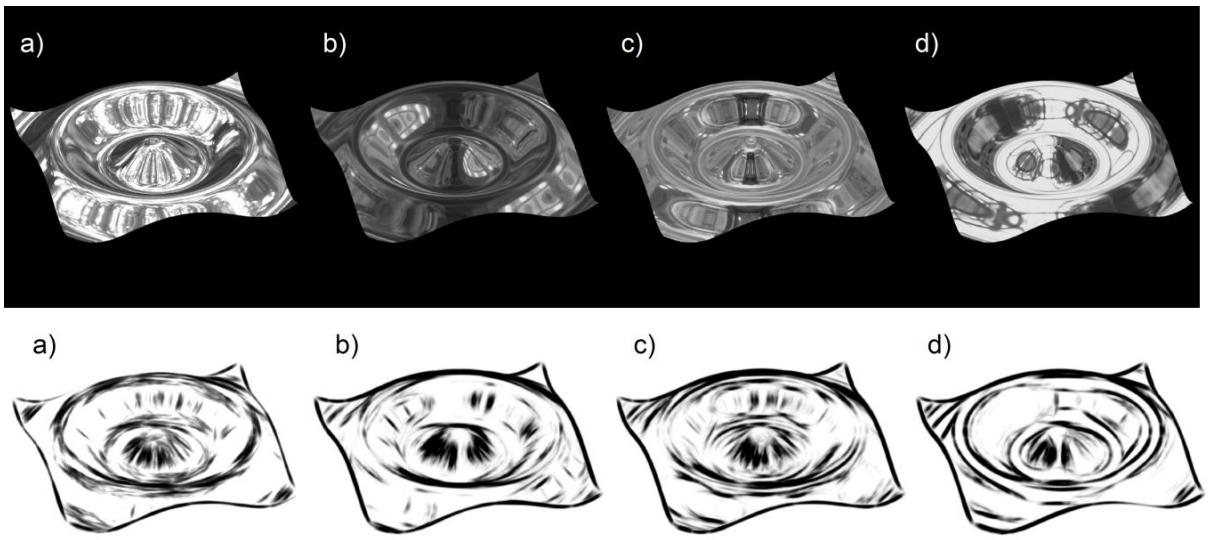


Figure 2.16: Input images rendered with 4 different environmental scenes (first row). The specular objects are rendered with the 'Eucalyptus Grove' scene (a) the 'St Peters Basilica' (b) the 'Galileo's Tomb' (c) and the 'Overcast Breezeway' (d). The reflection maps are taken from [Debevec *et al.*, 2000]. For each different scene we show the sketch output of the model after 10 cycles of feedback (second row). We can see that although the input images look quite different, the sketch output looks very similar in all four cases.

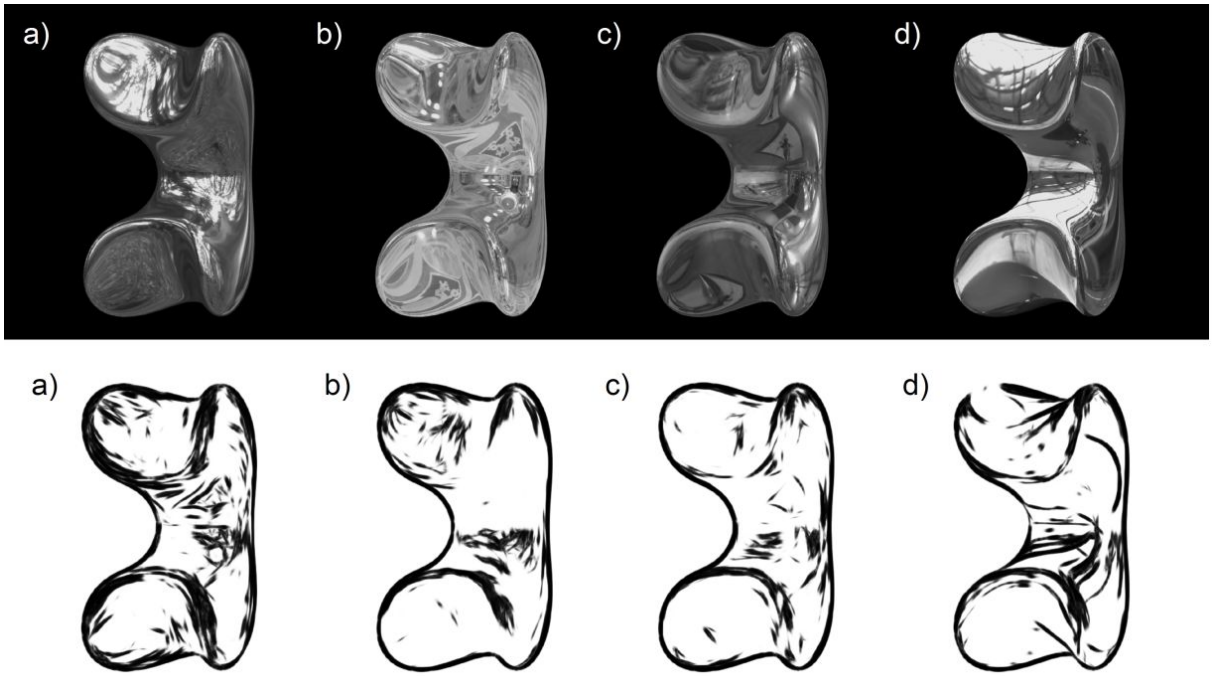


Figure 2.17: We show another mirrored object rendered under four different environmental scenes: (a) 'Eucalyptus Grove', (b) 'St Peters Basilica' (c) 'Galileo's Tomb', and (d) 'Overcast Breezeway'. Model parameters are the same as used in Fig. 2.16. The figure illustrates that the resulting sketches look very similar even though environmental scenes do not perfectly obey the assumption of isotropic scene statistics.

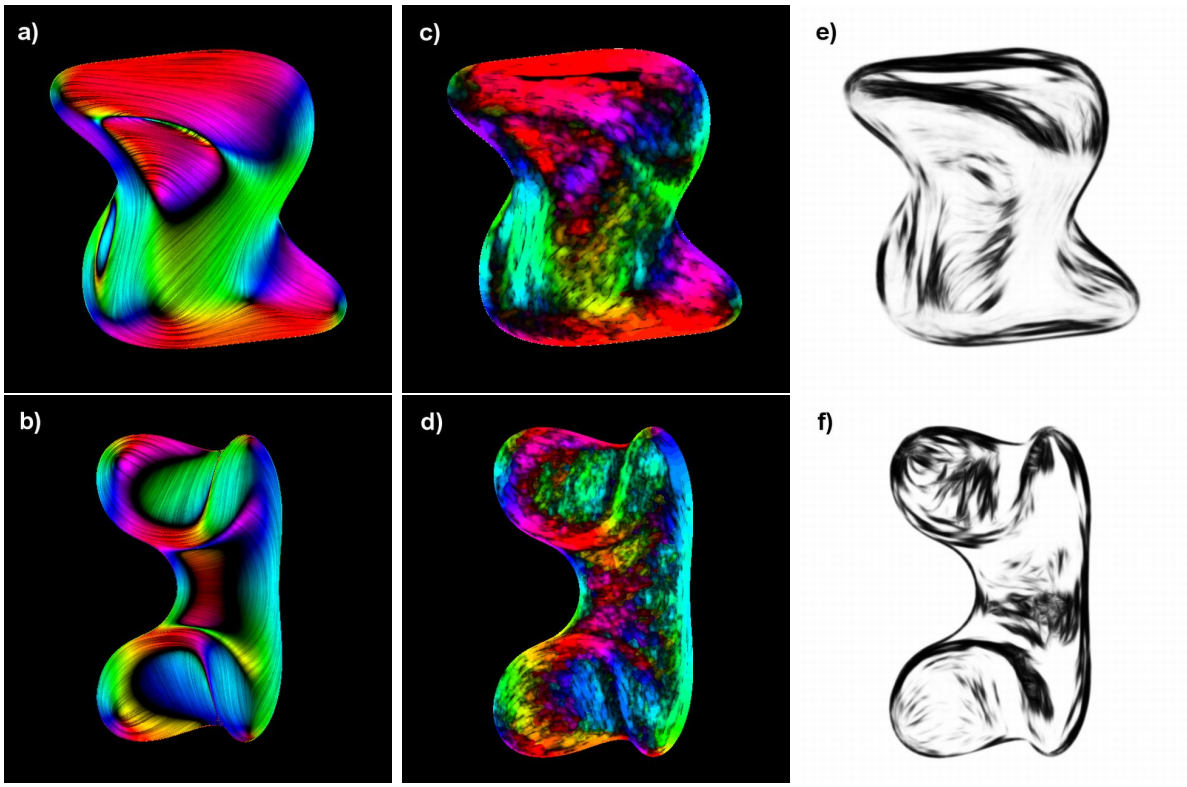


Figure 2.18: Model output for two different objects (rendered with the 'Eucalyptus Grove' scene and parameters ($\alpha = 10$, $\mu = 10$)). We show the estimated minimal curvature orientations in combination with the anisotropy of curvature after 10 iterations of feedback (c)+(d) in comparison to the ground-truth image images (a)+(b). We also show the sketch computed from each object (e)+(f).

Figure 2.19: The figure shows sketch results from images of real world scenes containing shiny surfaces. Here we demonstrate that the model also performs well on images taken from real world scenes. Note that overall the mirrored kettle has relatively isotropic curvature properties, and thus the model tends to produce patterns that depend on the environmental scene.

extracted sketch becomes more distinct.

The model dynamics were developed in the context of contour perception by [Neumann and Sepp, 1999]. We bring together and further extend previous proposals on perceptual grouping of surface features [Grossberg and Mingolla, 1985] with recent proposals of human shape perception of specular surfaces [Fleming *et al.*, 2004]. Our model employs simple biologically motivated mechanisms as it is well known that primary visual cortex contains cells that are tuned to different image orientations [Hubel and Wiesel, 1968].

It is important to note that our model output can only provide *constraints* on 3D shape, rather than a complete estimate of the shape model. Orientation fields are inherently ambiguous. For example, convex and concave surface patches cannot be distinguished locally. However, human vision almost certainly applies additional constraints to resolve these ambiguities, e.g. by enforcing boundary conditions based on occluding contours [Koenderink, 1984], or by enforcing global constraints, such as smoothness.

Our model naturally extracts object structures such as smooth occlusion contours and curvature ridge lines. These features have been described by [Todd, 2004] as typical features of line drawing showing 3D objects. Indeed, the fact that image orientations tend to align with these shape features may provide an explanation of *why* artists choose to depict these features in particular. Our simulations also demonstrate that the model sketch can be used as an abstract representation of the 3D object, invariant to the reflected scene (Fig. 2.16). The sketch tends to emphasize regions of high curvature. Previous work on the perception of 2D shape, has suggested that high curvature regions are the most informative locations on a 2D curve [Attneave, 1954; Feldman and Singh, 2005].

Previous work on 3D shape visualization has suggested the utility of aligning texture with the intrinsic principal directions of the surface [Interrante and Kim, 2001; Interrante *et al.*, 2002; Kim *et al.*, 2003]. Here, we suggest that the principal directions defined in *view-centered* coordinates may also be useful for conveying surface shape. Note that areas of high curvature tend to bear patterns of high spatial frequency, because this is where the reflection is most compressed. By emphasizing fine structures the extracted sketch also looks subjectively like a charcoal drawing.

### 2.4.1 Limitations of the model

Although our model behaves well for most objects and surrounding scenes there are of course some special cases where the model fails to provide accurate estimates. As mentioned earlier, our model implicitly assumes the surrounding scene is isotropic. If we infringe this assumption which we tacitly do when we use real-world scenes, this leads to errors in our curvature measures and also in the sketch generation (see Fig. 2.20). It has been shown by Fleming, Torralba & Adelson, [Fleming *et al.*, 2009] that this is no problem as long as the surface anisotropy is strong enough to overcome the anisotropies in the surrounding scene. However, when the surface is locally isotropic, such as in the central region of a sphere, the orientation measurements are dominated by the contours

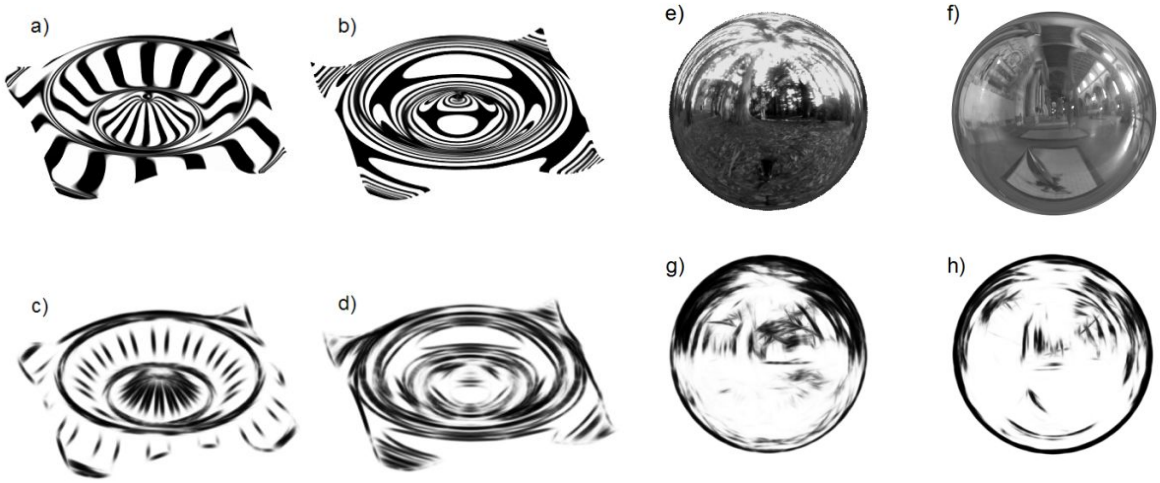


Figure 2.20: (a) and (b) show an object mirrored with an artificial scene that consists of vertical stripes (a) and horizontal stripes (b). This artificial environment represents a very unnatural scene with purely anisotropic scene statistics (which strongly infringes the assumption of an isotropic scene). Hence, the resulting sketches in (c) and (d) look very different from one another. Despite this, it is observable that on highly curved areas on the surface (curvature ridges) both sketches show similar striped patterns. (e) and (f) show a sphere illuminated with the 'Eucalyptus Grove' scene (e) and the 'Galileo's Tomb' scene (f) from [Debevec *et al.*, 2000]. Apart from the outer rim, a sphere has roughly isotropic curvature properties. Thus, the sketches in (g) and (h) show rather different orientation patterns, which are not invariant to the surrounding scene.

of the objects in the surrounding scene. Nevertheless, even weak curvature anisotropy biases the texture distortion towards the direction of minimal surface curvature. Furthermore, on complex smoothly curved surfaces, these isotropic conditions appear relatively rarely, generally only at singular points (in the center of humps or dimples) or along lines of inflection points of curvature (where one curvature component changes its sign). Assuming ordinary real world scenes, the reflections on such areas are not characterized by smoothly curved streaks but rather by a noisy orientation field. Consequently we designed our model to weaken such areas by introducing the grouping component in combination with the recurrent feedback signal.

### 2.4.2 Generalization of the model

Most objects in our environment do not have a perfectly specular surface. Indeed, this type of surface reflectance property where incoming light is only reflecting in one single direction is quite unusual. Most materials scatter light in many directions, leading to a continuous shading pattern on the surface. However, [Fleming *et al.*, 2004] showed that orientation fields also remain somewhat stable across changes in material. In future work, we intend to extend our model to deal also with rough or diffuse materials like brushed aluminum or chalk. Our model could also be extended to interpret specular highlights which they have the same characteristics as perfectly specular reflections except that they appear only at isolated locations.

In summary, we have shown a novel architecture of shape processing from specular objects that combines orientation selective filtering [Fleming *et al.*, 2004] with recent proposals on perceptual grouping and feedback processing [Neumann and Sepp, 1999]. Our model can be used to reliably extract the principal curvature directions of smoothly curved specular surface and to generate a non-photorealistic sketch which is invariant to the surrounding scene.

# Chapter 3

# Extraction of Surface-related Features

## 3.1 Introduction

Our visual system structures the visual world into surfaces that, if required, we recognize as familiar objects. A fundamental task of vision therefore is to find the boundary contours separating the regions corresponding to surfaces or objects. As our retina captures only a 2D projection of the 3D world, mutual occlusions are a natural consequence which can be interpreted by the visual system as a cue to relative depth. A vivid demonstration of surface-based depth perception is given by a painting of a professional artist who tries to depict a scene where the visual system generates surface segmentations in the presence of multiple occlusions (Fig. 3.1). However, it remains unclear what particular features are used by the visual system to detect occlusions and whether this information is derived locally or from more global criteria. Some recent evidence [Rubin, 2001b; Nakayama *et al.*, 1995; McDermott and Adelson, 2004] suggests that the human visual system might use surface-related features that are specific contour junctions that have a surface-based relevance in scene interpretation.

In this chapter, we propose a neural model that suggests how surface-related features can be extracted from a 2D luminance image. The approach is based on contour grouping mechanisms found in visual cortical areas V1 and V2. Our computational model comprises the extraction of oriented contrasts which are subsequently integrated by short- and long-range grouping mechanisms to generate disambiguated and stabilized boundary representations. We argue that the mutual interactions realized by lateral interactions and recurrent feedback between the cortical areas considered stabilize the representation of fragments of outlines and group them together. Moreover, we demonstrate that the model is able to signal and complete illusory contours over a few time-steps. Illusory contours are a form of visual illusion where contours are perceived without a luminance or color change across the contour. Such illusory contours can be induced by partially occluded surfaces where the contour of the occluded object is perceptually completed (amodal completion) or where the occluding object has the same luminance than parts

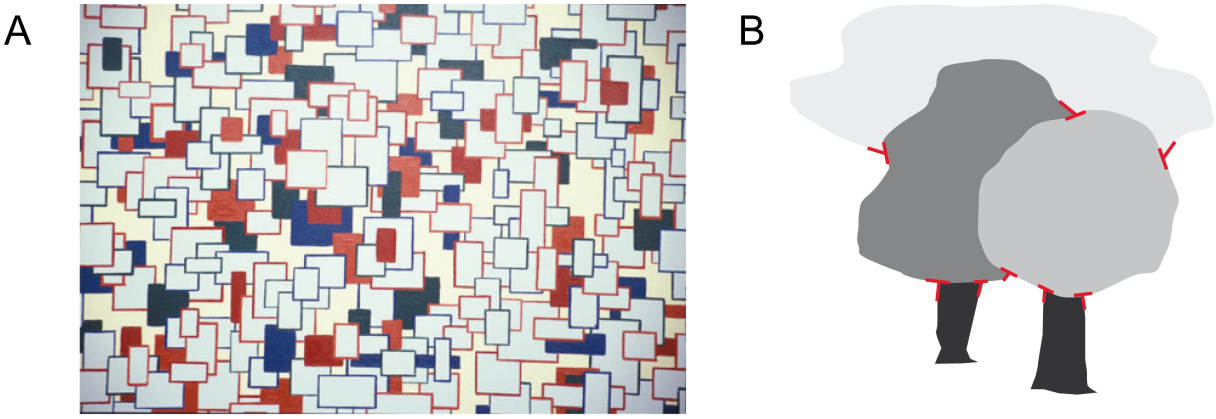


Figure 3.1: (A) Painting of a professional artist [Marrara, M., 2002, reproduction with permission from the artist] that leads to the perception of different depths induced by occlusion and color cues. Notice how hidden surface parts are perceptually completed by the human visual system in order to segregate surfaces apart from each other. Surfaces can also be associated with (parts of) objects in scenes depicted by trees and clouds in (B). A human observer could use local cues such as T-junctions (red) formed by the boundary contour of surface parts to detect surface occlusions and hence to infer depth from monocular scenes.

of the occluded background (modal completion). Illusory contours play a significant role in the perceptual interpretation of junction features. For instance, it was suggested by Rubin [Rubin, 2001b] that the perception of occlusion-based junctions (T-junctions) can be induced by L-junctions in combination with the presence of illusory contours.

Consistently, in our model junction signals are read out from completed boundary groupings which are interpreted as intermediate-level representations that allow for the correct perceptual interpretation of junctions, namely L-junctions features can be perceptually interpreted as T-junctions. This is unlike previous approaches which are based on purely feature-based junction detection schemes [Harris and Stephens, 1988; Smith and Brady, 1997].

Taken together, our proposed model suggests how surface-based features could be extracted and perceptually interpreted by the visual system. At the same time, this leads to improved robustness and clearness of surface-based feature representations and hence to an improved performance of extracted junction signals compared to standard computer vision corner detection schemes. Based on these perceptual representations, surface-related junctions are made explicit such that they could be interpreted to interact as to generate surface-segmentations in static or temporally varying images.

## 3.2 Model

In this section we give a short overview of the proposed model and its components. Our model focuses on the early processing stages of form processing in primate visual cortex, namely cortical areas V1 and V2, and incorporates hierarchical feedforward processing as well as top-down feedback connections to consider the signal flow along the reverse

hierarchy processing [Ahissar and Hochstein, 2002]. An overview of the model architecture is depicted in Fig. 3.2.

### 3.2.1 Overview of the model architecture

The model has been structured into three main components. The first component comprises initial feedforward processing. The monochromatic input image is processed by a cascade of different pools of model cells in V1, specifically simple, complex, end-stop, and bipole cells. Each cell population consists of cells that are tuned to different orientation selectivity. This is consistent with the representation of orientation selective cells found in V1 which are arranged in hypercolumns [Hubel and Wiesel, 1968]. Our model cell types are responsive for specific local image structures, i.e., simple and complex cells are sensitive to oriented contrast represented by edges or bar elements, end-stop cells respond best to contour terminations that occur, e.g., at line ends or corners, and bipole cells are sensitive for collinear arrangements of contour fragments with similar orientation. Model area V2 receives forward projections from V1 bipole cells and V1 end-stop cells. These signals are then integrated by long-range V2 bipole cells which have a larger extent than bipole cells in model V1. Bipole cells in V2 respond to luminance contrasts as well as to illusory contours, thus resembling functional properties of contour neurons in V2. This finding suggests that orientation selective mechanisms for contour integration in area V2 do not simply represent a scaled version of V1 mechanisms for lateral contrast integration. Their capability to integrate activities to bridge gaps and generate illusory contours makes an important step towards surface boundary segregation while V1 contrast integration is selective to stimulus feature processing.

The second model component comprises recurrent feedback processing between model areas V1 and V2. Neurons in V1 are also responsive to more global arrangements of the scene [Lamme and Rolfsemma, 2000]. These response properties possibly arise from recurrent processing and lateral connections from pyramidal neurons [Hupé *et al.*, 1998]. Whereas feedforward connections have mainly driving character, feedback connections are predominantly modulatory in their effects [Hupé *et al.*, 2001]. There is evidence that feedback originating in higher level visual areas such as V2, V4, IT or MT, from cells with bigger receptive fields and more complex response properties can manipulate and shape V1 responses, accounting for contextual or extra-classical receptive field effects [Hirsch and Gibert, 1991; Salin and Bullier, 1995; Sillito *et al.*, 2006].

We account for these findings by incorporating a recurrent interaction mechanism between model areas V1 and V2. In our model, activity in V2 serves as top-down feedback signal to iteratively improve initial feedforward activity in V1. The feedback signals that are delivered by descending cortical pathways multiplicatively enhance initial activities at earlier processing stages. Importantly, this type of feedback is not capable of generating new activity at positions with zero initial activity which could lead to an uncontrolled behavior of the overall system's functionality. Feedback can only *modulate* activity that is already present at V1 [Hupé *et al.*, 1998]. We shall demonstrate in our results that multiple

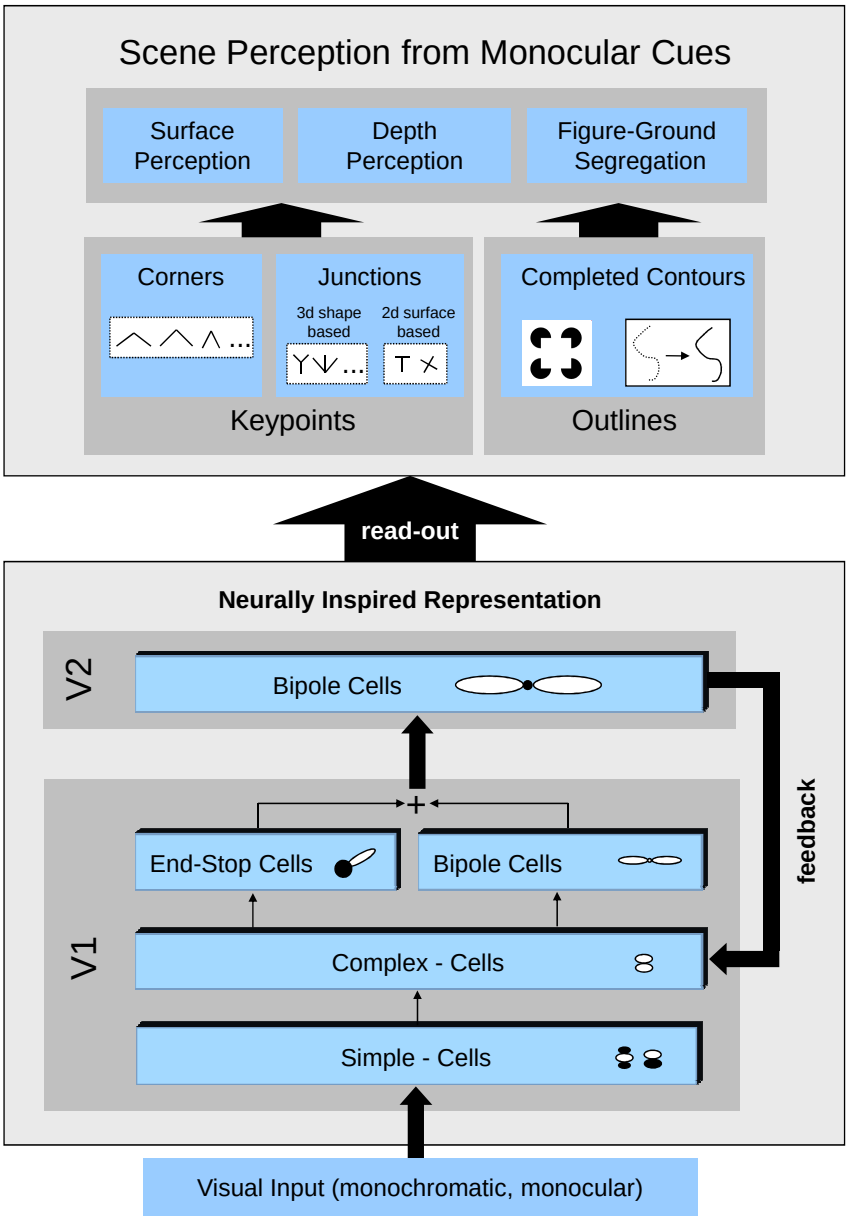


Figure 3.2: Overview of the core model architecture. The model consists of several stages that were designed to resemble properties of cells found in the early primate visual cortex. Visual input is processed by the hierarchy of different stages from visual area V1 to V2 and vice versa, that is feedforward and feedback. To enhance and complete initial contour signals, recurrent interactions between those two areas are performed iteratively until activities at all stages converge to a stable state. The converged activities can then be read-out from distributed representations to obtain specific maps that signal perceptually important image structures such as completed contours and different types of junction configurations. Such mid-level features provide important cues for occlusion detection or detection of transparencies. In addition, these mid-level features can also play a role in tasks such as border-ownership assignment which perhaps take place in higher visual areas such as V4 or IT.

iterations of feedforward-feedback processing between model areas V1 and V2 lead to clearly more consistent and stable results compared to purely feedforward processing schemes.

The third component of the model comprises the extraction or *read-out* of scene relevant information that is provided by different pools of cells within the two model areas V1 and V2. Fig. 3.2 presents an overview of the different types of mid-level features that can be extracted from the distributed representation of cell responses. This includes the extraction of several maps that signal contours, illusory contours, and key points characterized by different junction configurations. It has been stressed by several authors that specific junction configurations like T- or X-junctions provide important cues for the discovery of occlusions or transparency [Rubin, 2001b] in the context of surface segmentation. Therefore, we suggest that the visual system uses specialized mechanisms to read out separate maps for such configurations. It is important to note that our model is *not* only a simple key point detector. In fact, our model additionally provides structural information about key points represented by activities of model cell pools located at the key point.

### 3.2.2 Detailed description of model components

In this section, we explain the individual model parts in more detail. For a precise mathematical description of the model and its different processing stages the reader is referred to appendix S1. The detailed model architecture is illustrated in Fig. 3.3.

#### Model area V1

The initial feedforward stage represents early visual mechanisms in area V1 and V2 of the primate visual cortex. We do not simulate processing at earlier stages of the visual pathway such as LGN or the retina since this would not considerably influence our results.

In our model, a first step is to simulate pools of cells that encode at each position of the input image oriented luminance contrasts which are represented by *V1 simple* cells in the primary visual cortex [Hubel and Wiesel, 1968]. Each model cell represents the average responses (or firing-rate) of groups of neurons with similar response properties. We simulate two different types of simple cells with even and odd symmetrical receptive field (RF) properties. These orientation selective cells respond best to oriented line segments or edges, respectively. Simple cells are also selective to contrast polarity, such that they signal light and dark bars as well as light-dark and dark-light transitions. We do not, however, keep this information separate but combine this information to yield a representation of the local contrast energy, which is invariant against the sign of contrast. This is motivated by the fact that the model tries to explain computational stages to form unsigned boundaries, and thus should be invariant to contrast polarity [Grossberg and Mingolla, 1985]. In addition, such a convergence of activity coheres with current models of hierarchical feedforward processing of binocular input of disparity sensitive cells in primary visual cortex [Ponce and Born, 2008].

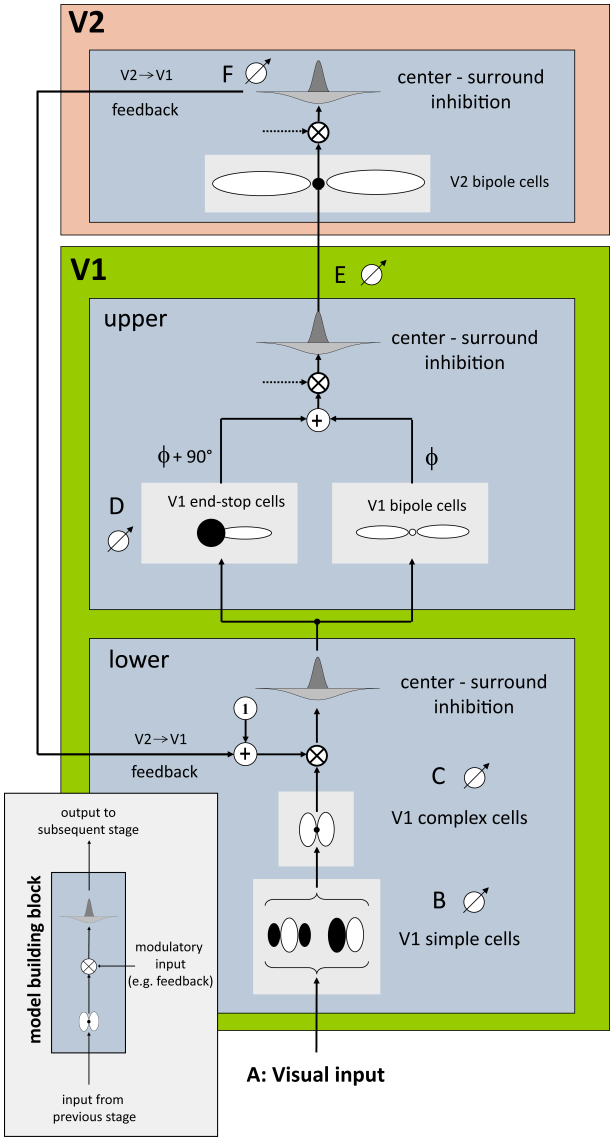


Figure 3.3: Detailed model architecture. Our model simulates cells of two areas in the visual cortex, visual areas V1 and V2. Each model (sub-) area is designed with respect to a basic building block scheme (right). The scheme consists of three subsequent steps, namely filtering, modulation and center-surround inhibition. This scheme is applied three times in our model architecture (left), corresponding to upper and lower area V1 and area V2. In this model, modulatory input (provided by feedback from area V2) is only used in lower area V1. Otherwise the default modulatory input is set to 1 (which leaves the signal unchanged). The lower part of area V1 is model led by simple and complex cells for initial contrast extraction. Note, that each cell pool consists of 12 oriented filters equally distributed between 0° and 180°. The upper part of V1 is model led by end-stop and bipole cells which both receive input from lower V1. The additively combined signals are further passed to area V2 where long-range lateral connections are model led by V2 bipole cells. Note, that "•" stands for a multiplicative connection of filter sub-fields as employed in V2 whereas "o" stands for an additive connection as employed in V1. Finally output of area V2 is used as feedback signal which closes the recurrent loop between areas V1 and V2.

*V1 complex cells* pool activity of two equally oriented V1 simple cells of opposite polarity. Thus, complex cell activity is invariant to contrast polarity which resembles response properties of real complex cells. The output signals of model complex cells subsequently undergo a center-surround inhibition that is realized by a lateral divisive inhibition mechanism. In the literature, this divisive type of lateral inhibition is also termed *shunting inhibition* [Sperling, 1970]. This mechanism leads to a competition of cell activities within a neighborhood that is defined over the spatial and orientation domain. High activity of multiple orientations leads to a suppression of overall cell activity whereas activity in a single orientation channel leaves cell activity relatively unchanged. As a consequence, responses in areas with undirected structure such as textured or noisy areas are weakened by this operation. On the other hand, responses in areas with directed structures, such as edges and lines are strengthened by this operation. Such a stage of divisive inhibition has been previously proposed to account for non-linear effects in contrast and motion responses of V1 cells [Heeger, 1992; Caradini and Heeger, 1994; Tolhurst and Heeger, 1997].

In the next step of the hierarchical processing scheme two different populations of cells receive forward projections from V1 complex cells. The first population of model cells resembles long-range lateral connections found in V1 [Gilbert and Wiesel, 1989; Schmidt *et al.*, 1997]. These long-range connections are modeled by *V1 bipole cells* which consist of two additively connected elongated Gaussian sub-fields. The spatial layout of the filter is similar to the bipole filter as first proposed by [Grossberg and Mingolla, 1985]. The spatial weighting function is narrowly tuned to the preferred orientation, reflecting the highly significant anisotropies of long-range fibers in visual cortex [Bosking *et al.*, 1997]. Here, we parameterize the length of a V1 bipole cell about 2 times the size of the RF of a complex cell.

The second population receiving input from complex cells are *V1 end-stop cells*. End-stop cells respond to edges or lines that terminate within their RF. This includes also corners or junctions where more than one contour ends at the same place. However, at positions along contours or at X-junction configurations, end-stop cells do not respond. Such types of cells have been first observed in cat striate cortex [Hubel and Wiesel, 1965]. More recently, evidence for end-stop cell properties of V1 neurons was found in several physiological studies [DeAngelis *et al.*, 1994; Jones *et al.*, 2001; Sceniak *et al.*, 2001].

In our model, end-stop cells are modeled by an elongated excitatory sub field and an inhibitory isotropic counterpart [Thielscher and Neumann, 2008]. Our model end-stop cells are direction sensitive and are therefore modeled for a set of directions between 0 and 360 degrees. Activities of end-stop cells corresponding to opposite directions are additively combined in order to achieve direction invariance. Finally, at the output of model area V1 activities from V1 bipole and V1 end-stop cells are merged and normalized by a center-surround inhibition stage before they are forwarded to model area V2.

### Model area V2

Visual area V2 is the next stage in the hierarchy of processing stages along the ventral stream. Several physiological studies on macaque monkeys have shown that cells in V2 respond to luminance contrasts as well as to illusory contours [Heitger *et al.*, 1998; von der Heydt *et al.*, 1984]. In contrast to complex cells in V1 they respond much stronger if the luminance contrast is continuous, and less if gaps are between inducers, thus resembling functional properties of contour neurons in V2. Moreover, they respond to moderately complex patterns such as angle stimuli [Ito and Komatsu, 2004]. However, the precise functional role of area V2 remains unclear.

In our model, we employ V2 bipole cells with elongated sub-fields which are collinearly arranged and centered at the reference position. The sub-fields sample the input activations generated by V1 bipole cells and V1 end-stop cells. Responses of the individual sub-fields are multiplicatively combined [Neumann and Sepp, 1999] such that the net effect of the contrast feature integration leads to an AND-gate of the individual sub-field activations. Thus, activity from both sub-fields of an integration cell is necessary to generate cell activity. This non-linear connection has the effect that activity can emerge between two or more like-oriented contour-fragments or line ends at positions where no initial luminance contrast is present which is indicative for the presence of an illusory contour.

At the same time, activity of a V2 bipole cell at an isolated contour termination would be zero as one sub-field does not receive any input. V2 bipole cells are additively combined with perpendicular oriented V1 end-stop cells. This has the effect that V2 bipole cells can integrate activity of end-stop cells along line terminations that are linearly arranged, which leads to the impression of an illusory contour. Such a mechanism of *ortho-grouping* has been proposed earlier by Heitger and colleagues [Heitger *et al.*, 1998].

### Recurrent V1-V2 feedforward-feedback interaction

In our model, lower area V1 and area V2 interact in two directions, that is feedforward and feedback. Feedforward interaction is realized by feeding bottom-up input activation from model V1 to V2 and was described in detail in the last sub-section. On the other hand, feedback interaction is realized by top-down *modulatory* feedback connections that deliver signals from model V2 to V1. The recurrent loop is closed at the feedback re-entry point in V1 where initial feedforward complex cell responses and V2 bipole cell responses are multiplicatively combined (Fig. 3.3).

In order to allow feeding input signals to be propagated, even in the case that no feedback signals exist, the feedback signal is biased by a constant unit value. This bias introduces an asymmetry for the roles of forward signals and feedback processing. Feedforward signals act as drivers of the hierarchical processing scheme, whereas feedback signals generate an enhancing gain factor which cannot on its own generate any activity at positions where no initial feeding input responses are present. This realizes a variant of the no-strong loop hypothesis [Crick and C., 1998] to avoid uncontrolled behavior of the

overall model dynamics and to limit the amount of inhibition necessary to achieve a stable network performance. Several physiological studies support the view that, e.g., feedback from higher visual areas is not capable of driving cells in lower areas, but modulates their activity [Hupé *et al.*, 1998; Salin and Bullier, 1995].

It is important to mention that modulatory feedback in a recurrent loop only works correctly in combination with a suitable inhibition mechanism. Otherwise, the feedback signal would lead to uncontrolled growth of model cell activities. Thus, both mechanisms, the modulatory V2→V1 feedback interaction and the subsequent shunting lateral inhibition work in combination in order to enhance distributed contour and junction representations in model V1 and V2 which mutually support each other considering a larger spatial context. At the same time, mainly through the action of the divisive inhibition mechanism, the overall activity in a pool of cells is kept within a maximum bound which stabilizes the network behavior and prevents the energy from getting too excited. In addition, those feeding activities that receive no amplification via feedback signals will be less energetic in the subsequent competition stage. Consequently, their activities will be finally reduced, which realizes the function of biased competition which has been proposed in the context of modulation in attention effects [Desimone and J., 1995].

### 3.2.3 Read-out and interpretation of model activities

From the distributed representation of cell responses in both model areas V1 and V2 several retinotopic maps can be extracted that signal perceptually relevant contour configurations. If not mentioned otherwise, these maps are extracted by computing at each position the mean activity of all orientation responses. An alternative method for reading out salience values was suggested by Li [Li, 1999], who choose to extract at each position the maximum activity over all orientations. In the following, we describe in detail how saliency maps for specific image structures, namely corners and junctions can be extracted by combining activities from different model cells pools.

In this paper, we define saliency maps as 2d maps that encode at each position the likelihood that a specific structure is present. A more broad discussion on the concept of salience and salience maps can be found in [Zhou *et al.*, 2000]. In Fig. 3.4 the structural configurations are sketched to present an overview of the output as signaled by the different orientation sensitive mechanisms of the proposed model. This summary indicates how the different visual structures of surface shape outlines and their ordinal depth structure might be selectively encoded neurally through the concert of responses generated by different (model) cell types.

The conclusions are two-fold. First, it is indicated that the presence of, e.g., a T-junction (which most often coheres with an opaque surface occlusion[Rubin, 2001b]) is uniquely indicated by the response pattern of V1 and V2 cells at one spatial location. The T-junction is represented by an end-stop cell response at the end of the T-stem, V1 bipole cell responses in the orientations of both the T-stem (signaled by one active subfield) and the roof, and finally a V2 bipole cell response in the orientation of the roof of

| | cue type | 2d-corner | 3d-corner | transparency | occlusion | contour | illusory contour |
|---|---|---|---|---|---|---|---|
| model area | structure / cell type | L | Y, Ψ | X | T | | |
| V1 | end-stop | 2 | 3 | - | 1 | - | - |
| | bipole | 2 | 3 | 2 | 2 | 1 | - |
| V2 | bipole | - | - | 2 | 1 | 1 | 1 |

Figure 3.4: Response properties of different model cell populations for different structural configurations together with their most likely interpretation (cue type). Numbers denote the modality of the response distribution across cell pools located at the position marked with a red dot for each structure. A bar means that the cell population is not responsive for this structure. Note, that each structure has a specific neural response profile across different model cell populations which can be used to extract separate saliency maps. For a better understanding, we sketched the configuration of filters together with the underlying structure. Remember, that V2 bipole sub-fields are connected multiplicatively (signaled by a "•"), leading to zero activity of the whole bipole cell if input from one sub-field is missing (symbolized by red crosses). On the other hand, V1 bipole sub-fields are additively connected (signal led by a "o") which has the effect that input from one sub-field is sufficient to create activity.

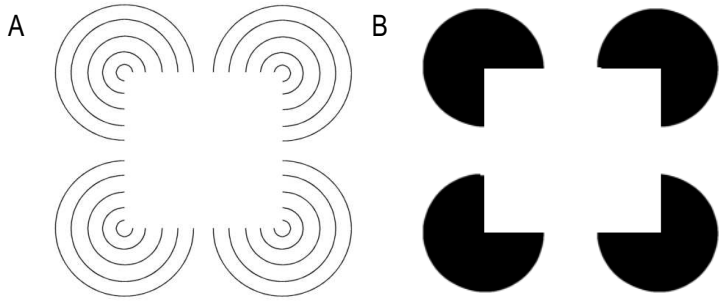


Figure 3.5: Stimuli used in our experiments that show illusory contours effects. In both, the Kanizsa shape and the five-line Varin shape, observers have the impression of seeing a white square which is partly formed by illusory contours. There is strong evidence that illusory contours are induced by horizontally and vertically arrange line-endings (A) or by collinearly arranged luminance contrasts (B).

the T (representing the occluding boundary). Second, we argue in favor that no explicit *detectors* are needed to represent those local 2D structures. Fig. 3.4 indicates that the explicit representation of different junction types necessitates a rich catalog of cells with rather specific wiring patterns. Below we propose specific read-out mechanisms in order to visualize the information we suggest is important for surface-related analysis of the input structure.

### Contours / Illusory contours

Contours are basic image structures which are important for the segmentation of surfaces by generating a likelihood representation of the locations and orientations of the shape outline boundaries. Furthermore, contours mark the border between two adjacent surfaces and play a major role in the process of figure-ground segregation and border-ownership [Zhou *et al.*, 2000; Zhaoping, 2005; Rubin, 2001a]. In our model, contour saliency is encoded in the response of V1 bipole cells. The contour saliency map can be extracted by summing activity of orientation selective bipole cells pools in V1. Illusory contours are a form of visual illusion where contours are perceived without a luminance change across the contour. Classical examples are the Kanizsa figure [Kanizsa, 1955] where an illusory square is induced by four flanking pac-man symbols or the Varin figure [Varin, 1971] where linearly arranged line-ends mark the borders of the illusory square (Fig. 3.5). There is evidence that illusory contours are represented by V2 neurons [von der Heydt *et al.*, 1984]. In our model, illusory contours are signaled by activity of V2 bipole cells.

### T-junctions

A T-junction is formed when a contour terminates at a differently oriented continuous contour. T-junctions most often provide local evidence for occlusions as they frequently occur when a surface contour is occluded by another opaque surface in front. At the point where the bounding contour of the occluded surface intersects with the bounding

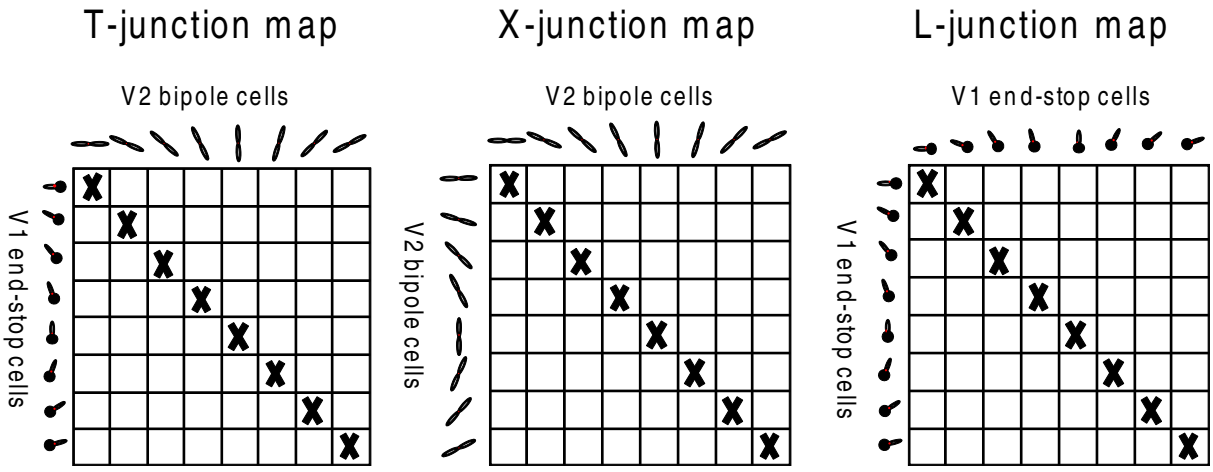


Figure 3.6: Extraction of junction signals. The figure depicts how local activity from model cells is combined to obtain specific junction maps. To extract a T-junction signal orientation specific V2 bipole cell activities are combined pair wise with orientation specific V1 end-stop cell activities. Pairs of cells that correspond to same orientations are not considered (diagonal marked with "X"). The combination of cell activities is done multiplicatively such that both cells have to be active in order to produce a response. This can be represented in a map where each entry corresponds to a specific T-junction configuration. X-, and L-junction maps are extracted similarly except that pair wise combination is based on V2 bipole cells (X-junctions) and V1 end-stop cells (L junctions). To summarize, the maps represent information about local image structure with respect to different junction types. Note, that end-stop cells are only indicated for one direction in the figure. However, end-stop cell responses of both directions that correspond to one orientation are additively combined for the extraction of junction maps.

contour of the occluding surface a T-junctions is formed in the image which is dependent on the position of the viewpoint. It has been suggested by Rubin [Rubin, 2001b] that T-junctions play a central role in monocular depth perception, surface completion, and contour matching.

To read-out T-junction signals we multiplicatively combine activities of V2 bipole cells with V1 end-stop cells. More precisely, we use a pair wise combination of orientation specific V2 bipole cell activities and orientation specific V1 end-stop cell activities. Pairs of cells that correspond to same orientations are not considered. The multiplicative operation realizes an AND-gate, implying that both cell populations have to be active in order to signal the presence of T-junctions. A saliency map that signals the presence of T-junctions is then extracted by summing over the orientation domain. In general, this map represents the priority of the gathered evidence in favor of the particular scene feature.

### L-Junctions

L-junctions, also termed V-junctions or corners, are formed by two contour segments which terminate in the same projected location. We extract corner signals in a similar way than T-junctions signals. Instead of combining V1 end-stop activities with V2 bipole activities we multiplicatively combine activities of differently oriented V1 end-stop cells among each other (Fig. 3.6). A combined orientation invariant corner saliency signal is then extracted

by integrating over all combinations. Corners are important cues for shape perception as they mark key points of the boundary contour. Importantly, an L-junction can also result from two occluding surfaces, assuming that one of the surfaces is partly formed by illusory contours [Rubin, 2001b]. Under such circumstances, an L-junction feature would as well suggest for occlusions. We shall show that our model initially detects an L-junction in this case, but over time when groupings could be established, then these types turn into perceptual T-junctions (irrespectively whether the boundaries are formed by physical luminance contrasts or by illusory contours). The perceptual representation of T-junctions in turn signals the presence of an occluding surface which is consistent with the impression that human observers report when they are confronted with illusory figures, such as Kanizsa squares [Kanizsa, 1955].

### X-junctions

X-junctions configurations appear at positions where two contours intersect. In scenes with overlapping surfaces, this pattern is created when an occluding surface has transparent material properties leading to a visibility of the occluded surface region through the surface at the occluded surface contour. Therefore, in the transparent occlusion situation, a T-junction turns into an X-junction. In our model, X-junctions are read-out by multiplicatively combining activities of pairs of differently oriented V2 bipole cells (Fig. 3.6). The saliency map is obtained by summing over the orientation domain. Again, the multiplicative connection acts like an AND-gate which extracts only those V2 bipole responses that have a bimodal activity distribution in the orientation.

### Y- and W-junctions

Y- and W-junctions are strong cues for 3D-corners induced by surface intersections of 3D objects which cut in a single location. For instance, a cube produces a Y-junction at the position where the corners of three visible surfaces meet. The same corner observed from another viewpoint turns into a W-junction. Notably, in rare cases, such junctions can be also produced by occluding 2D surfaces when a contour of an occluded surface meets a shape-based L-junction of an occluding surface. However, this can be seen as an 'accidental' and rather unstable configuration since even small changes in viewpoint would lead to vanishing of such occlusion-based junctions. Since the perception of 3D objects is not our primary focus in this thesis, we do not take Y- and W-junctions into further consideration.

### Competition between junction signals

In order to suppress ambiguous activations for more than one junction type at the same place, junction signals compete with each other through lateral inhibition (Fig. 3.7). If one junction type is activated the other junction signals in a local neighborhood are weakened. Finally, all junction activities are passed through a non-linear saturation function in order to have the same range for all activity signals. Note, that although we use a similar

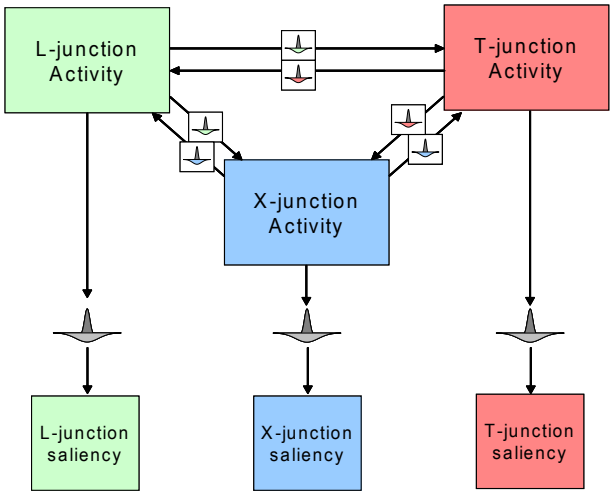


Figure 3.7: Competition between junction signals. Junction signals can locally compete with each other to avoid ambiguous signals. In addition, center-surround inhibition helps to suppress multiple junctions in a small neighborhood which could be induced by fine texture or noise.

inhibition scheme than for the model activities we do not claim that this kind of competition has a biological counterpart. This is just a necessary operation to disambiguate feature signals and has no biological relevance.

## 3.3 Results

In this section we present results of the model simulations in order to demonstrate the computational capabilities of the proposed model. We begin by presenting model results from the various neural cell pools that are simulated in the model. We also show how recurrent feedforward-feedback interaction helps to enhance and stabilize the initial responses of cells. Then we show that various feature maps can be extracted from the distributed cell activities suggesting that this representation is capable of providing cues that are perceptually relevant for fundamental visual processes such as occlusion detection.

### 3.3.1 Robustness to noise

In a first simulation an artificially created image of a noisy square is employed to demonstrate the robustness of the model against noise perturbations. The image was created such that the standard deviation of the additive Gaussian noise equals the luminance difference of the square against the background (so-called 100% noise). Fig. 3.8 shows that initial complex cell responses are strongly disrupted resulting from the additive noise pattern. However, recurrent feedforward-feedback iterations lead to a significant reduction of noise responses and at the same time to a strengthening of contour and contour-termination signals corresponding to V1 bipole and V1 end-stop cell activities, respectively. Note that the long-range interaction stage does not lead to activities beyond contour terminations at the corners of the square which would mistakenly lead to turn

L-junctions into X-junction.

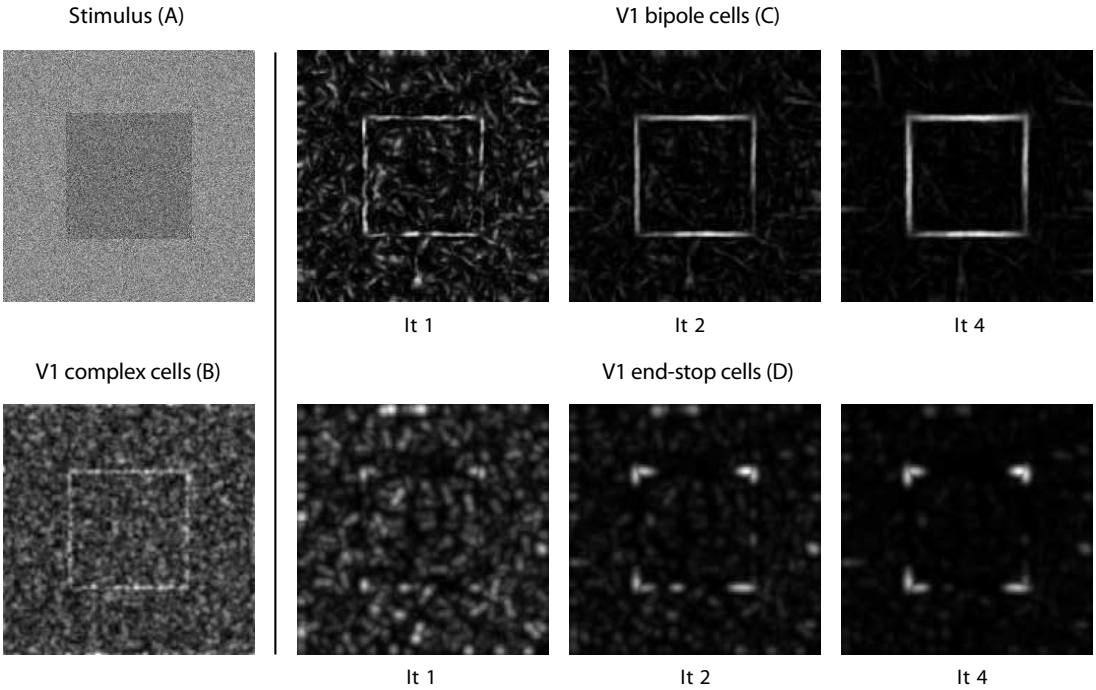


Figure 3.8: Robustness to noise. A square stimulus that has been corrupted with high amplitude Gaussian noise was used as model input (A). Initial complex cells responses are strongly influenced by the noise pattern (B). Recurrent feedforward-feedback processing significantly reduce activity of noise-induced responses of V1 bipole cells (C) and end-stop cells (D) (illustrated are responses after 1, 2, and 4 iterations). At the same time, surface contours and corners are enhanced over time.

### 3.3.2 Extraction of junction configurations

In a second simulation, we used a noise-free image of four occluding transparent and opaque squares as input for the model (Fig. 3.9a). The stimulus includes all three types of junction configurations and is therefore a good example to demonstrate the capabilities of the model at a glance. Initial feedforward responses from simple and complex cells shown in Fig. 3.9b and Fig. 3.9c demonstrate that these responses are not invariant to luminance contrast. Furthermore, they are not robust against noise (Fig. 3.8).

The results described in the following correspond to model activities after four recurrent feedforward-feedback cycles where model activities tended to converge to a stable state. The upper sub-area of model V1 is represented by bipole cells and end-stop cells. End-stop cells respond at positions where contours meet (e.g. T-, and L-junctions) or at positions where a contour ends (Fig. 3.4). However, they do not respond at X-junction configurations as can be observed in Fig 3.9e. Model V1 bipole cells are responsive for contours. They connect short like-oriented fragments and equalize contrast changes along the contour. At contour endings their activity is reduced (not shown) since they have additively connected sub-field (one sub-field is still activated, see Fig. 3.4). The additive

combination of V1 bipole and perpendicular oriented end-stop cells compensates for the reduction effect at corners or T-junctions. Fig. 3.9d show that contour activity is not reduced at corners or T-junctions. The combined bipole end end-stop activities from model area V1 are further processed by V2 bipole cells which are as well responsive for contours. However, as a result of their multiplicatively connected sub-field these cells do not respond at contour endings. Consequently, V2 cell responses are zero at T- and L-junctions, but not at X-junctions (see Fig. 3.9d and Fig. 3.4).

In Fig. 3.10, T-,L-, and X-junction maps were extracted and visualized based on converged model activities. In each map, the presence of the specific feature is signaled by patches of high activity. In order to prevent multiple features to be active at the same position, all signals undergo a competition stage where multiple signals in a local neighborhood compete with each other. The output of this stage is presented in Fig. 3.10c which is a combined map that represents different features, signaled by color. Finally, from this map, position and type of junctions are extracted by local maximum selection (Fig. 3.10e).

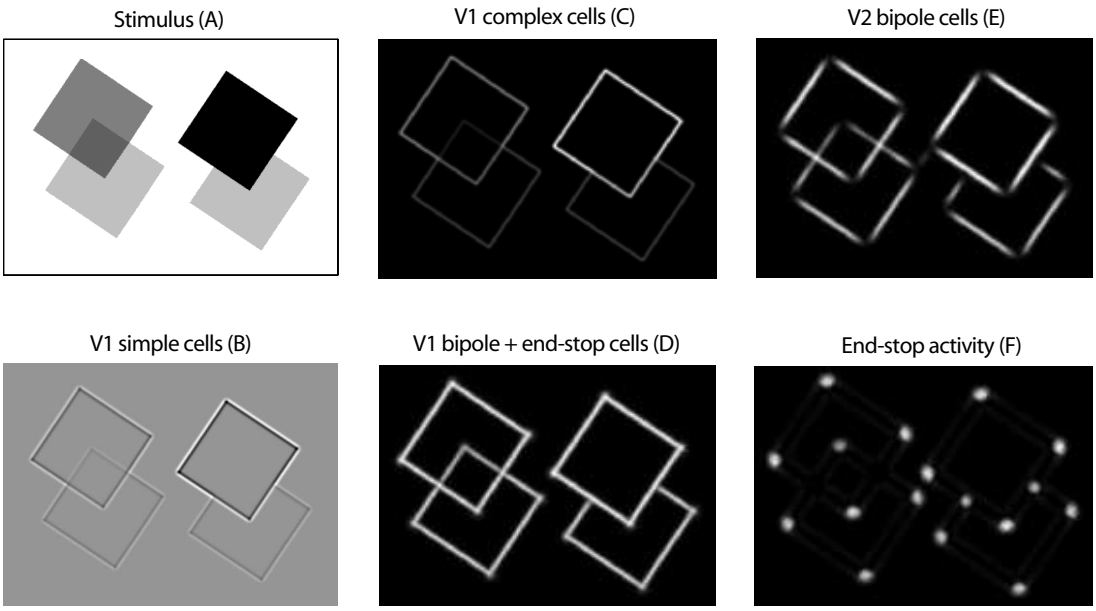


Figure 3.9: Model responses based on artificial input stimulus of overlapping squares. The stimulus (A) includes two different occlusion conditions: two overlapping opaque surfaces that produce T-junctions and two overlapping transparent surfaces which generate X-junctions. Furthermore, several L-junctions are visible at the corners of the squares. Model activities are summed over the orientation domain for all stages of our model. We show initial V1 simple and complex cell responses (B, C), combined V1 bipole/end-stop cell responses (D), end-stop activity (E), and V2 bipole cell activity (F) after 4 recurrent feedforward-feedback cycles. For clarification, the labels (A)-(F) correspond also to labels (A)-(F) in Fig. 3.3. Note that end-stop cells do not respond to X-junction configurations and that V2 bipole cells do not respond at L-junctions (cp. Table 1). Note also, that responses in (D)-(F) are invariant against image contrast.

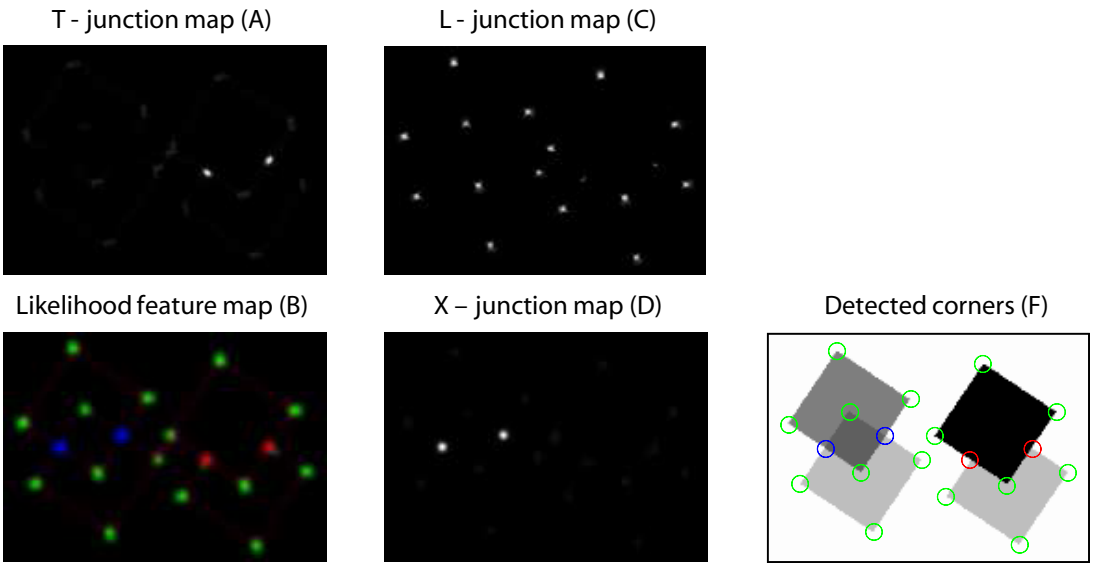


Figure 3.10: Saliency maps for different junction types. The junction maps (A), (B), and (D) were extracted from model activities presented in Fig. 3.9. Bright regions indicate positions where the ensemble of model activities suggests for the presence of the respective junction type (L-, T-, X-junction). A combined feature likelihood map(C) incorporates all features, color coded. Blue signals the presence of X-junctions, red signals T-junctions and green signals the presence of L-junctions. Moreover, position and type of detected features is visualized in (E) (based on maximum likelihood selection).

### 3.3.3 Processing of illusory contours

In a third simulation, we show the capabilities of the model to uncover illusory contours in scenes. As input we used two different versions of the Kanizsa square [Kanizsa, 1955](Fig. 3.5). The first image leads to the impression of an illusory square where the corners occlude black circles. The second image gives the impression of a white square in front of concentric circles. In both cases, only parts of the square are formed by luminance contrast. However, human observers mentally *see* the square as a coherent object. Our model results demonstrate that the invisible contour parts are uncovered by V2 bipole cell responses. Moreover, we show that subsequent recurrent feedforward-feedback cycles help to close large gaps between like-oriented contour elements or along linearly arrange contour endings (Fig. 3.11, Fig. 3.12). Importantly, this has a strong influence on the junction type that is signaled by the model responses. In Fig. 3.11, initial model responses suggest for L-junctions, which is in accordance with the physical luminance contrasts. But, after some recurrent iterations, V2 bipole cells close the gaps between contour elements which leads to a different, more global interpretation: the L-junctions along the illusory contour turn into T-junctions. The emergence of T-junctions in turn supports the perceptual interpretation of occlusions. A similar effect can be observed in Fig. 3.12 where line endings initially produce no junction signals. After a few feedforward-feedback cycles, however, the emerging illusory contour responses of V2 bipole cells lead to T-junction signals along the contour.

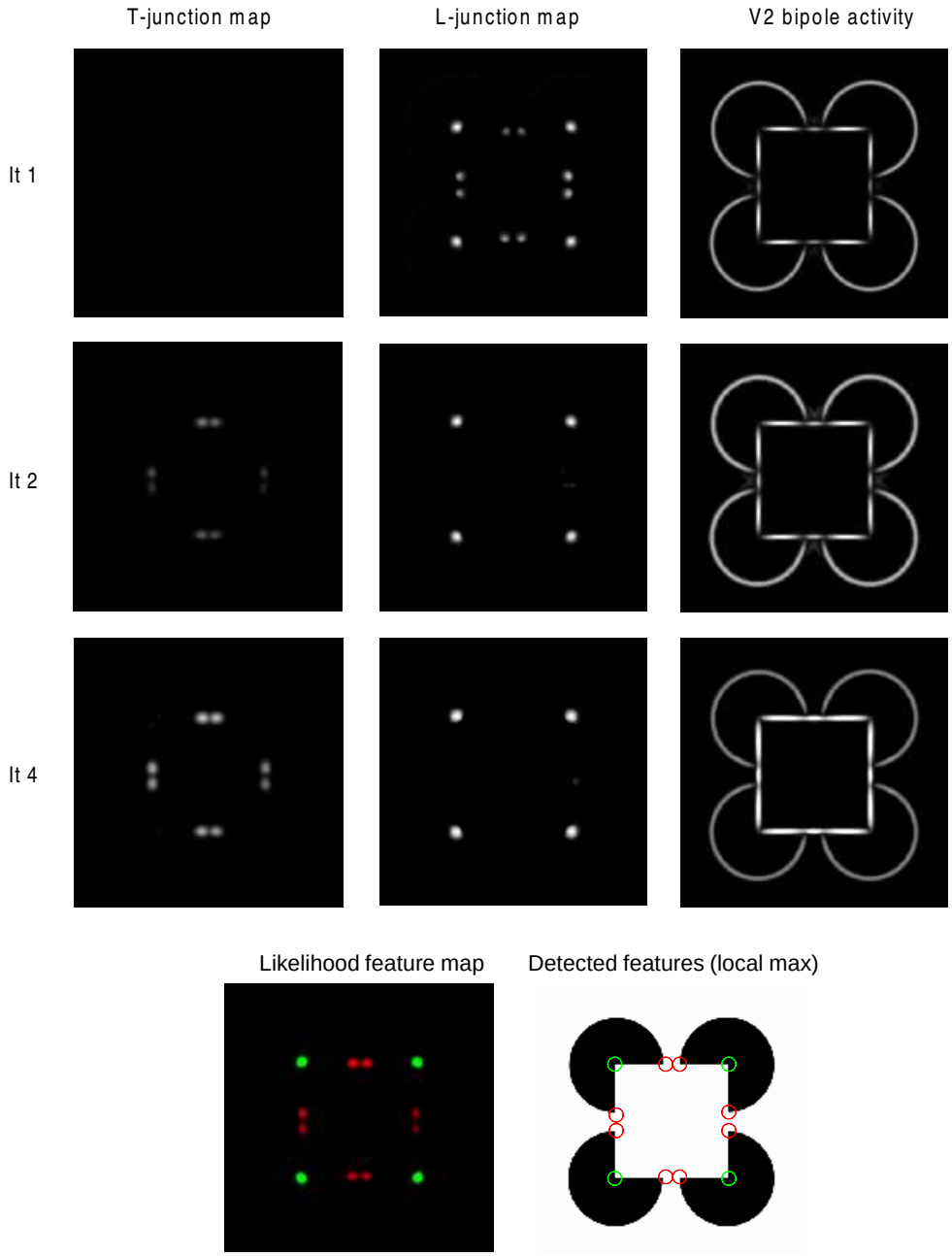


Figure 3.11: Recurrent processing of illusory contours. A Kanizsa figure is used as input for the model. V2 bipole cell activities and T-/L-junction signals are illustrated initially (no recurrent processing), after one recurrent cycle and after 3 recurrent cycles. The combined feature map show individual likelihoods co lour coded. Local maxima of the feature map are used to detect junction positions. Illusory contours are signal led by V2 bipole cell activities and are completed over time. Note, that as a result of the completion process of illusory contours, L-junctions signals turn into T-junction signals over time.

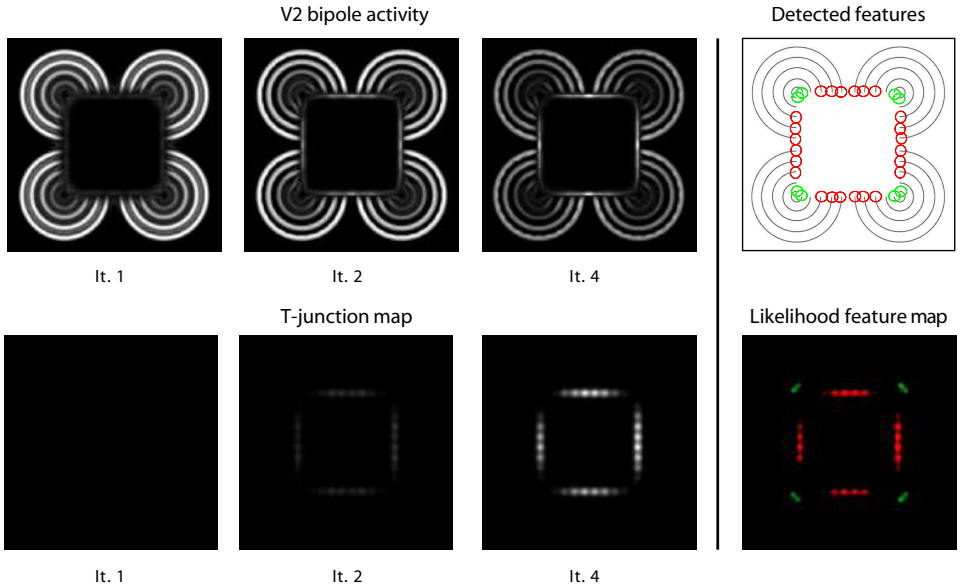


Figure 3.12: Recurrent processing of illusory contours formed by line ends. An alternative version of the Kanizsa figure is used where line ends lead to the impression of an illusory white square. Model responses are visualized according to Fig. 3.8. Note, that illusory contours signal led by V2 bipole responses develop over time. Note also, that over time this leads to the development of T-junction signals (red) which suggest for the presence of an occluding surface.

### 3.3.4 Processing of real-world data

In order to examine how the model performs for real-world camera images we used an image taken from a desk scene where several papers and a transparent foil are arranged such that they partly occlude each other (Fig. 3.13). Model activities and the extracted feature map demonstrate that the model is also capable of dealing with real-world images. Note, that one of the papers has a very low contrast ratio with respect to the background. Nevertheless, the model performs excellent in finding the contour and the respective junctions. This also underlines that the model is invariant to contrast changes and thus also stable against changes of illumination conditions.

### 3.3.5 Quantitative evaluation and comparison

In this section, we evaluate our model by comparing our recurrent junction detection scheme with results obtained by simply switching the recurrent feedback cycle off, which reduces the model to an ordinary feedforward model. Moreover, we compare our results to a standard computer vision corner detection scheme based on the Harris corner detector [Harris and Stephens, 1988]. In a comparative study of different corner detection schemes, the Harris corner detector provides the best results among five corner detectors [Schmid *et al.*, 2000]. As input for our comparison, we use corner test image adapted from Smith and Brady [Smith and Brady, 1997] that poses several challenges such as, e.g., low contrast regions, smooth luminance gradients, or obtuse and acute angles. Moreover, all types of junctions (L, T and X) considered are represented in the test image together with

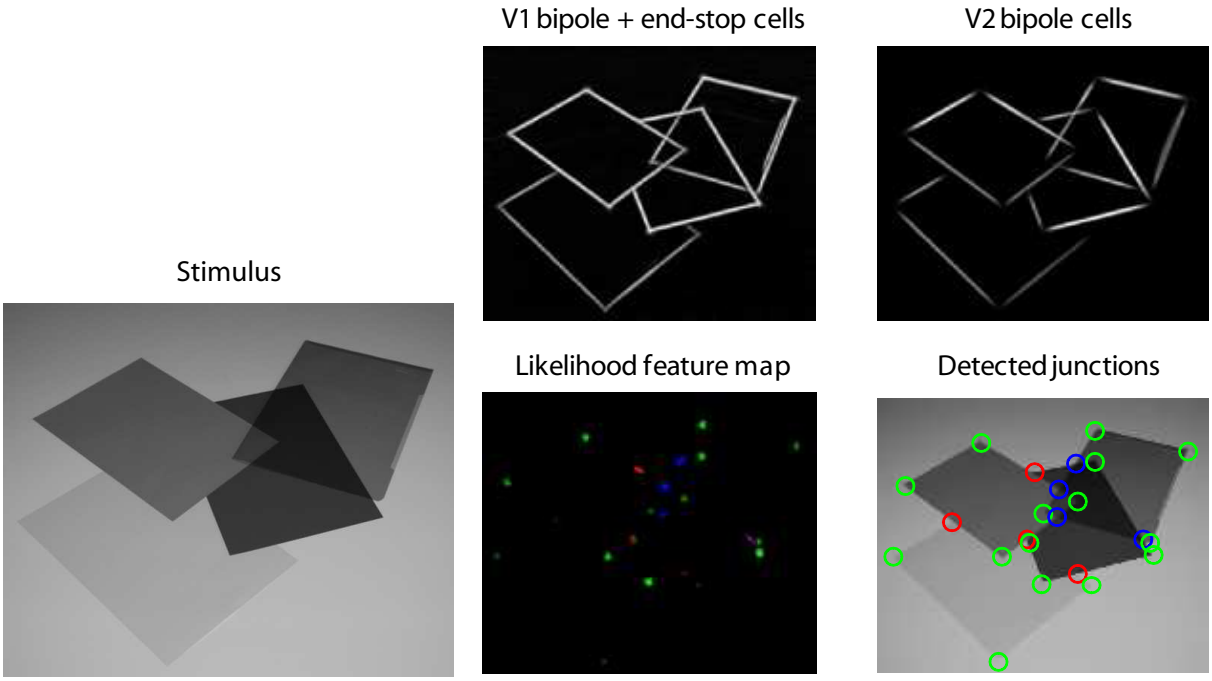


Figure 3.13: Processing of real-world input image. Model activities of V1 and V2 cell pool (right, first row) resulting from a real-world scene image that includes occluding opaque and transparent surfaces (left). A feature map was extracted from model activities, revealing position and type (color coded) of detected junction configurations (right, second row).

information about their exact position (ground truth information). Since the Harris corner detector is not able to discriminate between different junction types, our comparison is only based on the detection performance, irrespective of the junction category. To measure the performance of the different schemes we use receiver operator characteristic (ROC) curves. This method is frequently used to evaluate true positive rate or hit rate and the false positive rate of a binary classifier system as its discrimination threshold is varied. Here, we use the junction feature map as input for the ROC analysis. Fig. 3.14 shows the resulting ROC curves extracted from junction feature map given the test image as input. It is clearly visible that the ROC curve computed from the recurrent model responses lies well above the Harris corner detector curve and the initial feedforward model curve. This suggest for a significantly better detection performance of our recurrent model compared to feedforward processing-based junction detection schemes.

### 3.3.6 Simulations with dynamic input stimuli

We particularly designed this experiment to demonstrate a model prediction, namely, that feedback leads to a brief persistence of object and material appearance. This is usually unnoticed when the scene does not change at all. If, however, appearances of surface material change during recurrent interaction (while keeping the shape registered) the apparent surface property should stay more prolonged depending on whether it is transparent and changes into opaque or whether it is opaque and smoothly changes into

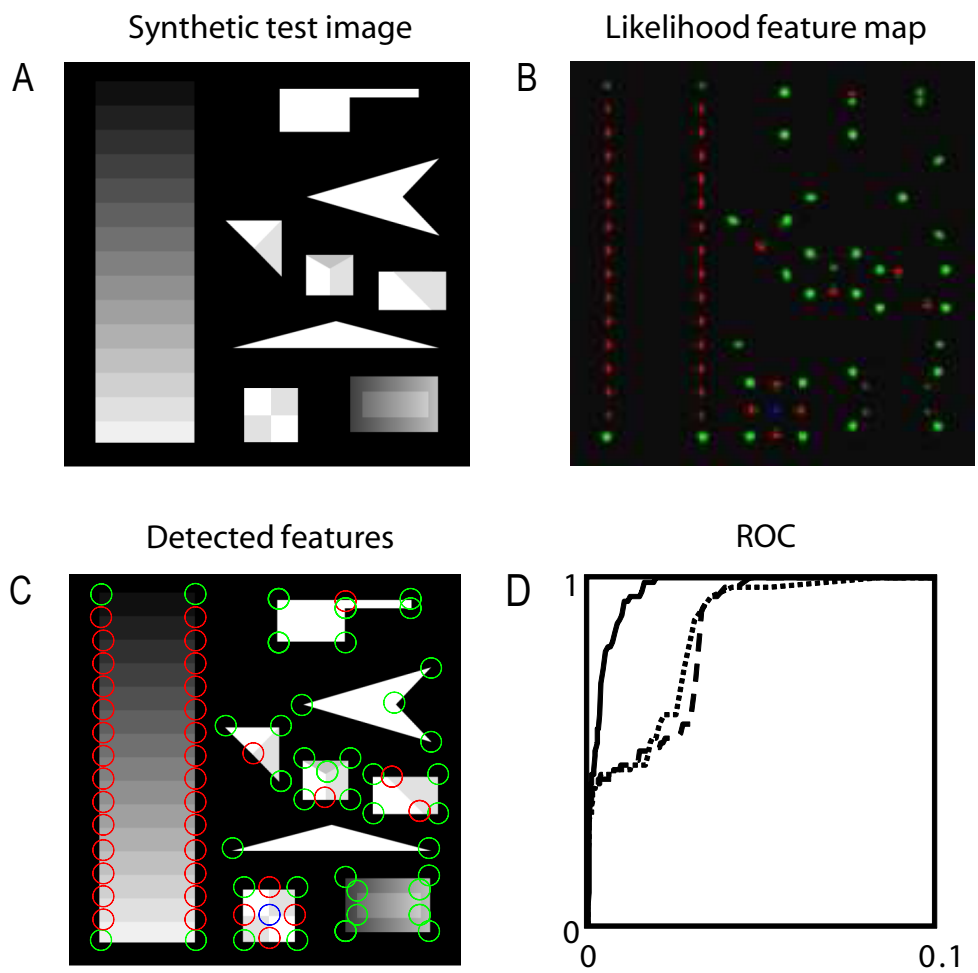


Figure 3.14: Evaluation of extracted junction signals on synthetic test image. A synthetic test image (A) reproduced from Smith and Brady (1997) was used to evaluate and compare extracted junction signals against a computational scheme (structure tensor) for corner detection proposed by Harris and Stephens (1988). The extracted junction saliency is visualized in a feature likelihood map (B) and detected junction positions/types are superimposed on input image (C). ROC curves are computed from structure tensor results (dashed), initial model responses (dotted) and from converged model responses after 4 recurrent cycles (solid) (D). Abscissa denotes the false alarm rate, and the ordinate denotes the hit rate. Note, that for better visibility the abscissa has been scaled to [0, 0.1].

transparent. Due to the action of modulatory feedback activation the registered boundary and junction activations continue to enhance the configuration signaled from previous input features for a short period of time. This appearance history (or memory in the processing architecture) is known as hysteresis effect. Here, we used temporal sequences of two occluding squares where the opacity of the topmost square was linearly altered between 100% opacity and 90% opacity, thus making the occluded region increasingly more visible (invisible). The input sequence consists of 10 frames static input (opaque) followed by 10 frames linear change from 100 % to 90% opacity followed by 10 frames static input (transparent, 90% opacity). The sequence was presented as described above (opaque-transparent) and in reverse temporal order (transparent-opaque). To investigate the hysteresis effect of feedback we presented both sequences to the full model (with feedback connection enabled) and to a restricted version of the model with feedback connections disabled. In the full model, feedback processing time is equal to stimulus presentation time, i.e., one feedback cycle is performed per stimulus time frame. In the restricted model, we switch all feedback connection off which constrains the model to perform only feedforward processing.

Throughout the simulation, model activities indicating T- and X- junctions are extracted at positions where occluding contours of both squares intersect each other (Fig. 3.15). The results generated by the full model show that feedback leads to a sequence directional hysteresis effect by temporally locking the prediction of a junction type (T- or X-junction). Moreover, initial ambiguities induced by predictions for different junction types are resolved by top-down feedback during the first few iterations. In contrast, when the model is restricted to feedforward processing no hysteresis effect can be observed, i.e., the model activities are equal for both input sequences, namely opaque-transparent and transparent-opaque transitions. Furthermore, no disambiguation between different junction type predictions takes place.

## 3.4 Discussion

In this section we begin by summarizing our main findings. Then, we compare our model with other proposed models that are related to our work. Moreover, we show that all core mechanisms employed in our model are biological plausible. We also discuss how junction signals that are extracted from our neural representation could be used by other cortical areas to solve visual tasks such as depth ordering, figure-ground segregation or motion correspondence finding. Finally, we briefly discuss some examples where our model fails to produce accurate results.

### 3.4.1 Summary of findings

We presented a recurrent model of V1-V2 contour processing utilizing long-range interactions in combination with short-range lateral inhibition which as been adapted from [Neumann and Sepp, 1999]. We have shown quantitatively and qualitatively that recur-

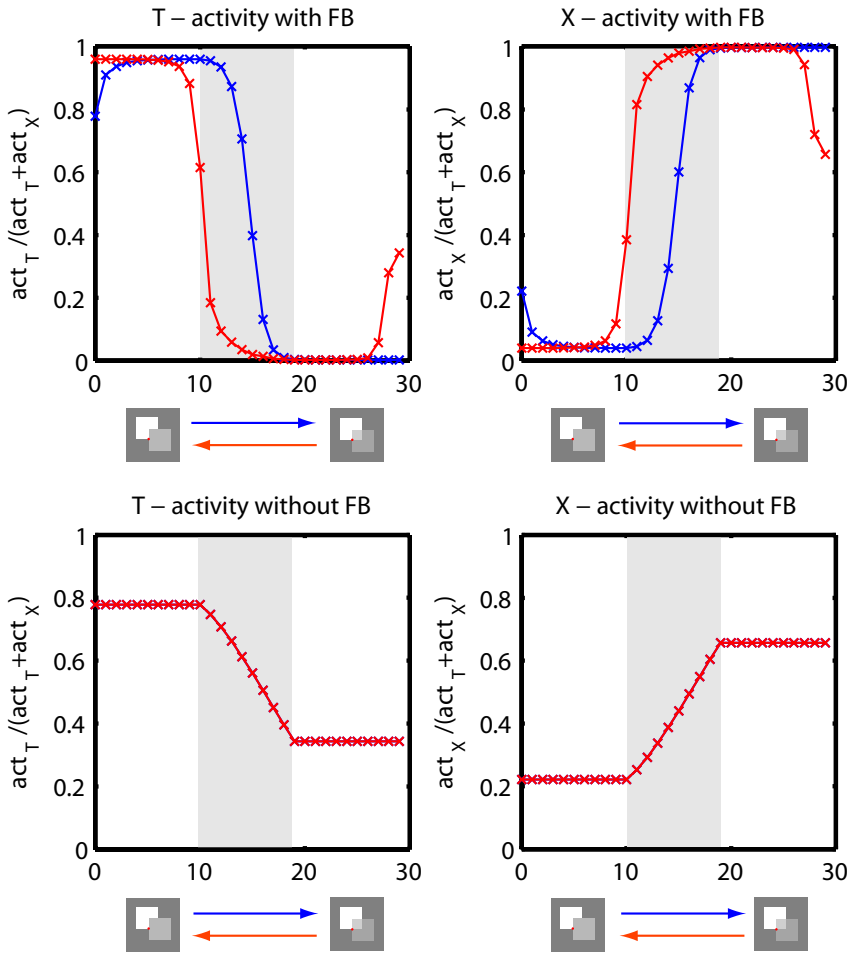


Figure 3.15: Hysteresis effect induced by feedback. The figure illustrates T- and X-junction activities extracted from the model based on temporal input sequences of two occluding squares where the topmost square changes material appearance from opaque (100% opacity) to transparent (90% opacity) (blue lines) over time. Furthermore the sequence was presented in reverse temporal order leading to a change of material appearance back from transparent to opaque (red lines). Activities were extracted at positions indicated by the red dot on the stimulus. Illustrated are results based on model simulations with feedback (top row) and without feedback connections (bottom row). Gray shaded areas indicate periods where stimulus properties linearly change. The results demonstrate that feedback leads to a hysteresis effect by temporary locking the prediction for a junction type (T- or X-junction). Without feedback the hysteresis effect disappears and both input sequences produce exactly the same results (only red curve is visible). Furthermore, the results demonstrate that feedback helps to disambiguate initial junction signals by amplifying the most likely prediction and suppressing weak predictions over the first few iterations.

rent combination of contextual features substantially enhances the initial estimates of local contrast. We have also shown that the model V2 cells are capable of generating illusory contour groupings which strongly influence the interpretation of different junction type estimates of local contrast. In addition, we demonstrated that cell activities represented in both model areas can be combined and extracted to robustly signal three different sorts of junctions (L-,T-, and X-junctions). These junctions provide important cues for fundamental visual processes such as surface completion and figure-ground segregation [Rubin, 2001b,a; Koffka, 1935]. Furthermore, our model predicts a hysteresis effect between opaque-transparent and transparent-opaque transitions which could be experimentally validated in a psychophysical experiment. Finally, we have demonstrated in a quantitative analysis that our model responses outperform a state-of-the-art computer vision corner detection scheme.

### 3.4.2 Related work

A number of different models have been proposed for contour integration. A comprehensive review can be found in [Shipley and Kellman, 1990]. Our contour integration model which utilizes interaction of feedforward and feedback, in particular modulatory feedback, has been applied successfully to a number of different tasks of visual processing such as optical flow estimation [Bayerl and Neumann, 2004], texture processing [Thielscher and Neumann, 2005], selective attention [Hamker, 2005], cortico-thalamic enhancement [Gove *et al.*, 1995], and linking synchronization [Eckhorn *et al.*, 1990].

One of the first computational models in the context of contour grouping that incorporate principles of long-range interactions and *interlaminar* recurrent processing has been proposed by Grossberg and colleagues [Grossberg and Mingolla, 1985] by introducing the Boundary Contour System (BCS). A more recent version of the BCS focuses more on the *intercortical* processing between areas V1 and V2 [Ross, 2000]. Grossberg and coworkers propose that V2 is mainly a slightly modified version of V1 operating at a coarser scale. Thus, they suggest that both areas, V1 and V2 share the same functional properties. Unlike them, we argue that V1 and V2 have different functional roles, e.g., corner selective cells occur in V1, bipole cells responding to illusory contours occur in V2. A more fundamental difference in the model of Grossberg and colleagues is that they use additive feedback connections with the effect that new activity can be generated in model area V1 at positions where initial bottom-up signals are zero. To compensate for this, they have to incorporate thresholds which lead to more complex balancing processes.

However, we use modulatory feedback connections, which implies that initial bottom-up activity is *required* to generate activity. Thus, in our model, illusory contours characterized by zero luminance contrast can only be signaled in V2, but not in V1. This is consistent with electrophysical studies of Peterhans and von der Heydt [Peterhans and Heydt, 1989] who concluded that illusory contour cells are virtually absent in V1. Concurrently, there is strong evidence for illusory contour selective cells in V2 [von der Heydt *et al.*, 1984]. A recurrent model of V1-V2 interactions based on modulatory feedback was

first proposed by Neumann and Sepp [Neumann and Sepp, 1999]. Mundhenk and Itti [Mundhenk and Itti, 2003] presented a multi-scale model for contour integration that is motivated by mechanisms in early visual cortex (V1). Similar to our model, the authors try to extract saliency values from contour representations. Individual contour saliency maps from different scales are combined by weighted averaging. Differences to our model are that they do no incorporate feedback mechanisms and that they do not consider illusory contour extraction as they do not model V2 neurons.

Interestingly, only few computational models of contour grouping address the computation and representation of corners and junctions. The model proposed by Heitger et al. [Heitger *et al.*, 1998] is closely related to our model because they use several elements that we incorporated into our model (e.g. complex cells, end-stop and bipole operators). A key element of their model is the concept of *ortho*- and *para*-grouping to generate illusory contour representations. *Ortho* grouping applies to terminations of the background, which tend to be orthogonal to the occluding contour. *Para* grouping applies to discontinuities of the foreground and is used to interpolate the contour in the direction of termination. However, a major shortcoming of their model is that it relies on a purely feedforward scheme which would presumably produce erroneous results when given a degraded input image (e.g. by noisy or low contrast). This also contrasts with the findings of several authors [Rubin, 2001b; Hupé *et al.*, 1998; Sillito *et al.*, 1995; Zhou *et al.*, 2000; Baylis and Driver, 2001] that feedback and recurrent interactions play an important role in visual processing for figure-ground segregation.

The model of Hansen and Neumann [Hansen and Neumann, 2004] is also closely related to our model since it is based on the same biological principles such as modulatory feedback and long-range interactions for the extraction of corners and junctions. However, the model is restricted to interlaminar interactions in V1 to explain contrast detection and subsequent enhancement effects. Our model, on the other hand, incorporates several extensions. First, our model takes illusory contours into account by additionally modeling area V2. Second, we show that our neuronal representation is further processed to extract different junction types such as L, T- and X-junctions which are perceptually important features that provide basic cues for global scene interpretations.

Recently, a recurrent model for surface-based depth processing was proposed by Thielscher and Neumann [Thielscher and Neumann, 2008]. In their proposed model, depth information derived from monocular cues is propagated along surface contours using local recurrent interactions to obtain a globally consistent depth sorting of overlapping surfaces. The model differs in several aspects from our model. In contrast to our model, they use additional recurrent interactions in V2 to propagate border-ownership information derived from detected T-junctions along contours. This propagated information enables them to obtain a globally consistent interpretation of depth relations between surfaces. Unlike this approach to monocular depth segregation, we focused on the extraction and perceptual interpretation of junction configurations.

In summary, variations of mechanisms employed in our model can be found in several other models of visual processing. But only few of them have concerned for combined

boundary extraction and junction extraction as well as their distinction. Unlike previous proposals which treat localized junction configurations as 2D image features [Harris and Stephens, 1988; Smith and Brady, 1997], we link them to mechanisms of apparent surface segregation. As a consequence, we demonstrate how junctions can change their perceptual representation depending on the scene context and the spatial configuration of boundary fragments.

### 3.4.3 Biological plausibility of model components

Our model architecture is inspired by biological mechanisms and is based on neural representations of early visual cortex. We now put individual model components into a physiological or psychophysical context and discuss for their plausibility.

In the initial stages of our model we simulate V1 simple and complex cells [Hubel and Wiesel, 1968]. Model V1 bipole cells are inspired by horizontal long-range connections that link patches of neurons of similar orientation preference [Gilbert and Wiesel, 1989; Bosking *et al.*, 1997]. Consistently, model V1 bipole cells pool activities of appropriately aligned complex cells from the lower part of model area V1 (Fig. 3.3) which resembles intracortical layer 4 of area V1. Evidence for non-local integration also comes from psychophysical experiments for contrast detection [Kapadia *et al.*, 1995] and contour integration [Field *et al.*, 1993]. Additionally, we model end-stop cells that selectively respond to contour terminations. The existence of V1 neurons reacting to end-stop configurations has been confirmed by several electrophysiological studies [Hubel and Wiesel, 1965; Maffai and Fiorentini, 1976; Peterhans, 1997]. As a consequence, end-stop cells were also modeled by several authors in the context of contour integration [Thielscher and Neumann, 2008; Peterhans and Heydt, 1989; Lesher and Mingolla, 1993].

In model area V2, we employ modified bipole cells with nonlinear response properties. As V2 neurons have larger receptive fields than V1 neurons, our bipole filters employed in model area V2 have a larger extent that those used in the upper part of model area V1 (Fig. 3.3). Evidence for contour selective cells in V2 comes from von der Heydt [von der Heydt *et al.*, 1984] where the authors probed V2 neurons with illusory-bar stimuli. They selectively respond to coherent arrangements having both fragments of an illusory bar intact. If one fragment is missing, the cell response drops to spontaneous activity [Peterhans and Heydt, 1989]. To be consistent with these findings we modeled V2 bipole cells with multiplicatively connected sub-fields which leads to similar effects than those reported by von der Heydt.

### 3.4.4 Evidence for representation of junctions and corners in visual cortex

Although it seems obvious that junctions play a crucial role in several perceptual processes [Rubin, 2001b] only little evidence was found that specific cells in the visual cortex are particularly responsive to junction features. Several studies suggest the presence of a

neural organization in V1 that may represent a mechanism for detecting local orientation discontinuity [Kapadia *et al.*, 1999; Knierim and van Essen, 1992; Sillito *et al.*, 1995]. Their results indicate a facilitatory interaction between elements of V1 circuitry representing markedly different orientations in contradiction to the common believe that functional connectivity is only seen between cells of like orientation [Ts'o *et al.*, 1986]. However, it is still unclear to what extent this selectivity is used for junction processing. In a study by Kobatake and Tanaka [Kobatake and Tanaka, 1994] critical features for the activation of cells reaching from V2, V4 up to posterior inferotemporal cortex (IT) were determined. V2 cells were found to react to stimuli such as concentric rings or tapered bars. Cells that respond selectively to junction-like features like crosses were only found in V4 and posterior IT.

More recently, two studies [Ito and Komatsu, 2004; Anzai *et al.*, 2007] report on cells in monkey visual area V2 that seem to explicitly encode combinations of orientations as represented by junctions or corners. Thus, such V2 neurons may provide important underpinnings for the analysis of surfaces [Nakayama *et al.*, 1995]. In a straightforward model approach it was shown that these V2 neurons may simply sum the responses from orientation selective V1 neurons [Boynton and Hedgé, 2004]. However, the fact that only little evidence exists for junction selective cells in V2 could also motivate the hypothesis that junctions are not explicitly encoded by specific cells in V2 but higher visual areas such as V4 or IT link responses from cell types selective for lower-level features, such as complex, end-stop, and bipole cells. Thus, the extraction of junction signals from combinations of model cell responses, described as *read-out process* in our model, follows the idea mentioned above, that, e.g., V4 cells could pool signals from several cortical areas, particularly from V1 and V2. Notably, we do not claim that junction signals are encoded by V4 or IT neurons, but we demonstrate that our model performs well assuming that junctions are processed from distributed activities of neurons at early cortical stages.

### Model predictions for psychophysical experiments

Our model incorporates recurrent feedback processing from higher to lower stages. This leads to temporal model dynamics depending on bottom-up feedforward signal and top-down feedback signal. Without a change of the input signal (e.g., static input) model activities tend to converge after a few iterations. However, when the input signal temporally changes this leads to a conflict between bottom-up and top-down signals. Thus, the system acts like a short-term memory maintaining the actual state for a few time steps. Consistently, if the material appearance changes from opaque to transparent over time one would expect that the perception of the apparent stimulus is more prolonged in time. From this it follows that our model simulations predict a perceptual hysteresis effect for discrimination between opaque-transparent and transparent-opaque transitions induced by a top-down feedback mechanism. Such a hysteresis effects has been already observed psychophysically for motion direction disambiguation (leftward motion vs. rightward motion) [Williams and Phillips, 1987]. Since both motion and form processing are based on the same neural principles we expect that the predicted hysteresis effect can

also be measured psychophysically. Therefore, we are currently planning to investigate experimentally whether our model predictions are validated or not.

### 3.4.5 The role of junctions in visual perception

Our core model integrates like-oriented contrasts to simulate the process of contour perception in the visual system. When contours of different orientation meet at the same place non-collinear orientation combinations, namely junctions, are formed. The formation of junctions provides important cues, e.g., for the occurrence of occlusions and transparencies. Occlusions occur in almost every real word scene, and thus, surface completion is a fundamental visual process. In the following, we discuss the individual role of some basic junction types that can be extracted by our model.

#### Transparency

It has been suggested that the perception of transparency is triggered by X-junctions formed by junctions of contours of the transparent and opaque regions at the overlapping area [Kersten, 1991]. However, the presence of X-junctions is necessary but not sufficient to elicit a strong transparency effect. In addition, the luminance contrast around the X-junction must follow the two rules: (1) the direction of luminance contrast across an opaque border cannot change in the transparent region; (2) the luminance difference across an opaque border must be reduced in the transparent region [Metelli, 1974; Anderson, 1997]. A violation of these rules strongly diminishes the perception of transparency.

Wolfe and collaborators [Wolfe *et al.*, 2005] explored in a series of visual search experiments which cues are relevant to guide attention in a search for opaque targets among transparent detractors or vice versa. One of the experiments showed that performance is impaired when X-junctions are removed from transparent items. Another experiment showed that efficient search is still possible if X-junctions are merely occluded (i.e. an occluding bar is used that disrupts the X configurations). In summary, these findings show that indeed X-junctions play an important role in the perception of transparencies, but there seem to be many other factors that play an additional role for transparency perception. Nevertheless, our proposed architecture facilitates the perceptual interpretation of X-junction as proposed by Wolfe and colleagues.

#### Occlusion

When an opaque surface occludes another surface of different luminance a T-junction is formed at the position where the boundary contours intersect each other. If the surface in front has the same luminance than the background the T-junctions collapses to an L-junction. We have shown that our model initially detects such configurations as L-junctions. After a short period of time, when contours are completed over gaps, such L-junctions are recognized by the model as T-junctions. This is consistent with the more context driven interpretation, as observed by Rubin [Rubin, 2001b]. Rubin investigated in

psychophysical experiments how local occlusion cues, such as T-junctions and more global occlusion cues, specifically relatability and surface similarity, play a role in the emergence of amodal surface completion and illusory contour perception. Two contour fragments are *relatable* when they can be connected with a smooth contour without inflection points [Wolfe *et al.*, 2005]. Rubin proposes that local T-junction cues can initiate completion processes and that relatability plays a part at later stages.

Interestingly, in rare cases, T-junctions can also support the perception of X-junctions [Watanabe and Cavanagh, 1993]. In their psychophysical experiments, the T-junctions were perceived as having an additional illusory contour leading to the perception of an X-junction (termed "implicit X-junction"). This special case shows that T-junctions do not always lead to the perception of occluding opaque surfaces but can itself be altered in the more global context when prototypical surface patches are formed which may lead to a reinterpretation of local features.

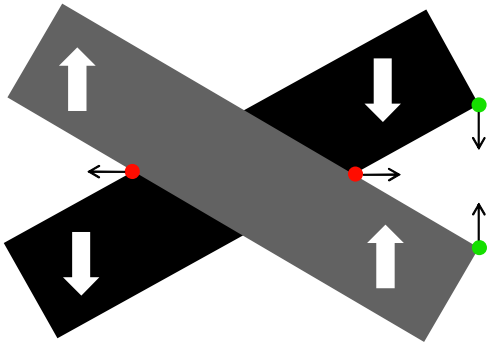

Figure 3.16: Dynamic stimulus where two occluding bars move in opposite direction. Tracking of L-junctions (green) leads to correct motion estimates of the two bars while tracking of T-junctions leads to erroneous motion estimates. Thus, the visual system might use form information, e.g., surface-based occlusion cues to selectively discount local motion estimates for moving T-junctions.

### Figure-ground segregation

Separating figure from background is one of the most important tasks in vision. Figure and ground information in an image can be represented by assigning ownership of the border between two surfaces. The figure which occludes parts of the background leads to specific boundary configurations, in particular T-junction configurations which can help in the assignment of figure and background. In more detail, the stem of the "T" is formed by the boundary contour of the background surface while the top of the "T" corresponds to the boundary contour of the figure. Motivated by physiological evidence for cells that are selective for border-ownership information [Zhou *et al.*, 2000] some models were proposed where cues signaled by T-junctions are used to generate consistent representations of layered surfaces [Thielscher and Neumann, 2008; Zhaoping, 2005]. This underlines the importance of T-junction cues for the visual system.

### Motion perception

Junctions do not only play a role in static scenes, they are also important in the context of motion perception. An object's motion cannot be determined from a single local measurement on its contour which is commonly known as the aperture problem [Wallach, 1935]. However, at positions where multiple oriented contrasts (i.e. two-dimensional features, such as corners and junctions) are present the ambiguity can be resolved and further be propagated along object contours to get a more global motion percept [Pack and Born, 2001]. Thus, tracking of two-dimensional features over time is a fundamental task in the analysis of motion signals.

In a study by Pack et al. [Pack *et al.*, 2003] it is suggested that end-stopped V1 neurons could provide local measures of two-dimensional feature correspondences in motion by responding preferentially to moving line endings. However, the results of Guo et al. [Guo *et al.*, 2006] contrast with the suggestion that end-stop neurons can determine global motion directions. They propose that lateral and feedback connections play a critical role in V1 motion information integration. But still, it remains unclear whether cortical neurons represent object motion by selectively responding to two-dimensional features such as junctions and corners. On the other hand, motion of specific junction configurations, in particular T- and X-junctions generates erroneous motion trajectories. An example is shown in Fig. 3.16, where T-junctions generated by two occluding bars that move in opposite directions lead to incorrect local motion estimates [Lorenceau and Shiffrar, 1992]. Thus, static form cues could be selectively discounted in the process of motion interpretation [Nowlan and Sejnowski, 1995].

### Limitations of the model

Although we have shown that our model is able to produce results that are in line with several empirical findings, there are also some shortcomings of the model. For instance, consider a Kanizsa figure such as illustrated in Fig. 3.17 where the gaps between contour elements are so large that the V2 bipole filter cannot bridge the gap. In this case the model would fail to produce an illusory contour signal in V2. Nevertheless, human observes still have a weak impression of seeing an illusory triangle. We suggest that higher visual areas such as V4 also play a role in illusory contour processing. Evidence comes, e.g., from Pasupathy [Pasupathy and Conner, 1999] who found responses to contour features in macaque area V4.

Another restriction of the model is that it does not account for different image scales. Consequently, the model focuses on fine details and suppresses coarse structures. However, the model could be provided with a pyramid of differently scaled version of the same input image. This would correspond to simply replicating the model at multiple scales. An alternative approach would be to employ scaled versions of Gabor wavelet filters in the input stage. Finally, our model does not explain how occlusion-based junctions can be distinguished from texture-based junctions. In this model, we only used stimuli that have homogeneous surface reflectance properties. Thus, contours are interpreted as

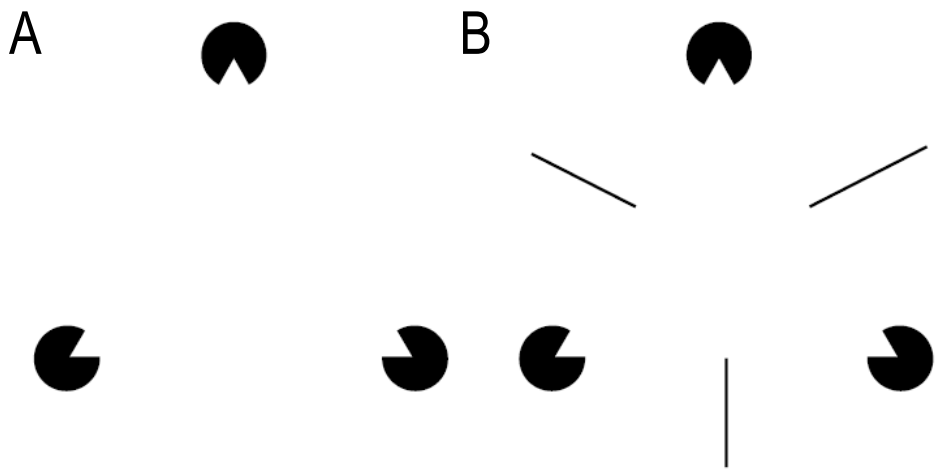


Figure 3.17: Stimuli where the model fails to produce illusory contour activations. Distances between the starting points of the contours are too large to be bridged by a V2 bipole cell.

surface borders by the model. In natural scenes, surfaces are often textured due to surface material properties which would also lead to junctions signals and thus to ambiguities in the interpretation. However, such ambiguities could be solved in higher visual area such as V4 [Fellemann and van Essen, 1991] which are not in the scope of this paper. In addition, stereo information can also help to correctly identify occlusion-based junctions.

# Chapter 4

# Learning of form and motion patterns in social interaction

## 4.1 Introduction

Understanding the behavior of other individuals is an essential ability that is crucial for survival of many species. For instance, the natural behavior is often used by higher primates to socially interact with conspecifics. In particular, humans have developed excellent skills in recognizing and interpreting the actions and intentions of others. It has been target of considerable research to investigate the neural mechanisms underlying our ability to understand others intentions from the mere observation of their motor actions.

Neurophysiological and neuroimaging studies have revealed that neurons in the superior temporal sulcus (STS) region of the cerebral cortex selectively respond to social actions that are characterized by movements of the eyes, face, hands and body [Allison *et al.*, 2000; Jellema and Perrett, 2007; Peelen and Downing, 2008]. Interestingly, in some situations even a single static image provides enough information to *read* someone's intentions or to recognize an action being performed (motion from form).

The painting in Fig. 4.1 demonstrates impressively that we can recognize several actions and intentions taking place without having explicit motion information. Thus, it seems obvious that mid- and high-level form information may as well play a crucial role in the perception and interpretation of social body actions.

By combining findings from physiology and brain imaging it can be demonstrated that the primate brain converts information about spatiotemporal sequences into meaningful actions through interactions between early and higher visual areas processing form and motion [Kourtzi *et al.*, 2008]. However, the precise interplay between form and motion signals in higher cortical regions of the primate brain still remains unclear.

Figure 4.1: Social interaction: The painting from Georges de La Tour 'The Fortune-Teller', (1630) give a quite strong impression what a single image can tell us about intentions and actions of other people. The man in the middle attends to the old woman while two other women attempt to rob him. Note that eye gaze, head pose and body posture seem to play a fundamental role for the correct interpretation of the scene. Figure adapted from [Allison *et al.*, 2000].

### Goals & Model approach

In recent years a considerable quantity of empirical evidence has been gathered in this particular research field, but only little effort has been made to develop computational models for validating existing data, and for testing the consistency of possible explanations.

In this chapter, we present a computational framework to simulate neural mechanisms at several cortical stages of the form and motion pathway that are supposed to be involved in the visual processing of socially relevant body actions. Moreover, we link information from form and motion pathways in a separate processing area by lateral interactions of from- and motion-driven sequence neurons. This contrasts with a trivial superposition (i.e., linear combination) of both signals. In our proposed model, we interpret artificially generated sequences of a character performing articulated body and head turns towards or away from an observer. Such articulated body actions in which at least one body part (e.g. the head) moves with respect to the remainder of the body typically communicate social signals, such as attentiveness or disinterest.

In the model form pathway static representations of the body, so called *snapshots*, are extracted while in the model motion pathway typical velocity patterns appearing from goal-directed head and body movements are captured. In both pathways, we employ a Hebbian learning principle to establish prototypical representations of form and motion appearances from sequences of body actions so that they capture different features of such body actions. Furthermore, in the next model stage we propose a neural mechanism how a combined signal is generated from model cells that are tuned to temporal sequences of form and motion activity patterns.

Our goal is to demonstrate that our proposed model performs significantly better in discriminating two socially relevant body actions when signals from both pathways are

integrated compared to the condition when input from one pathway is omitted. In this case, we expect to find a drop in discrimination performance. Also, the open question which pathway, form or motion, provides more information in this specific scenario is addressed in our model simulations.

## 4.2 Model

In this section, we describe in detail the individual stages of a biologically inspired model for the recognition of specific body actions that are important in the context of social interaction.

### 4.2.1 Biological motivation and model overview

Our proposed model for the processing of body actions is inspired by the cortical architecture of the primate brain where visual stimuli are processed hierarchically along two mainly dissociated pathways, namely the dorsal and the ventral pathway. Traditionally, a separate organization of the two pathways is proposed [Ungerleider and Mishkin, 1982; Desimone and Ungerleider, 1989; Goodale and Milner, 1992]. The dorsal '*where*' stream primarily deals with the analysis of spatial locations while the ventral '*what*' stream mainly deals with the shape and identity of objects. However, the strict 'where-what' dichotomy is questioned by several authors, e.g. [Goodale and Milner, 1992; Milner and Goodale, 1995], among others. In their view, the occipito-temporal ventral stream subserves the purpose of *perceptual* tasks (corresponding to form processing) while the dorsal occipito-parietal stream subserves the visual guidance of *action* (corresponding to motion processing).

Nevertheless, the extent of functional interaction between dorsal and ventral stream is still unknown. Though, there is evidence that interaction between both processing streams takes place specifically in higher cortical areas such as the superior temporal sulcus (STS) , in order to recognize the behavior of others in terms of key postures and animations [Oram and Perrett, 1996; Tanaka *et al.*, 1999].

In our proposed model both pathways are modeled at several cortical stages. In the ventral pathway we model the visual areas V1, V2, and inferotemporal cortex (IT). The dorsal pathway is represented by model areas V1, medial temporal cortex (MT/V5) and medial superior temporal cortex (MST). Finally, signals from both pathways interact in model area STS via lateral interactions. Signal flow between cortical areas is realized via feedforward and feedback connections between early visual areas (V1-V2 and V1-MT) and exclusively by feedforward between mid- and high-level cortical areas (V2-IT-STS and MT-MST-STS). We are aware that there is physiological evidence for feedback connections to also exist between higher cortical areas. However, for simplicity we restrict our model architecture to be purely feedforward at mid- and higher cortical levels. In model area STS we implement plastic connections between sequence neurons that are realized by Hebbian synapses. Given training sequences of a person performing an action these synapses can

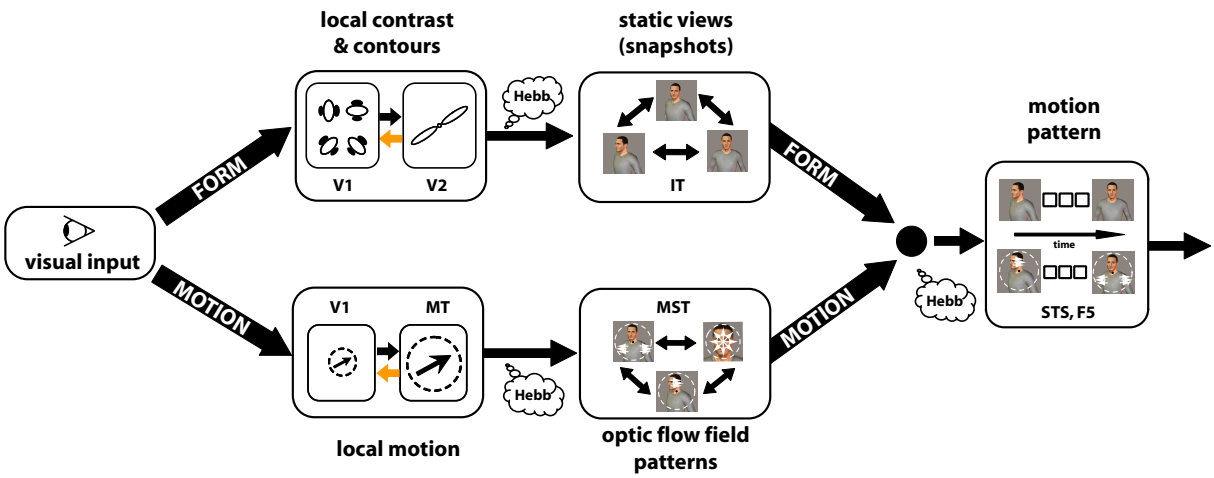


Figure 4.2: Overview of the proposed model architecture. Visual input is processed by two dissociated pathways. The form pathway is represented by model areas V1, V2, and IT while the motion pathway is represented by areas V1, MT, and MST. Both pathways project into model area STS where signals from both streams interact to generate a final output signal. Feedforward (black arrows) and feedback (orange arrows) are used to illustrate the signal flow direction. Furthermore, Hebbian synapses between cortical areas are indicated by a balloon symbol.

adapt to the stimulus such that sequence neurons become selective for specific form and motion patterns depicting certain actions.

### 4.2.2 Processing in the form pathway

#### Extraction of local form features

The initial stage of feature processing starts in model area V1 where oriented local contrasts such as small bars or edges are extracted. These local features are represented as activities for each position and feature type, i.e. contrast orientation, in the image. Further processing is performed in model area V2 where features are grouped into smoothly connected contours. These contours represent the outlines of objects as well as junctions that occur a positions where objects are partly occluded. Recurrent interactions (feedforward and feedback in turn) between model area V1 and V2 lead to an enhancement of features that lie along smooth contours of scene elements. It has been shown that this processing mechanisms leads to robust feature representations that are stabilized against varying illumination conditions or degradations by image noise [Weidenbacher and Neumann, 2009]. For a more thorough description of the proposed V1-V2 model architecture the reader is referred to chapter 3.

#### Learning of static views

The goal at this model stage is to learn prototypical representations of activity distributions that encode static views of persons that appear while performing a body action.

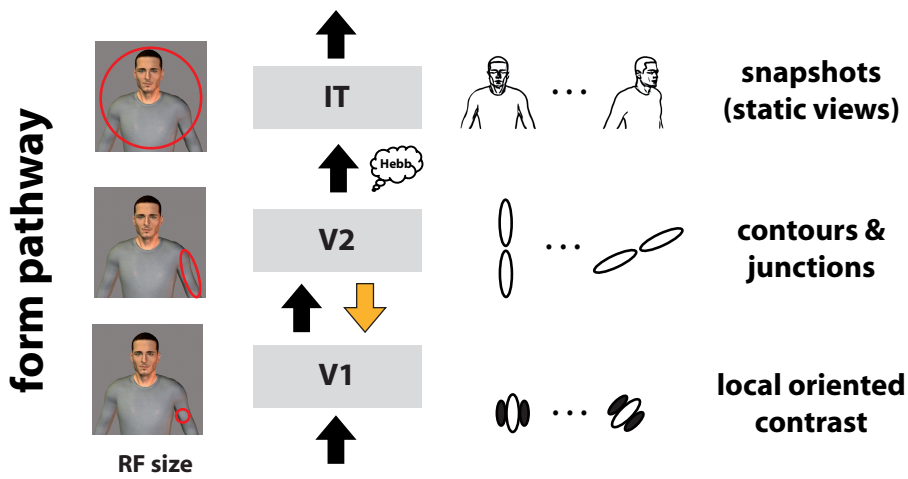


Figure 4.3: Overview of the model form pathway. Model area V1 starts with initial extraction of local oriented contrasts. In model area V2, like-oriented contrasts that are located for instance at object boundaries are grouped into contours. Feedforward (black arrows) and Feedback (orange arrow) processing between V1 and V2 helps to iteratively complete contours and to become invariant against variations in illumination. From this representation, IT neurons are trained by Hebbian learning to become selective for different body configurations that correspond to different snapshots of a person while performing an action.

Evidence for such view-tuned representations was found by Logothetis and colleagues [Logothetis *et al.*, 1995] in IT of monkeys.

Here, model V2 activities are forwarded and serve as input for learning synaptic connections between model V2 and IT neurons. For a given input sequence of body actions, activities $A^{V2}_{x,y,\varphi}$ from model area V2 are represented by local features $\varphi$ at each spatial image position $(x, y)$. Here, $\varphi_i$ is described by 8 different feature orientations equally distributed between 0 and 180 degrees. Note, that activity of model V2 neurons also depends on time as the input to the model is a temporal image sequence. However, at this stage, no temporal interactions are performed. Thus, we omit the index $t$ in the notation of activities for more clarity.

Pose specific representations are modeled by snapshot neurons that are fully connected to each feature neuron in model area V2. The activity of a snapshot neuron $A^{IT}_j$ is computed by a weighted sum over all positions and features:

$$A^{IT}_j = \sum_{x,y,\varphi} w_{(x,y,\varphi)j} \cdot A^{V2}_{x,y,\varphi}$$

where $w_{(x,y,\varphi)j}$ are the weights between model V2 neurons given space-feature selectivity (x,y,$\varphi$) and IT neuron $j$. Importantly, at this model stage, temporal correlations between frames of the sequence are not considered. However, we simulate lateral inhibition by applying a MAX-operation over all snapshot neurons [1], such that

[1]The MAX-operation can be implemented in various biologically plausible ways, e.g., by a feedforward network with shunting inhibition or by a recurrent feedback network. An analysis of biological plausible implementation of the MAX-operations can be found in [Yu *et al.*, 2002].

$$A_j^{IT} = \begin{cases} A_j^{IT} & if\ j = argmax\left(A_j^{IT}\right) \\ 0 & otherwise \end{cases} \tag{4.1}$$

There is neurophysiological evidence for MAX-like mechanisms in area V1 [Lampl *et al.*, 2004] and area IT [Sato, 1989]. Furthermore, it is postulated to be a fundamental operation involved at many stages of the primate visual processing system [Riesenhuber and Poggio, 1999].

In order to learn the synaptic connections represented by the weights $w_j$ we employ a Hebbian learning rule with a quadratic normalization term [Oja, 1982].

$$\Delta w_{(x,y,\varphi)j} = \eta \cdot \left(A_{x,y,\varphi}^{V2} \cdot A_j^{IT} - \left(A_j^{IT}\right)^2 \cdot w_{(x,y,\varphi)j}\right) \tag{4.2}$$

where $\eta$ is a small constant that represents the learning rate. It has been shown for this learning rule that the length of the weight vector is bounded and converges to a unit length of one [Oja, 1982].

Initially, the weights $w_j$ are set to equally distributed random values in range $[0, 1]$. Input activities $A_{x,y,\varphi}^{V2}$ are normalized, such that $\sum_{x,y,\varphi} \left(A_{x,y,\varphi}^{V2}\right)^2 = 1$. Hereby, we force each input activity distribution to have the same amount of energy and hence to have initially the same chance to win the competition in the MAX-operation (Eq. 4.3).

### 4.2.3 Processing in the motion pathway

#### Extraction of local motion features

The model architecture for motion processing is based on the organization of the dorsal stream, where areas V1, MT, MST are organized in a hierarchical way (see Fig. 4.2). The areas V1 and MT are connected bidrectionally such that signal flow is bidirectional, bottom-up and top-down. Connections between model areas MT and MST are modeled unidirectionally. Furthermore, these connections are modeled by plastic Hebb synapses such that selectivity of MST cells for short-term optic flow patterns can be learned.

Local motion feature extraction is performed by a model of V1-MT interaction that has been first proposed by [Bayerl and Neumann, 2004]. Here, initial motion detection is performed by spatiotemporal correlation of Gabor filter responses [Daugman, 1988] implemented by extended Reichhardt detectors [Reichardt, 1987]. These initial local motion estimates in model area V1 are proposed to generate local velocity space representations. Model areas MT and V1 are connected via feedforward (FF) and feedback (FB) connections. Recurrent FF and FB processing between model areas V1 and MT leads to enhanced motion features by solving, e.g., the aperture problem [Wallach, 1935; Pack and Born, 2001]. This is achieved by modulatory top-down feedback signals that deliver contextual motion information represented at model area MT. Activity in V1 that is consistent with the MT feedback signal is strengthened while inconsistent V1 activities are weakened by this process.

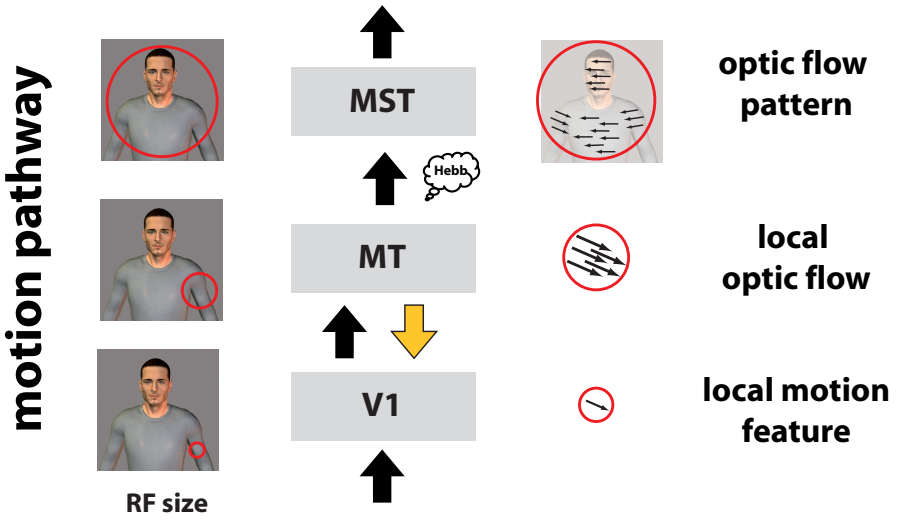


Figure 4.4: Overview of the model motion pathway. Model area V1 starts with the extraction of local motion features. In model MT, local optic flow is estimated by integrating activities that are forwarded from model V1. Recurrent feedforward-feedback processing between model areas V1 an MT helps to solve the aperture problem. In model area MST, neurons are trained via Hebbian plasticity to become selective for specific optic flow patterns induced by head and body movements.

Processing in both model stages (V1 and MT) is based on a three-level cascade of basic operations. In a nutshell, the cascade consists of a filter stage where activities are integrated, a modulatory coupling stage where feedback connections can modulate driving FF input activations, and a normalization stage where center-surround shunting inhibition is performed. Importantly, this generic scheme is also employed for form feature extraction in model areas V1 and V2 of the ventral pathway (see 3.2.1). For a more detailed description of the proposed generic processing mechanisms the reader is referred to section 3.2.1 of chapter 2.

The output of the V1-MT model is given by a distributed representation of neural activities represented by a pool of neurons that are tuned to different directions and speeds for each spatial location of the input image. We sample the discrete space of 7 speeds and 16 directions in a 2d velocity space corresponding to 112 neural activities. These local activities generated at the stage of model area MT are integrated by model MST cells to receive a mean velocity vector. This is implemented by computing at each MT location the sum over all discrete feature vectors weighted by their corresponding activity.

### Learning of optic flow patterns

In the next processing stage, model MT responses are pooled by model MST cells. The receptive field of model MST neurons covers the whole input image. We use plastic synapses between model MT and MST neurons that are learned by the same Hebbian

learning rule [Oja, 1982] that was used in the form pathway between V2 and IT neurons. Hebbian learning is used to achieve selectivity of model MST neurons for specific optical flow patterns that are represented by activities of model MT cells.

Neurons tuned to specific flow field pattern have been found in several areas of the dorsal processing stream, specifically in area MSTd [Tanaka *et al.*, 1986; Duffy and Wurtz, 1991; Saito, 1993; Graziano *et al.*, 1994]. As input for the learning stage, we present mean velocities from model area MT to train neurons in model area MST such that they selectively respond to a specific part of a body action.

The activity of model MST neurons $A_j^{MST}$ is further computed by a weighted sum over all positions and features:

$$A_j^{MST} = \sum_{x,v} w_{(x,v)j} \cdot \overline{v}_x^{MT}$$

where the input flow field vector $\overline{v}_x^{MT}$ is multiplied by the weight vector $w_{(x,v)j}$ at the position $x$.

Like in the ventral pathway, lateral inhibition is simulated by applying a MAX-operation over all MST neurons

$$A_j^{MST} = \begin{cases} A_j^{MST} & if\ j = argmax\left(A_j^{MST}\right) \\ 0 & otherwise \end{cases} \tag{4.3}$$

In order to learn the synaptic connections represented by the weights $w_j$ we employ a Hebbian learning rule with a quadratic normalization term [Oja, 1982].

$$\Delta w_{(x,v)j} = \eta \cdot \left(\overline{v}_x^{MT} \cdot A_j^{MST} - \left(\overline{v}_x^{MT}\right)^2 \cdot w_{(x,v)j}\right) \tag{4.4}$$

where $\eta$ represents the learning rate. Remember that for this learning rule the length of the weight vector is bounded and converges to a unit length of one [Oja, 1982]. Again, this is consistent to the corresponding learning step in the proposed model ventral pathway such that they are structurally homologue.

### 4.2.4 Combination of motion and form signals

In the last subsections we described how selectivity for motion and form patterns can be learned and represented by neurons in model area IT and MST. A specific view or optic flow pattern captures only a short temporal interval of the whole body action.

In the next model stage we extend the temporal context to learn selectivity for temporal activity patterns of IT and MST neurons. Importantly, at this model stage both, form and motion information are integrated to learn sequence selectivity.

We suggest model STS sequence neurons that each receive forward projections from a corresponding IT or MST neuron. In addition, these sequence neurons are laterally interconnected by plastic connections. A Hebbian learning mechanism is then used to strengthen connections between sequence neurons that respond successively for a given

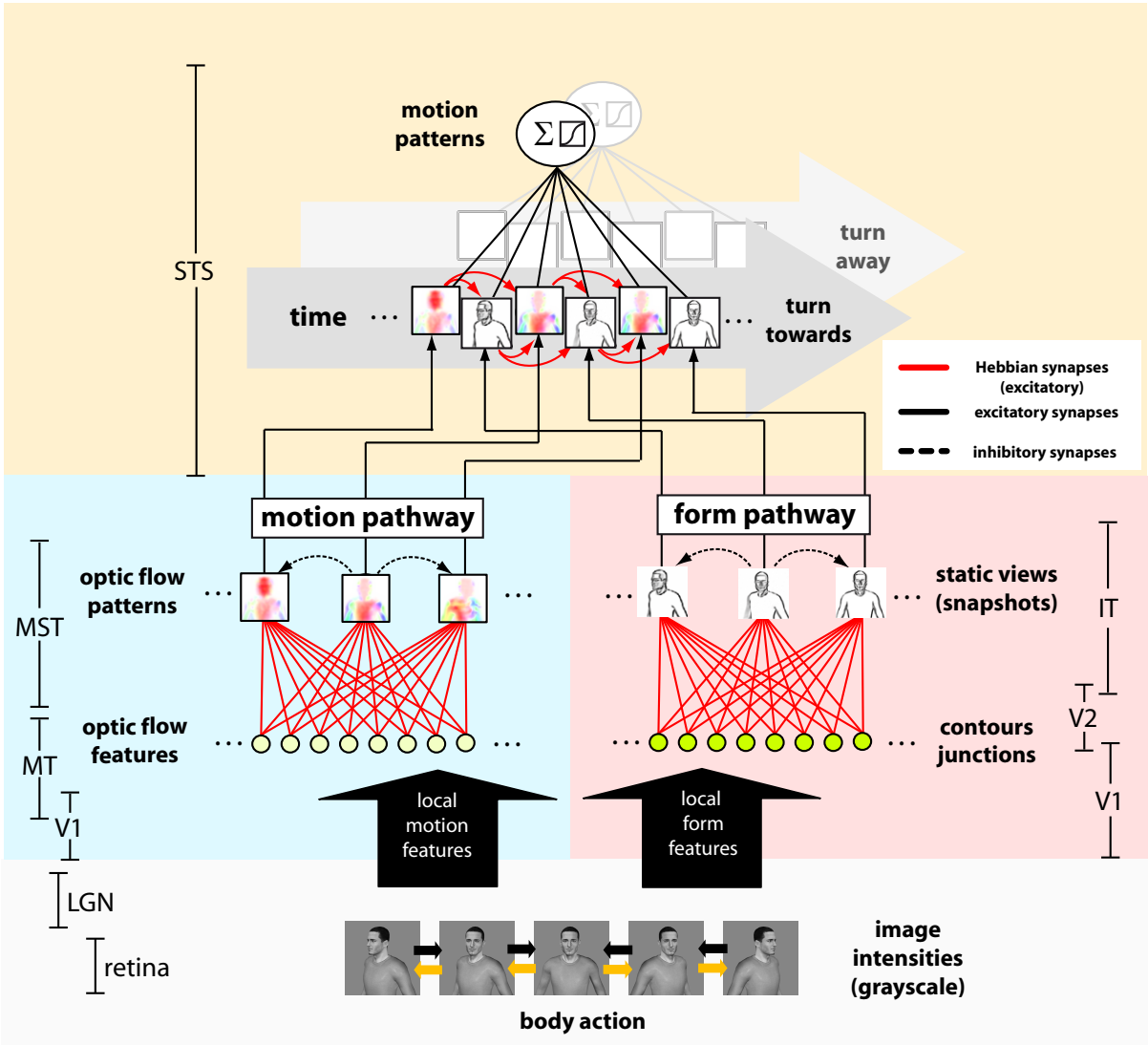


Figure 4.5: Detailed model description. Local motion and form features are extracted from an artificially created input sequence of a character performing a body action. In the motion pathway, MT neurons that encode local optic flow features are fully connected to a population of MST neurons via Hebbian synapses (red) to learn selectivity for short-term optic flow patterns. Likewise, in the form pathway, V2 neurons that encode local form features such as contours and junctions are fully connected via Hebbian synapses to a population of IT neurons to learn selectivity for static views (snapshots). At this stage, lateral inhibition within populations of MST and IT neurons is simulated by a MAX-pooling operation. In a next step, both processing streams are combined in area STS, where sequence neurons receive signals from the form and the motion pathway. Selectivity for temporal response patterns of both, IT and MST neurons, is realized by plastic lateral connections that are adapted by a Hebbian learning rule. The learned asymmetric lateral connections lead to an overall increased mean activity of sequence neurons for the preferred input motion pattern. On top of the model, activity from sequence neurons is integrated by motion pattern neurons.

input sequence of a specific body action (see Fig. 4.5 for a detailed model description).

For instance, an active sequence neuron that encodes a specific body posture or optic flow field pattern can pre-excite sequence neurons that encode the predicted body posture for subsequent time steps. At the same time, sequence neurons that encode unpredicted views are inhibited. The prediction of future events can be simultaneously learned by a Hebbian mechanism, i.e., by strengthening lateral connections between neurons that respond successively and weakening connections to all other neurons. In summary, this mechanism leads to an overall increased activity of sequence neurons for the preferred snapshot sequence. These activities are pooled by motion pattern neurons (see Fig 4.2 and Fig. 4.5) in the next hierarchical processing stage to extract signals for different types of body actions.

### Mathematical description of sequence learning

More formally, we define a population of sequence neurons as $A^{seq}$. Each sequence neuron receives feedforward input $A^{inp}$ from an afferent neuron of a lower model area, i.e., area MT for the motion pathway or area IT for the form pathway. Moreover, all sequence neurons are laterally connected with each other via Hebbian synapses. Recurrent lateral interaction between these sequence neurons leads to increased activity of sequence neurons for specific input sequences from IT and MST neurons.

To model the recurrent interactions between sequence neurons the signal $A^{inp}$ is used as input for a recurrent neural network that has been originally proposed by Amari [Amari, 1972] in the context of pattern sequence learning in self-organizing networks. The temporal course of activation is described in the following ordinary differential equation (ODE)

$$\tau_{seq}\frac{d}{dt}A_j^{seq} = -A_j^{seq} + \alpha A_j^{inp} + \beta\sum_{i=1}^{N} w_{ij}^{+} f(A_j^{seq}) - h \tag{4.5}$$

where $w_{ij}$ describes the lateral coupling strength between sequence neuron $i$ and $j$, $\tau_{seq} = 0.03$ describes the time constant that characterizes the rise time of the cell's membrane potential, $f$ is a sigmoidal nonlinear activation function, and parameter $h = 1$ is a constant that simulates tonic inhibition which leads to a negative resting potential of the cell. The constants $\alpha, \beta = 5$ are constant parameters that have been experimentally determined. Eq. 4.5 can be extended to incorporate a mechanism of *shunting inhibition* [Grossberg, 1988] by replacing the constant $h$ with a negative kernel function

$$h = \left(\gamma + \psi f(A_j^{seq})\right)\sum_{i=1}^{N} w_{ij}^{-} A_j^{inp} \tag{4.6}$$

However, in our simulations we use constant global inhibition which corresponds to $w^{-} = w_{\infty}^{-}$ and $\psi = 0$.

The quadratic matrix $W^{+}$ which includes the lateral coupling strength between sequence neurons is initially set to small random values equally distributed between 0 and 0.1. The matrix $W$ is adapted according to the Hebbian principle in the following equation

$$\tau_w \frac{d}{dt} w_{ij}^{+} = f(A_j^{seq}) \left( A_i^{inp} - f(A_j^{seq}) w_{ij}^{+} \right) \tag{4.7}$$

which implements Oja's rule [Oja, 1982]. Importantly, here we use the same learning rule as in the learning of static views in section 4.2.2 and learning of flow patterns in section 4.2.3. The learning rule constrains the lateral weights $w_{ij}$ to be bounded. The learning rate $\tau_w$ in the Eq. 4.7 has to be chosen sufficiently larger than the time constants of the equations describing the network dynamics. Note that the weights cannot become negative which means that only excitatory lateral connections are learned. However, all sequence neurons receive a tonic inhibition signal represented by constant h in Eq. 4.5 which leads to a relaxation of neural activity.

Finally, activity from sequence neurons ($A^m p$) is accumulated to encode specific motion patterns by a simple leaky integrator described in Eq. 4.8

$$\tau_{mp} \frac{\partial}{\partial t} A^{mp} = -A^{mp} + \sum_{j=1}^{N} f(A_j^{seq}) \tag{4.8}$$

## 4.3 Results

In this section we present results from model simulations at different stages of the form and the motion pathway. Then, we demonstrate that a combination of information from both pathways leads to an improved recognition performance compared to using information of a single pathways.

### 4.3.1 Model input

As model input we use artificially generated sequences of a virtual character performing head and body turns towards and away from the camera. To generate the video sequences we use the animation software *Poser 6*. We create four different sequences of a male virtual character performing turns towards and away from the camera. Turns towards the camera start at half-left or half-right profile view while turns away start in frontal view. The movement of the artificial actor is generated such that every action begins with a rotation of the head and is smoothly followed by a rotation of the body. Such stimuli of articulated body actions were also used in neurophysiological experiments with monkey [Jellema and Perrett, 2003a].

Fig. 4.6a illustrates an input sequence showing a turn away and a turn towards the camera from/to half-left profile. Moreover, the plot in Fig. 4.6b illustrates the sigmoidal course of the function that is used to model the relative temporal evolution of head and body movements. One can observe that the body rotation is temporally delayed with respect to the head rotation. Each sequence consists of 90 image frames that have a size of 256 x 256 pixels.

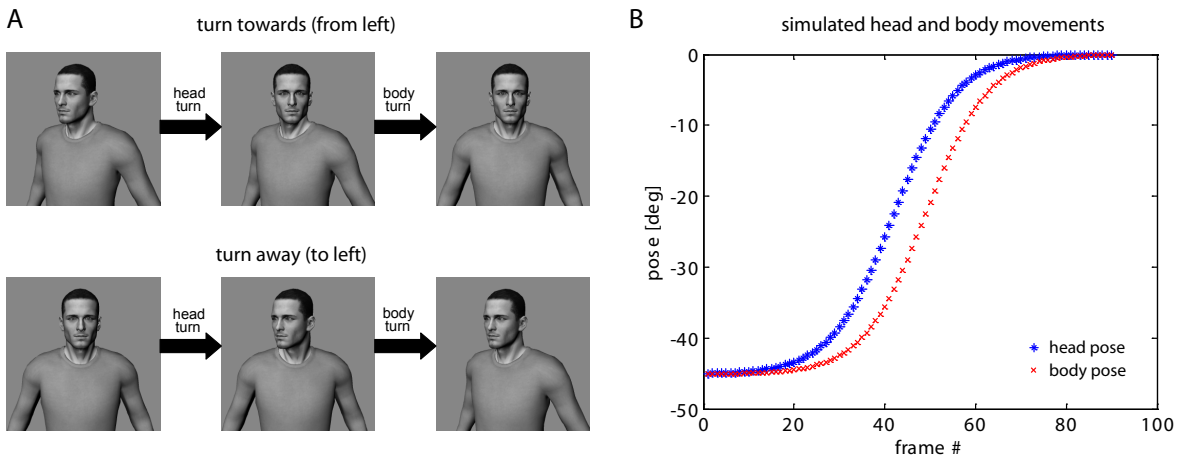


Figure 4.6: We use 4 different artificially generated input sequences of a virtual character who performs head and body turns towards and away from the camera from/to both sides (A). In each sequence, the action starts with a head turn and is smoothly followed by a body turn. The precise progress of head (blue, stars) and body (red, x) poses across image frames for a turn towards the camera is illustrated in (B).

### 4.3.2 Motion pathway

#### Initial motion feature extraction

In the motion pathway, the first processing step comprises the extraction of optic flow field information from the input sequence. Here, we use a model of V1-MT interactions to robustly extract local motion features [Bayerl and Neumann, 2004]. Fig. 4.7 depicts an optic flow field that has been extracted at model area MT. At each location 112 motion features are represented in a polar grid of 16 different directions and 7 speeds. Illustrated in Fig. 4.7 is the mean velocity vector that results from a superposition of all feature vectors weighted by their activity.

In the next step of the processing hierarchy receptive field (RF) properties of model MST neurons are learned unsupervised via Hebbian synapses from input of model MT neurons. Here, the RF size of a MST neuron is equal to the input dimensions, which means that the MST cell covers the whole upper body including the head, upper arms, and chest. We use a predefined number of 4 MST neurons to learn prototypical velocity patterns of head and body movements. The model is provided with sequences of body actions that are randomly selected from one of the 4 different input sequences:

a) turn **towards** from **right** half-profile view to frontal view (**towards/right**)

b) turn **towards** from **left** half-profile view to frontal view (**towards/left**)

c) turn **away** from frontal view to **right** half-profile view (**away/right**)

d) turn **away** from frontal view to **left** half-profile view (**away/left**)

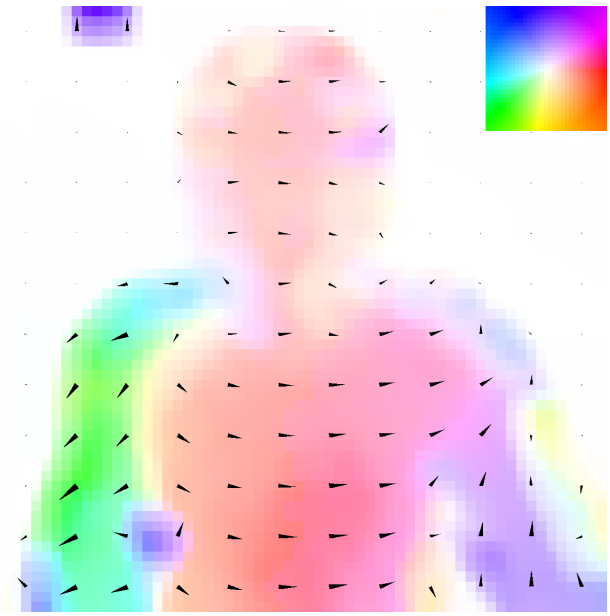

Figure 4.7: Optic flow field extracted at model area MT. Local velocities are color coded (see colormap). The hue channel indicates the direction component while the saturation indicates the speed of the estimated velocity vector. Overlaid is a sub-sampled vector field of arrows to get a more vivid impression of the underlying optic flow field.

### Results of learned optic flow patterns

Learning of optic flow prototypes is performed by repeatedly presenting velocity patterns from model MT neurons in correct temporal order until the weights converge to a stable state. The result of the Hebbian learning stage is demonstrated in Fig. 4.8. One can observe that MST neuron M2 and M6 are tuned mainly to rotations of the head while another MST neuron M1 and M3-M5 are tuned mainly to rotations of the body. This result reflects the asynchrony between head and body rotation that was specified for the motion of the virtual character.

A close view on the results in Fig. 4.8 reveals that the first part of the response patterns is dominated by responses of neurons M6 (rightward head turn) and M2 (leftward head turn), however independent from the **turn away** or **turn towards** class. This means that discrimination performance between the two classes **turn away** and **turn towards** is supposed to be weak for the first part of the sequence. In the second part of the sequence the other four neurons M1, M3, M4, and M5 selectively respond for each of the four inputs (A-D) suggesting for a better discrimination between the two classes in the second part of the sequence which is mainly dominated by upper body rotation.

However, in the second part of the sequence, the upper body rotation, captured by the remaining four prototypes seems to have more discriminative power. This is visible in the response properties of M1 and M3-M5 where each prototype selectively responds for a specific body action category.

In summary, the result of pattern selective processing in the motion pathway suggest that motion information of the head *alone* would not suffice to discriminate between **turn**

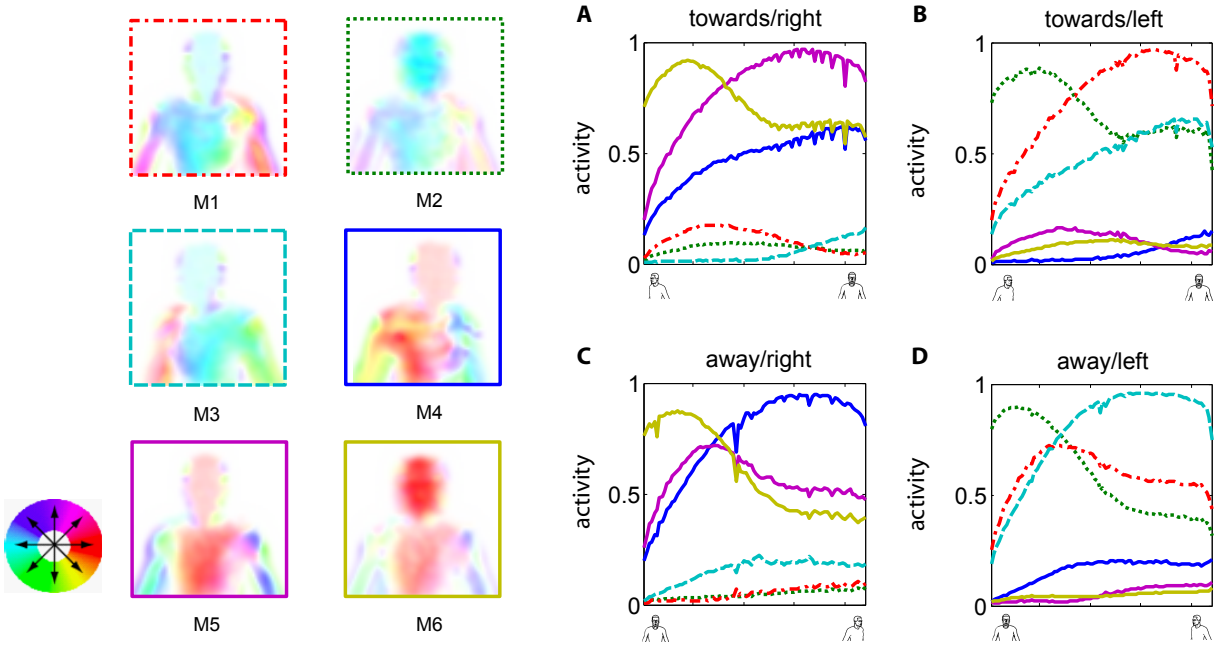


Figure 4.8: Receptive field properties of MST neurons M1-M6 learned via Hebbian learning given a set of 4 different input sequences of a persons that performs head and body turns towards and away from the observer from both sides (left and right). The hue component corresponds to the most selective direction (see color wheel) and the saturation component corresponds to the strength of the weights. The responses of the MST neurons M1-M6 are illustrated for all four types of input sequences (A)-(D). The corresponding prototype is indicated by the colored frames in. If we compare the responses of (A) with the responses of (B) and the responses of (C) with the responses of (D) we can observe that they are mirror images of each other but with different neurons and selectivity. Note, that neurons M6 and M2 encode mainly a head rotation while the other neurons M1, M3, M4 and M5 are mainly selective to body rotations. Note also, that both sequence types, turn away and turn towards, begin with the same head rotation direction, thus making a discrimination task rather difficult based on the responses of M2 and M6.


Figure 4.9: Example of initial form processing. A frame of the input video stream (left) is processed by orientation selective Gabor filters in model area V1. We illustrate the mean activity of V1 and V2 cells (over all orientations) after 5 iterations of FF-FB processing between model areas V1 and V2 (middle and left). Notice that salient contours are extracted and enhanced. The strength of the responses is relatively invariant against local variations in image contrast and brightness.

**towards** and **turn away** action. On the other hand, motion information extracted from body rotation seems to include enough information to discriminate between turn towards and turn away motion, at least in this example. Finally, the selectivity in all cases is rather poor since at least three MST neurons are coactive (have values greater than 0.5) for large parts of the sequence. In this case, robust classification performance based solely on signals from the motion pathway cannot be expected.

### 4.3.3 Form pathway

#### Initial form feature extraction

In the form pathway processing starts with the extraction of local image features. Initially, contrast information is extracted by computing the energy from 8 differently oriented Gabor filters which simulates V1 cells in primary visual cortex. Next, like-oriented oriented features are grouped along smooth contours in model area V2. Furthermore, recurrent feedforward-feedback processing between model areas V1 and V2 in combination with local normalization enhances smoothly curved contours and provides invariance against variations in illumination (see chapter 3 for a more comprehensive explanation of the V1-V2 interaction). Model V1 and V2 activities after 5 iterations of FF-FB processing are illustrated in Fig. 4.9 for one frame of the input sequence. It is clearly visible that salient contours are extracted from the input image that serve as primary features for further processing stages. Notice, that V1 responses are more detailed than V2 responses as V1 cells have smaller receptive fields (RF). V2 cells have larger elongated RF's that consist of two sub-fields to group local features that have similar orientation.

**Results of learned static views (snapshots)**

Each frame of the input sequence is processed by the recurrent V1-V2 model stage. We reduce the distributed representation of 8 orientation-specific responses at each image location to a Cartesian vector representation that consists of a horizontal and a vertical component [2]. The reduced V2 activities are now forwarded to model IT neurons. Importantly, here we use plastic connections between model areas IT and V2. In this model, we manually define a population of 9 IT neurons. The activity of IT neurons is computed by a weighted sum over all activities from afferent V2 neurons. The adjustable weights simulate synaptic plasticity and are adapted by a Hebbian learning rule [Oja, 1982]. Importantly, in one learning step, only weights from afferents to the IT neuron with maximal activity are adapted (winner-take-all strategy). In every trial, a randomly selected input sequence[3] of 88 frames is presented to the model. This is repeated for 100 trials until all weights have converged to a stable state. The resulting weight matrices are illustrated in Fig. 4.10. It can be observed that each IT neuron has developed a preference for a specific body and head pose. In Fig. 4.10, we also illustrate the activities of IT neurons for all input sequences. Indeed, the unimodal distribution of the IT responses demonstrates that each IT neurons selectively responds to a specific body pose. Consequently, we call these neurons '*snapshot*' neurons, as they capture specific body configurations that appear for a short time interval.

### 4.3.4 Combination of form and motion information

In our proposed model, body movements are characterized by specific temporal sequences of key body postures, indicated by model IT neurons of the form pathway. Furthermore, short-term optic flow information is represented by model MST neurons in the motion pathway.

The goal of this model stage is to integrate and learn feedforward activities from the form and motion pathway by adapting recurrent lateral connections between form-driven and motion-driven sequence neurons.

As input to model stage STS we select in each discrete time step the IT neuron and the MST neuron with the strongest response, all other activities are inhibited by explicitly setting them to zero (MAX-rule). This is a simple way to simulate strong lateral inhibition within a pool of neurons and is also employed in other biologically motivated models (e.g. [Riesenhuber and Poggio, 1999]). The resulting input to the next model stage is illustrated in Fig 4.11 for different input sequences.

[2]The reduction is computed by summing up the 8 corresponding vectors, weighted by their actual activity.

[3]A sequence from one of the four possible input sequences, namely turn away left/right and turn towards left/right

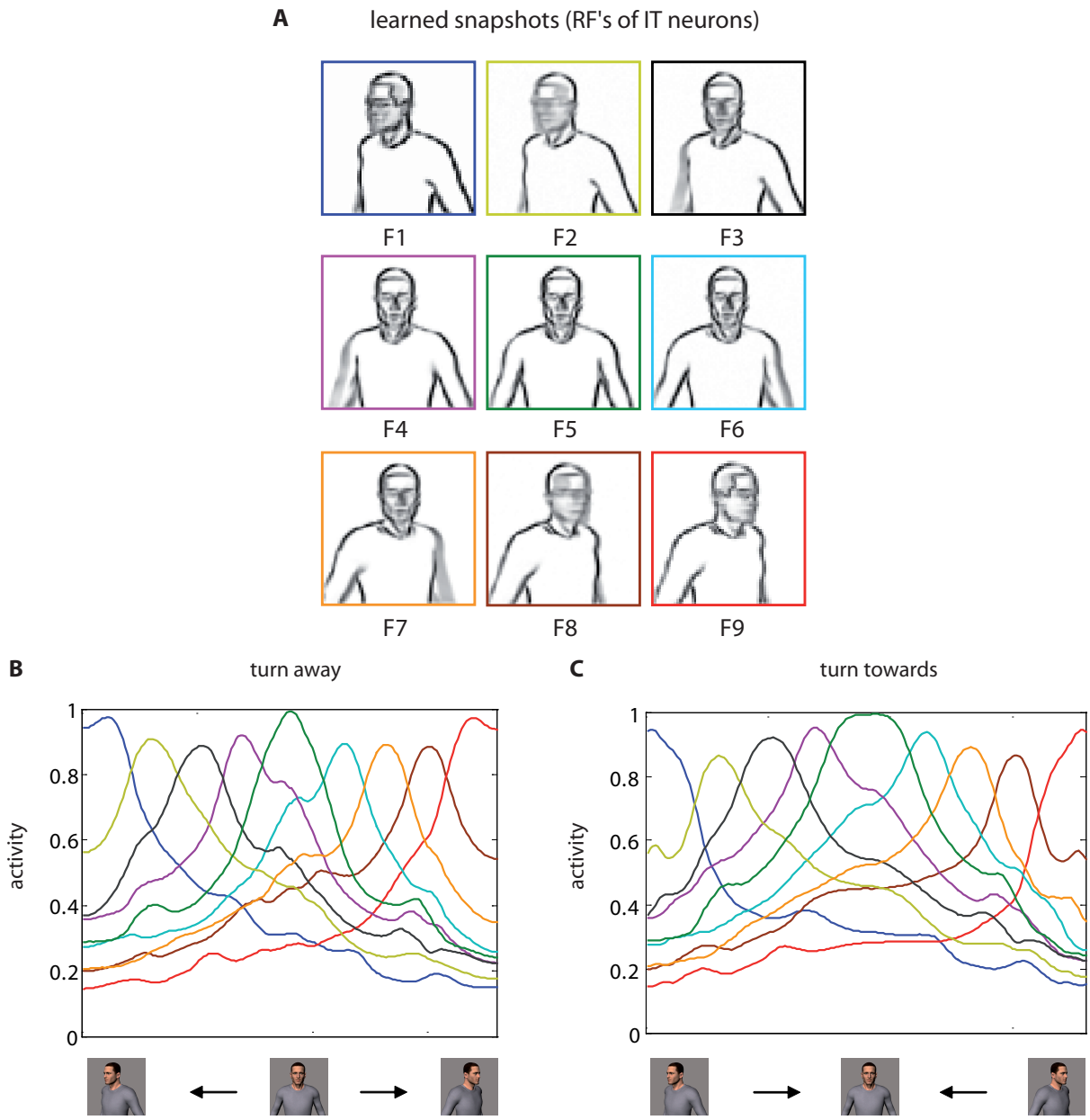


Figure 4.10: Prototypical form representations of body postures ('*snapshots*') encoded by IT neurons (F1-F9) learned via Hebbian plasticity (A) from feedforward input of V2 activities. The labels F1-F9 were assigned manually after the learning stage such that the numbers correspond to the order of responses. At the bottom responses of IT neurons are illustrated for a turn away (B) and a turn towards (C) input sequence. Colors correspond to the IT neurons F1-F9 indicated in (A). Notice, that each neuron selectively responds to a specific body configuration independent of the movement direction. As a consequence the resulting graphs depicted in (B) and (C), respectively, look almost identically. Furthermore, the population of IT neurons covers the whole space of body poses.

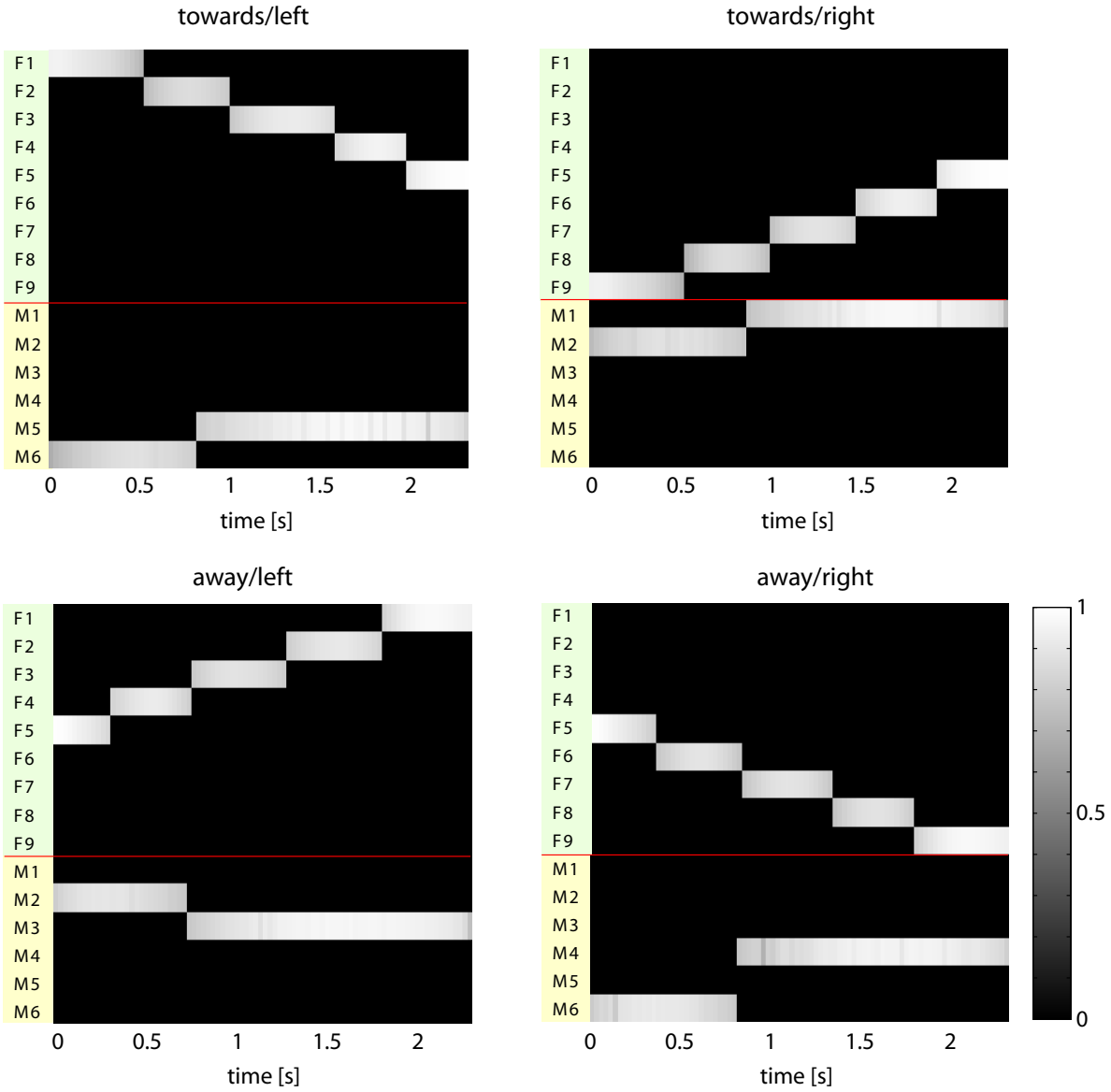


Figure 4.11: Incoming feedforward activities for different input sequences from form and motion pathway after a MAX operation in both channels. IT neurons of the form pathway are denoted with F1-F9, MST neurons from the motion pathway are denoted with M1-M6. The level of activity of individual model neurons is coded in grayscale intensities (see bar). Note that the labels of IT neurons (F1-F9) are assigned such that they are consistent with the order of responses. The labels of MST neurons (M1-M6) are not reordered, such that the numbering does not have any specific meaning and individual selectivity has evolved randomly by the unsupervised learning process.

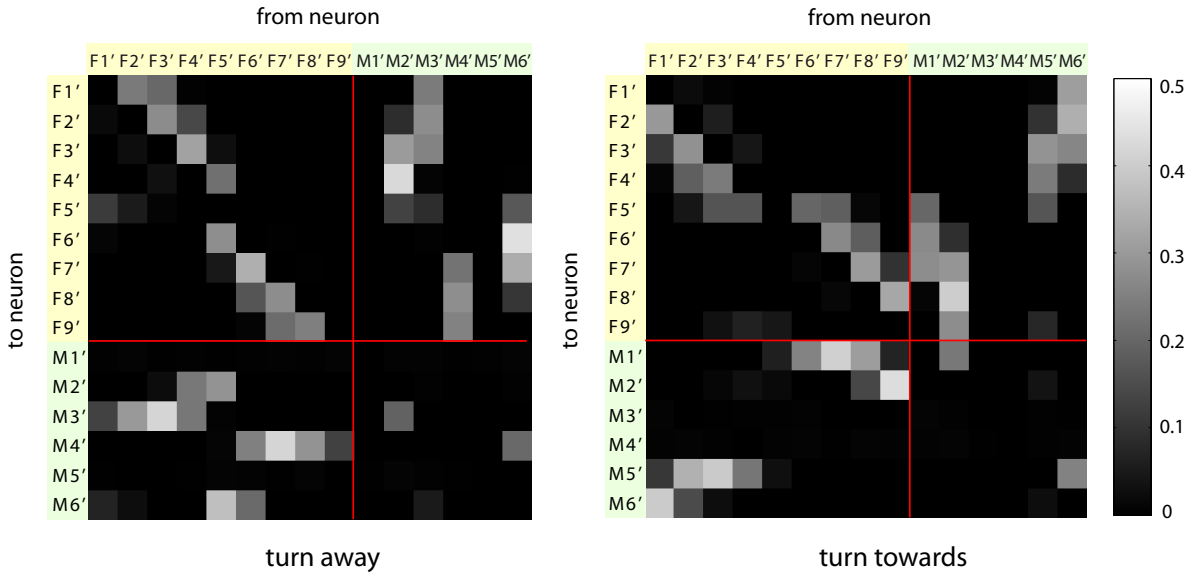


Figure 4.12: Learned lateral connection matrices of sequence neurons for two classes of body action, turn away and turn towards. Note that connections between form-driven sequence neurons (F1'-F9') can be found in the upper left part of the matrix. Connections between motion-driven sequence neurons can be found in the lower right part of the matrix. Connections between form- and motion-driven sequence neurons can be found in the lower left and the upper right part of the weight matrix. Note that specifically these interconnections between form and motion may help to better discriminate two classes.

### Results of sequence learning

To learn the temporal sequence pattern of form and motion responses, we employ a family of sequence neurons that are fully interconnected via dynamic excitatory lateral connections. Each sequence neuron receives input from a corresponding IT or MST neuron. The lateral weights are initialized with small random values[4]. We start training the lateral weights for a specific class of body motion by presenting input from the target class. This can be either turn towards (left/right) or turn away (left/right). The adaptation of lateral connections follows a Hebbian learning principle that has the following characteristics: If output activity of a sequence neuron $S_i$ and input activity that is forwarded from the previous stage to sequence neuron $S_j$ $(i \neq j)$ are both at high level, then the lateral connection between these two sequence neurons $S_i$ and $S_j$ are strengthened. In Fig. 4.12 we show the resulting weight matrices for the two different classes of body action after 500 presentations, respectively. The lateral connections between subsequently responding sequence neurons (F1'-F9',M1'-M6') are strengthened.

The effect of the trained lateral connections on the activity of sequence neurons is illustrated in Fig. 4.13. At the beginning of the preferred input sequence, the first sequence neuron is excited by feedforward input. At the same time, the neurons that are predicted to follow up for the preferred input sequence are pre-excited due to the learned lateral feedback connections. Furthermore, pre-excited sequence neurons can reach significantly

[4]We use random values from a uniform distribution in the interval [0,0.01]

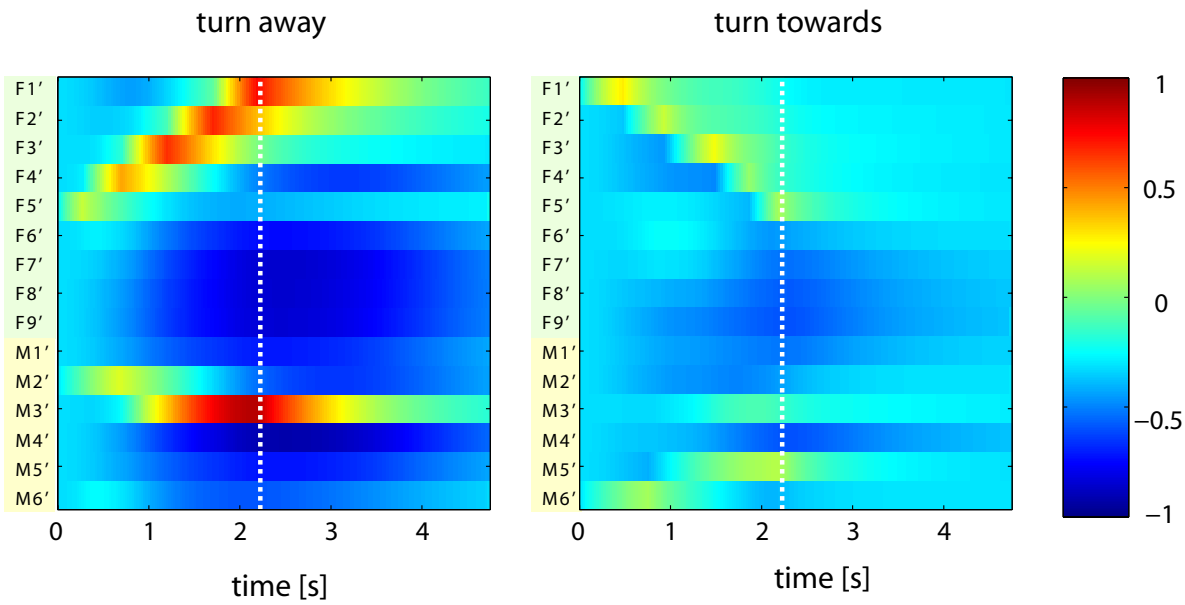


Figure 4.13: Activity pattern of sequence selective neurons (F1'-F9' and M1'-M6') that receive feedforward input from IT and MST neurons (F1-F9 and M1-M6), repectively (see also Fig. 4.5). The learned lateral couplings between sequence neurons in combination with input activity forwarded from IT neurons of the form pathway (F1-F9) and MST neurons of the motion pathway (M1-M5) induce high activations for the preferred turn away sequence (left). The same pool of sequence neurons provided with a non-preferred input of a turn towards sequence shows only weak activations as the actual input pattern does not match to the predicted input encoded by the learned lateral connections. The dashed white line indicates the point in time when the artificial actor has finished his body movement. Afterward, the model input is set to a homogeneously gray image.

higher activations when they also receive feedforward input. In summary, this leads to considerably higher activations for the whole pool of sequence neurons for a preferred input sequence.

In contrast, when a different input sequence is presented the overall activation should be considerably lower as the lateral feedback signal does not match to the predicted feedforward input. In Fig. 4.13 we show the activities of two pools of sequence selective neurons. The first pool of neurons was trained with motion- and form-based signals from a turn away input sequence while the second pool was trained with turn towards sequences. It can be observed that the pool of neurons that receives the correct input reaches much higher activations as the other pool of neurons that was trained on a turn towards sequence.

### Multi vs. single pathway model

Activations from each pool of sequence neurons are further accumulated by motion pattern neurons. In our simulations, we use two motion pattern neurons that integrate activities from two pools of sequence neurons that are trained for turn towards and turn away motion patterns, respectively. We show activity signals based on input from four artificial motion

sequences (turn towards from left/right and turn away to left/right) of two different classes of body action. In order to demonstrate that both form and motion information is important to discriminate between two different motion patterns, we run a simulation under three different model configurations. In the first configuration, sequence neurons are provided with motion and form signals. In the second and third configuration, we simulate the absence of either form or motion signals by simply setting the motion or form input activity to zero.

The temporal progress of neural activity for both motion pattern neurons is illustrated in Fig. 4.14. The figure clearly demonstrates for motion pattern neurons that the ratio of activation between the preferred and non-preferred input pattern is maximal when the sequence neurons are provided with signals from both, form and motion pathway. If the model is only provided with form information the activation of the preferred neuron drops down to a level of about 60 percent. However, the signal for the non-preferred input pattern remains at the same level, indicating that the discrimination performance is poorer when the model is solely provided with form information.

Finally, we simulated the inverse case where signals from the form pathway are suppressed. This means that only motion information is used for the processing of sequence selectivity. In this case the ratio of motion pattern signals is significantly lower than in the other two cases suggesting that motion information alone would not be sufficient to robustly discriminated between turn away and turn towards body actions. This result is not surprising as we have already noticed from responses of MST neurons that flow field information of a head turn in one direction is suggestive for both, either a turn towards the profile view or a turn away from the profile view. Thus, motion information in this specific discrimination task only gives additional cues but does not suffice to robustly discriminate between both classes of body action.

A comparison between maximal signal ratios is illustrated in Fig. 4.15. In the case of form and motion input, the maximal motion pattern signal for the preferred input is about 12 times higher than the signal for the non-preferred input. When only form information is processed (realized by suppression of input from the motion pathway) the signal ratio drops to about 5 and when only motion information is provided (realized by suppression of input from the form pathway) the signal ratio is even lower at a value of approximately 2. In summary, this indicates that form and motion information appropriately combined leads to a significantly better performance compared to single pathway based approaches.

### Implied motion effects

In a next simulation, we have investigated the effect of presenting a static body pose prior to the animated body action. Here, the hypothesis was that the presented static body pose should induce slightly increased activity for form- and motion-based sequence neurons that are predicted to follow up next. Thus, when the animated sequence starts activity of neurons that encode future events should emerge more quickly compared to the scenario when the action sequence starts immediately after a blank screen.

In fact, the results in Fig. 4.16 confirm that the motion pattern neuron trained for

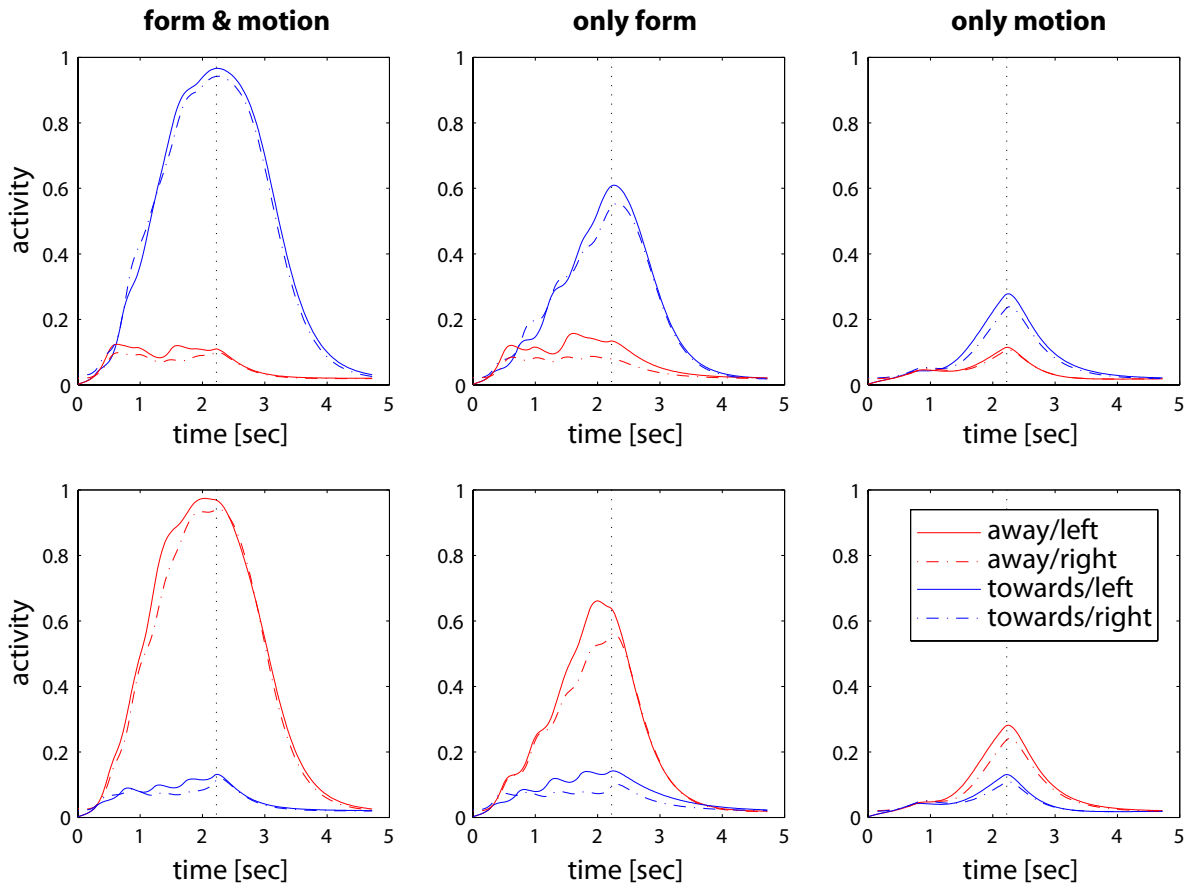


Figure 4.14: Activities of motion pattern neurons selective for turns towards (top row) and turns away (bottom row) for different model configurations. The first row shows temporal response curves from the motion pattern neuron selective for turns towards while the second row shows responses for the motion pattern neuron selective for turns away. If both, form and motion information is processed by sequence neurons, the mean activation level and the ratio between the signals for the preferred and the non-preferred stimulus is high. If feedforward signals from one processing stream are suppressed, the mean activation level for the preferred stimulus and the signal ratio are both reduced significantly. In this specific scenario, form information seems to be more meaningful than motion information as indicated by the increased response level.

turns towards the observer more reaches its peak activation when the static averted pose is presented at the beginning of the sequence. At the same time, the peak activation is higher compared to the peak activation in the simulations without static input. Interestingly it can be also observed that the same motion pattern neuron initially shows increased activity for a static frontal pose. However, when the static frontal pose is followed by a turn away sequence (which is in contrast to the prediction encoded in sequence neurons), the activity quickly drops down to activations close to zero.

This interesting effect can be explained by mechanisms in model area STS. Sequence neurons that are provided with input from IT cells of the form pathway are activated by the static input. Other sequence neurons that encode predicted form and motion appearances are also pre-excited through lateral connections (previously adapted via Hebbian

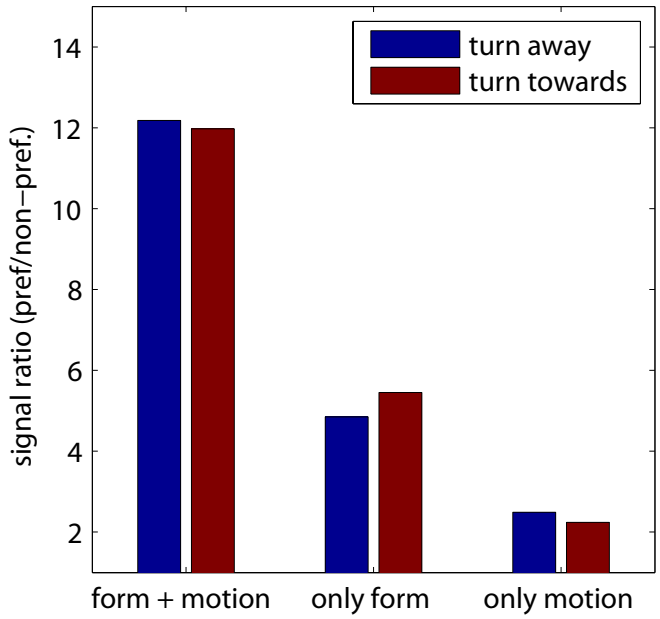


Figure 4.15: Discrimination performance for different model configurations indicated by the maximal signal ratio between activities of two motion pattern neurons that have been adapted to be responsive to turn away and turn towards sequences. A larger signal ratio indicates more robustness against perturbations of the input signals. The figure shows clearly that the highest signal ratios are generated when input from motion and form pathway is forwarded to the integration stage. However, when the model is modified such that signals from one pathway are suppressed a significant drop of signal ratios can be observed. Also, an important observation is that form information seems to be more informative than motion information for this specific scenario. Finally, the results indicate that the combined signal originates from a non-linear combination of form and motion signals (since the combined signal is obviously not just generated by the sum of form-based and motion-based signal ratios).

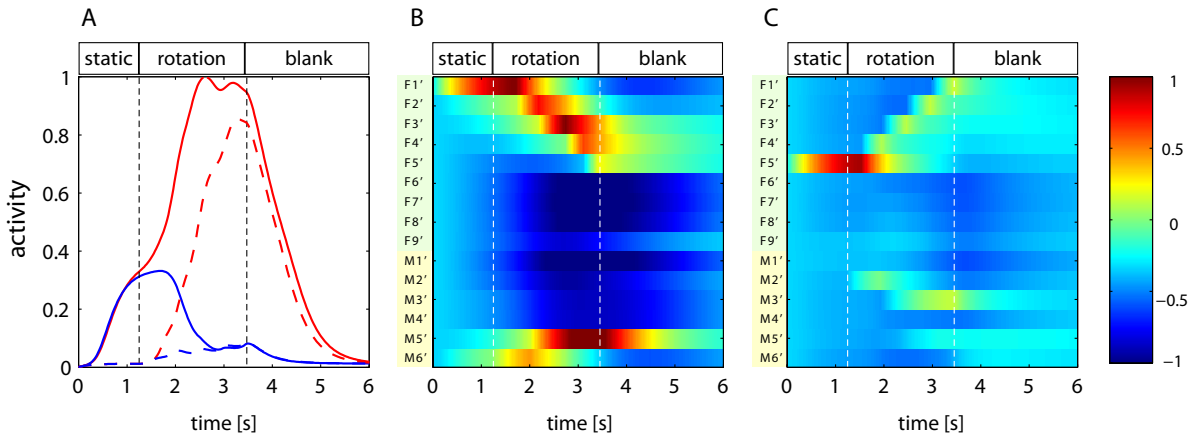


Figure 4.16: Results of model simulations that are based on a slightly modified input sequence. Prior to the beginning of the animated body action we show a static body pose that corresponds to the first frame of the input sequence. After the character has reached his final pose, we show a blank screen until the activities have converged to zero. Here we show results for a family of sequence neurons that have been adapted to body turns towards the observer. The activities of sequence neurons are based on a turn towards (B) and a turn away (C) input sequence. Activities for the corresponding motion pattern neuron are illustrated for a turn towards sequence (solid red) and a turn away sequence (solid blue). For comparison, we also show the corresponding activities based on input sequences used in previous simulations (Fig. 4.14) where the static input at the beginning was replaced with a blank screen (dashed).

learning). Thus, when the actual motion sequence starts these pre-excited sequence neurons receive additional feedforward input which leads to a faster response characteristic.

## 4.4 Discussion

In this section we summarize our main findings. We compare our model with other proposed models that are related to our work. Furthermore, we demonstrate that all core mechanisms employed in our model are biological plausible. We also discuss alternative model architectures to combine motion and form information. Finally, we briefly discuss some limitations of our model.

### 4.4.1 Summary of findings

We have presented a biologically motivated model that simulates processing of body action sequences in two parallel cortical pathways of the human visual system. Several cortical areas were simulated, that is V1, V2, IT in the form pathway and V1, MT and MST in the motion pathway.

For the robust extraction of initial form and motion features we have employed a generic model building block, consisting of feedforward and feedback processing in combination with lateral shunting inhibition between model areas V1 and V2 in the form pathway [Weidenbacher *et al.*, 2006; Weidenbacher and Neumann, 2009] and between

model areas V1 and MT in the motion pathway [Bayerl and Neumann, 2004; Raudis and Neumann, 2010].

We have shown that intermediate neural representations can be learned in both pathways by a Hebbian learning mechanism. Here, different key body postures that occur during a body action were extracted and represented in the form pathway. Likewise, in the motion pathway different short-term optic flow patterns were learned using the same mechanisms.

Furthermore, we have proposed a mechanism to combine information from both pathways in a higher model area that is sensitive to the sequence of incoming responses from mid-level feature neurons from both model pathways. To learn the temporal sequence from incoming feedforward responses of form and motion pathways we used time-dependent Hebbian plasticity in combination with recurrent lateral interactions between sequence neurons.

We have shown that lateral connections within a pool of motion- and form-driven sequence neurons were adapted by Hebbian plasticity such that significant activity can only emerge, when the temporal sequence of incoming motion and form signals fits to the prediction that was learned based on the perceptual history. This is in line with physiological experiments of [Jellema and Perrett, 2003a] where they show responsive cells in the anterior part of the superior temporal sulcus (STSa) of the macaque monkey code for specific articulated body actions and the consequent articulated static view.

Our experiments have also demonstrated that the combination of motion and form signals leads to significantly enhanced signal ratios between activities of motion pattern cells which encode different body actions compared to simulations where only signals from one pathway are forwarded (see Fig. 4.15). This suggests for a better discrimination performance between different body actions if signals from both pathways are processed in model area STS.

Moreover, our results indicate that for this specific task of action recognition, a solely form-based model that represent series of key body postures performs better in discriminating different body action than a solely motion-based model that is based on prototypical optic flow patterns. A plausible explanation for this effect is that optic flow information in this scenario is less informative as it encodes mainly the direction of rotation of individual body parts (head/body). However, the task of discriminating between turn towards and turn away actions is under-determined given only the direction of rotation. Here, the incorporation of form information helps constraining the problem by additionally providing information about the current body pose (see Fig. 4.17).

Finally, we have shown that a combination of static and dynamic input leads to interesting model behavior. Our model results suggest that a body action is *implied* from a single static body posture that is presented prior to the actual motion sequence. This is indicated by a steeper pre-activation profile in the response of motion pattern neurons when they have '*seen*' the static pose before the motion sequence starts.

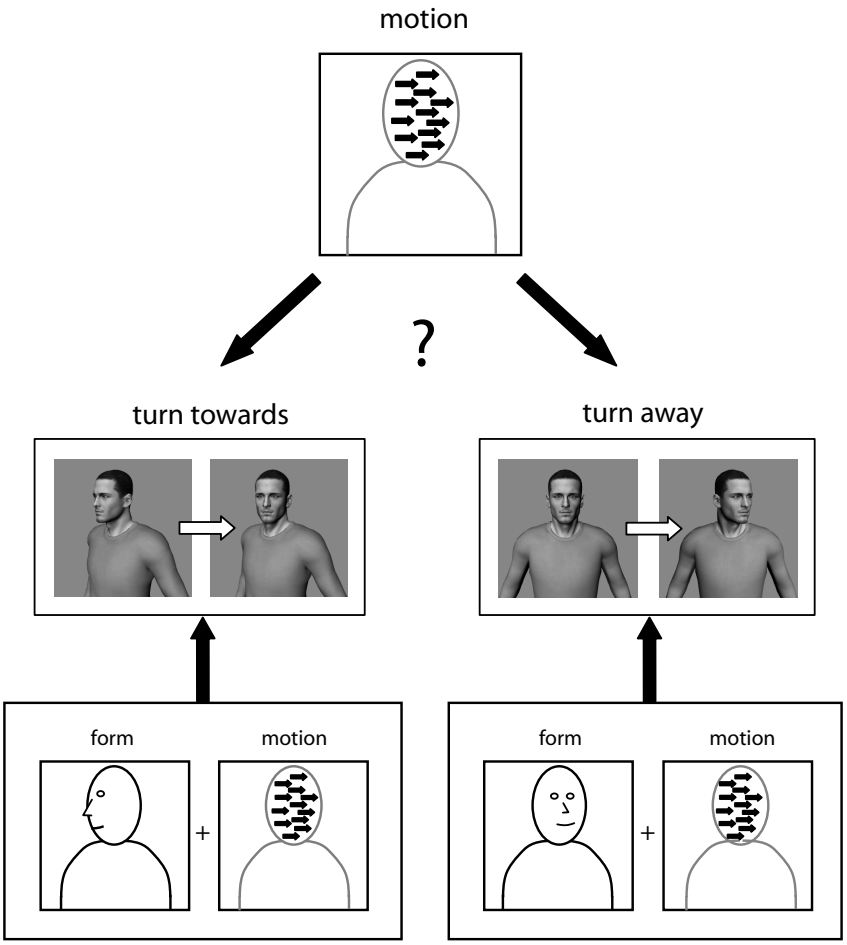


Figure 4.17: Motion information represented as optic flow pattern is not sufficient to robustly discriminate between articulated body actions such as turns towards and turns away from an observer. Knowledge about the current body pose is required to constrain the problem. It is therefore plausible that a combination of both information sources should lead to better discrimination results.

### 4.4.2 Related Work

Several neural network models have been proposed for the recognition of action sequences [Goddard, 1992; Rosenblum *et al.*, 1996; Little and Boyd, 1998]. Most of them are solely based on motion feature processing but only few models exist where form and motion information are both incorporated. For instance, [Schindler and van Gool, 2008] present a biologically inspired system for action recognition from very short sequences ("snippets") of 1–10 frames. They systematically evaluated their model on standard data sets of biological motion sequences. It turned out that even local shape and optic flow for a single frame are enough to achieve ≈90% correct recognitions, and snippets of 5-7 frames (0.3-0.5 seconds of video) are enough to achieve a performance similar to the one obtainable with the entire video sequence. However, they use a simple concatenation technique of form and motion feature vectors which are passed as input to a supervised support vector machine (SVM) classifier. In contrast, in our model all implemented processing mechanisms, including unsupervised learning, have a biological relevance.

Another biologically motivated system for the recognition of actions from real video sequences is proposed by [Jhuang *et al.*, 2007]. Their model approach is based on a hierarchical feedforward architecture inspired by [Riesenhuber and Poggio, 1999] where they use motion feature detectors of increasing complexity. Performance was systematically tested on standard data sets. Again, the classification stage is realized by a multi-class SVM. The model only accounts for the motion pathway.

[Niebles and Fei-Fei, 2007] present a model for human action categorization where video sequences are processed by extracting static and dynamic interest points. A hierarchical model is proposed that can be characterized as a constellation of bags-of-features and that is able to combine both spatial and spatial-temporal features. Their model simulations show that using both dynamic and static features provides a richer representation of human actions when compared to the use of a single feature type, which is consistent with our findings.

[Giese and Poggio, 2004] review psychophysical, neurophysiological and imaging studies of movement recognition. Furthermore, a computational model is proposed that is consistent with experimental data. The model addresses the question of what are the roles of form and motion pathways for the recognition of biological movements. Inspired by the model architecture of Giese & Poggio, we realized the idea of a neural architecture that incorporates several cortical stages of both form and motion pathways. Moreover, we have extended the model to address the outstanding question of *how* form and motion pathways could be combined in a biologically plausible way.

[Lange and Lappe, 2006] propose a computational model to explain the possible contributions of form and motion signals to biological motion perception. The form-based model consists of two stages: a first stage for the analysis of the body posture (form) and a second stage for the analysis of the temporal order of stimulus postures (motion). The approach resembles form processing in model areas IT and STS of our model. In contrast to our model they used computer generated stick-figure sequences of a walking human as input stimulus. Moreover, their model treats the body as global figure without

taking into account local stimulus features. Also, the templates that are used as prototypical representations of different body postures are predetermined. Instead, we have employed unsupervised learning mechanisms to automatically extract such prototypical representation from the input data.

### 4.4.3 Biological Plausibility

In this subsection we put individual model components into a physiological context and discuss for their plausibility.

#### Low- and mid-level features representations in form and motion pathways

Our proposed model simulates biological mechanisms of form and motion processing at several stages of the ventral and dorsal pathway of the visual cortex.

The primary visual cortex (V1) is the basis of two different pathways, namely the dorsal and the ventral pathway [Ungerleider and Mishkin, 1982; Mishkin *et al.*, 1983]. In area V1, local form features such as oriented contrasts are extracted and represented in hypercolumns [Hubel and Wiesel, 1962, 1968]. Cells in V1 also encode local motion features such as direction and speed selectivity [Mikami *et al.*, 1986; Pack *et al.*, 2003]. Thus, our representations in model area V1 are consistent with these physiological findings.

The form pathway is continued in area V2 [Fellemann and van Essen, 1991], where long-range lateral interactions have the effect grouping local oriented contrast into smooth contours [Peterhans and Heydt, 1989; von der Heydt *et al.*, 1984]. In the motion pathway, V1 projects to the medial temporal cortex (MT/V5) where neural activities are pooled from a broad spatial context [Movshon *et al.*, 1985; Pack and Born, 2001].

Recurrent feedback processing mechanisms between early visual areas V1 / V2 in the form pathway and V1 / MT in the motion pathway are incorporated in our model to improve the robustness of initial form representations [Weidenbacher and Neumann, 2009] and to disambiguate initial motion estimates [Bayerl and Neumann, 2004]. Physiological evidence for such modulatory feedback mechanisms was found in monkey visual cortex [Bullier *et al.*, 1988; Hupé *et al.*, 2001]. More comprehensive discussions on the role of feedback can be found in [Bullier, 2001; Sillito *et al.*, 2006].

The next level in the motion pathway of our model corresponds to medial superior temporal cortex (MST) where neurons are tuned to optic flow field patterns learned through Hebbian synaptic plasticity [Krikwood and Bear, 1994]. Neurons selective to such complex motion stimuli were identified in several physiological studies, e.g. [Duffy and Wurtz, 1991; Graziano *et al.*, 1994; Geesaman and Andersen, 1996].

Likewise, in the form pathway model neurons simulate view-tuned neurons found in monkey inferotemporal cortex (area IT) [Tanaka, 1996; Kobatake and Tanaka, 1994], which become tuned to complex shapes through learning [Logothetis *et al.*, 1995].

### High-level visual representations of observed actions

A brain area that is known to receive projections from area MST of the dorsal stream and area IT of the ventral stream is situated in the superior temporal sulcus (STS). Several studies have investigated the responsiveness of cells in monkey STS to particular body actions and body postures in the context of social cognition (for a review see [Jellema and Perrett, 2007] or [Peelen and Downing, 2008]).

One of the first studies that reports on cells in the STS of macaque monkey which showed sensitivity to articulation and rotation of the body and head was conducted by [Perrett *et al.*, 1985a]. Similarly, [Oram and Perrett, 1996] found cells in the anterior part of superior temporal sulcus (STSa) that are both sensitive to form (head and bodies) and motion direction. Some of these cells are selective for both motion and form of a single object, not simply the juxtaposition. The majority of responses were characterized as showing nonlinear dependencies between form and motion inputs. Indeed, model simulations confirm these observation by showing that the combination of form and motion signals must be of non-linear fashion (see Fig. 4.15)

More recently, [Jellema and Perrett, 2003b] showed that the response of cells in the temporal lobe of the macaque to the sight of static head and body postures is controlled by the sight of immediately preceding actions. Moreover, their results support the view that cells in the temporal cortex could support the formation of expectations about impending behavior of others.

In another paper [Jellema and Perrett, 2003a], cells in STS were tested with articulated (two or more body parts move independently from each other) and non-articulated (the body moves as a rigid object) actions of the upper body and head. The cells studied did not respond to non-articulated static posture but vigorously to the articulated posture that form the end-point of the action.

Together, these findings suggest that cell populations in the banks of the STS are sensitive to the perceptual history which might enable the prediction of future actions. Also, there is converging evidence that interaction between form and motion signals in area STS plays a crucial role in the recognition of socially relevant body actions. Thus, although the role of STS in biological motion recognition is undisputed, the contribution of signals feeding into the STS and their combination is unclear.

We have addressed these physiological findings by simulations of a model STS. A key element of this model area are the lateral connections between form- and motion-driven sequence neurons that are adapted using Hebbian plasticity. Our simulations demonstrate that a preceding static view of a person can lead to earlier responses of motion pattern STS cell when the following action sequence constitutes the expected continuation of the static view. Finally, we have shown that a purely motion- or form-driven model performs significantly worse compared to the full model that integrates both, motion and form signals.

### 4.4.4 Limitation of the model

There are several limitation in our model that are worth mentioning.

First, training in our model is based on a single virtual character performing action in front of a homogeneous background. This means that the model cannot cope with background clutter or occlusions at the current state. Second, we assume that the character is placed approximately in the center of the image at a predefined distance to the camera. Thus, our model is not invariant against translation or scale. Third, the model is only tuned to a specific speed of the body action.

In general, most of these invariance could be achieved by replicating the model at different locations or spatial and temporal scales in combination with a MAX-pooling mechanism.

Importantly, the goal of this contribution is not to compete with other models in terms of high recognition rates or invariance performance. Here, we present a biological plausible architecture to address the open question of *how* form and motion signals could interact and what the roles of form and motion information are in the processing of social actions. In order to demonstrate the corresponding basic functionality, we have implemented in detail the main cortical processing stages including low-, mid- and high level processing of form and motion information.

### 4.4.5 Open questions

Our model architecture leaves several questions open. For example, in our model architecture we employ a fixed number of snapshot neurons and optic-flow patterns neurons. Unsupervised learning methods could be incorporated that dynamically acquire a suitable number of neurons depending on the complexity of the data.

We have proposed one possible mechanism to combine form and motion signals at the level of sequence neurons in model area STS. Alternative architectures could be tested where form and motion pathways (additionally) interact at a higher level. For instance, this could be interaction between form-driven and motion-driven motion pattern neurons. Likewise, it would be interesting to examine form-motion interactions one level below between IT and MST neurons.

Finally, our model architecture includes feedforward, feedback and recurrent lateral connections. However, mid- and high level processing is mainly based on feedforward processing. Future model versions could also incorporate feedback connection between mid- and higher model areas to investigate top-down effects for both form and motion delivered by a reverse signal flow from representations at STS.

# Chapter 5

# Summary

## 5.1 A survey of major results

Throughout this thesis we have presented a neural model of form processing that makes the following major contributions:

- We have demonstrated that the model successfully extracts a robust representation of low level features, invariant against illumination an small changes in viewpoint [Weidenbacher *et al.*, 2005c].
- A sketch-representation was extracted from model activities that visualizes perceptually relevant surface features from perfectly mirrored objects which remain largely invariant under different environmental scenes [Weidenbacher *et al.*, 2006a].
- We have shown that the model is able to identify and recognize salient image structures such as smooth contours and different junction configuration which play an important role for the detection of occlusions between scene elements. A comparison based on input of benchmark images has shown that our our model results outperform state-of-the-art machine vision approaches for corner detection [Weidenbacher and Neumann, 2009a]
- We have recorded a comprehensive head pose and gaze database that consists of over 2000 images of faces from 20 subjects. A large amount of different combinations between eye gaze and head pose (horizontally and vertically) was recorded. [Weidenbacher *et al.*, 2007]
- An extended version of the model was presented that incorporates learning of prototypical representations 'snapshots' of faces and body configurations. Here, we proposed a novel approach to combine form-based representation with motion-based representation to detect typical body actions in human social interaction. We have shown that a combined signal based on interactions between both pathways leads to significantly better results compared to model simulations where only one pathway was deactivated [Weidenbacher and Neumann, 2009b].

## 5.2 Relevant publications

Weidenbacher, U., Fleming, R., Bayerl, P., and Neumann, H. (2005a). Perception of mirrored objects. *Journal of Vision Supplement*, Proceesings of the 5th Annual Meeting of the Vision Sciences Society (VSS '05), **5**(8), 526.

Weidenbacher, U., Bayerl, P., Fleming, R., and Neumann, H. (2005b). Perception of perfectly mirrored objects. *Perception ECVP Abstract Supplement*, **34**, page 177.

Weidenbacher, U., Bayerl, P., Fleming, R., and Neumann, H. (2005c). Extracting and Depicting the 3D Shape of Specular Surfaces. *ACM SIGGRAPH Conf. on Applied Perception in Graphics and Visualization*, **95**, 83-86.

Weidenbacher, U., Bayerl, P., Fleming, R., and Neumann, H. (2006a). Sketching Shiny Surfaces. *ACM Transactions on Applied Perception*, **(3)**3, 262-285.

Weidenbacher, U., Layher, G., and Neumann, H. (2006b). Detection of head and gaze direction for human-computer interaction. *Perception and Interactive Technologies*, Springer LNAI 4021, 9-19.

Weidenbacher, U., Bayerl, P., and Neumann, H. (2006c). Generation of sketch-like feature encodings in oriented faces - A neural model. *Journal of Vision Supplement*, Proceedings of the 6th Annual Meeting of the Vision Sciences Society (VSS '06), **6**(6), 1069.

Weidenbacher, U., Layher, G., and Neumann, H. (2007). A comprehensive head pose and gaze database. *IET International Conference on Intelligent Environments*, 455-458.

Weidenbacher, U. and Neumann, H. (2008a). The first spike counts: A model for STDP learning pose specific representations for estimating view direction. *Journal of Vision Supplement*, Proceedings of the 8th Annual Meeting of the Vision Sciences Society (VSS '08), **8**(6) , 161, 161a.

Weidenbacher, U. and Neumann, H. (2008b). Unsupervised learning of head pose through spike-timing dependent plasticity. *IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, Springer LNAI **5078**, 123-131.

Weidenbacher, U. and Neumann, H. (2009a). Extraction of surface-related features in a recurrent model of V1-V2 interactions. *PLoS ONE* 4(6): e5909. doi:10.1371/journal.pone.0005909

Weidenbacher, U. and Neumann, H. (2009b). Learning of motion and form patterns from head and body movements for the analysis of human visual communication. *Perception ECVP Abstract Supplement*, **38**, page 51

# References


Abbott, L. (1999). Lapicque's introduction of the integrate-and-fire model neuron (1907). *Brain Research Bulletin*, **50**, 303–304.

Abraham, W. C. (1997). Metaplasticity, a new vista across the field of synaptic plasticity. *Progress in Neurobiology*, **52**, 303–323.

Ahissar, M. and Hochstein, S. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, **36**, 791–804.

Albright, T. (1984). Direction and orientation selectivity of neurons in visual area mt of the macaque. *Journal of Neurophysiology*, **52**, 1106–1130.

Allison, T., Puce, A., and McCarthy, G. (2000). Social perception from visual cues: role of the sts region. *Trends in Cognitive Sciences*, **4**, 251–291.

Amari, S. (1972). Learning patterns and pattern sequences by self organizing nets of threshold elements. *IEEE Transactions on Computers*, **21**, 1197–1206.

Anderson, B. L. (1997). A theory of illusory lightness and transparency in monocular and binocular images: The role of contour junctions. *Perception*, **26**, 419–453.

Anzai, A., Peng, X., and van Essen, D. C. (2007). Neurons in monkey visual area v2 encode combinations of orientations. *Nature Neuroscience*, **10**, 1313–1321.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, **61**, 183–193.

Bair, A., House, D., and Ware, C. (2005). A perceptually optimizing textures for layered surfaces. *Applied Perception in Graphics and Visualization*, pages 67–74.

Bayerl, P. and Neumann, H. (2004). Disambiguating visual motion through contextual feedback modulation. *Neural Computation*, **16**, 2041–2066.

Baylis, G. and Driver, J. (2001). Generalization over contrast and mirror reversal, but not figure-ground reversal. *Nature Neuroscience*, **4**, 937–942.

Beck, J. and Prazdny, S. (1981). Highlights and the perception of glossiness. *Perception & Psychophysics*, **30**(4), 407–410.

Benson, P. and Perrett, D. (1991). Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology*, **3**(1), 105–135.

Biederman, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review*, **94**(2), 115–147.

Bienenstock, E., Cooper, L., and Munro, L. (1982). Theory for the development of neuron selectivity, orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, **2**, 32–48.

Blake, A. (1985). Specular stereo. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 973–976.

Blake, A. and Brelstaff, G. (1988). Geometry from specularity. In *Proceedings of International Conference on Computer Vision*, pages 394–403.

Blake, A. and Bülthoff, H. (1990). Does the brain know the physics of specular reflection? *Nature*, **343**, 165–168.

Blake, A. and Bülthoff, H. (1991). Shape from specularities: Computation and psychophysics. *Philosophical Transactions of the Royal Society of London B*, **331**, 237–252.

Bosking, W. H., Zhang, Y., Schofield, B., and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *Journal of Neuroscience*, **17**, 2112–2127.

Boynton, G. M. and Hedgé, J. (2004). Visual cortex: The continuing puzzle of area v2. *Current Biology*, **14**, R523–R524.

Bruckstein, A. M. (1988). On shape from shading. *Computer Vision, Graphics and Image Processing*, **44**(2), 139–154.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews*, **36**, 96–107.

Bullier, J. H., McCourt, M. E., and Henry, G. H. (1988). Physiological studies of the feedback connection to the striate cortex from areas 18 and 19 of the cat. *Experimental Brain Research*, **70**, 90–98.

Cabral, B. and Leedom, L. C. (1993). Image vector fields using line integral convolution. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 263–270. ACM Press.

Caradini, M. and Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science*, **264**, 1333–1336.

Carpenter, G. and Grossberg, S. (1988). The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, **21**, 77–88.

Chen, H., Liu, Z., Rose, C., Xu, Y., Shum, H.-Y., and Salesin, D. (2004). Example-based composite sketching of human portraits. In *Proceedings of the 3rd international symposium on Non-photorealistic animation and rendering*, NPAR '04, pages 95–153. ACM.

Collomosse, J. and Hall, P. (2003). Cubist style rendering from photographs. *IEEE Transactions on Visualization and Computer Graphics*, **9**(4), 443–453.

Crick, F. and C., K. (1998). Constraints on cortical and thalamic projections: The no-strong-loop hypothesis. *Nature*, **391**, 245–250.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

Curtis, C., Anderson, S., Seims, J., Fleischer, K., and Salesin, D. (1997). Computer-generated watercolor. In *Proceedings of the 24th annual conference on Computer-Graphics and Interactive Techniques (SIGGRAPH)*, pages 421–430, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, **2**, 303–314.

Daugman, J. G. (1988). Complete discrete 2d gabor transforms by neural networks for image analysis and compression. *Transactions on Acoustics, Speech, and Signal Processing*, **36**(7), 1169–1179.

Dayan, P. and Abbott, L. (2001). *Theoretical Neuroscience*. MIT Press.

DeAngelis, G. C., Freeman, R. D., and Ohzawa, I. (1994). Length and width tuning of neurons in the cat's primary visual cortex. *Journal of Neurophysiology*, **71**, 347–374.

Debevec, P., Hawkins, T., Tchou, C., Duiker, H.-P., Sarokin, W., and Sagar, M. (2000). Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '00, pages 145–156.

DeCarlo, D. and Santella, A. (2002). Stylization and abstraction of photographs. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, volume 21, pages 769–776.

Desimone, R. and J., D. (1995). Neural mechanisms of selective visual attention. *Annual Reviews of Neuroscience*, **18**, 193–222.

Desimone, R. and Ungerleider, L. (1989). *Handbook of Neurophysiology*, chapter Neural mechanisms of visual processing in monkeys, pages 267–299. Elsevier.

Deussen, O. and Strothotte, T. (2000). Computer-generated pen-and-ink illustration of trees. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 13–18.

do Carmo, M. (1976). *Differential Geometry of Curves and Surfaces*. Prentice Hall.

Duffy, C. and Wurtz, R. (1991). Sensitivity of mst neurons to optic flow stimuli: I. a continuum of response selectivity to large-field stimuli. *Journal of Neurophysiology*, **65**(6), 1329–1345.

Durand, F., Ostromoukhov, V., Miller, M., Duranleau, F., and Dorsey, J. (2001). Decoupling strokes and high-level attributes for interactive traditional drawing. In *Proceedings of the 12th Eurographics Workshop on Rendering*.

Dwyer, F. M. J. (1967). Adapting visual illustrations for effective learning. *Harvard Educational Review*, **37**, 250–263.

Eckhorn, R., Reitboeck, H. J., Arndt, M., and Dicke, P. (1990). Feature linking via synchronization among distributed assemblies: Simulations of results from cat visual cortex. *Neural Computation*, **2**(3), 293–307.

Feldman, J. and Singh, M. (2005). Information along contours and object boundaries. *Psychological Review*, **112**, 243–252.

Fellemann, D. J. and van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, **1**, 1–47.

Field, D. H., Hayes, A., and Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local "association field". *Vision Research*, **33**(2), 173–193.

Fleming, R., Torralba, A., Dror, R. O., and Adelson, E. H. (2003). How image statistics drive shape-from-texture and shape-from-specularity. *Journal of Vision*, **3**(9), 73–73.

Fleming, R., Torralba, A., and Adelson, E. H. (2004). Specular reflections and the perception of shape. *Journal of Vision*, **4**(9), 798–820.

Fleming, R., Torralba, A., and Adelson, E. (2009). Shape from sheen. Technical report, MIT Computer Science and Artificial Intelligence Laboratory.

Fraisse, P. and Elkin, E. H. (1963). Etude génétique de l'influence des modes de présentation sur le seuil de reconnaissance d'objets familiers. *Année Psychologique*, **63**, 1–12.

Förstner, W. (1986). A feature based correspondence algorithm for image matching. *Int. Arch. Photogramm. Remote Sensing*, **26**, 176–189.

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, **36**, 193–202.

Geesaman, B. and Andersen, R. (1996). The analysis of complex motion patterns by form/cue invariant mstd neurons. *Journal of Neuroscience*, **16**, 4716–4732.

Giese, M. and Poggio, T. (2004). Neural mechanisms for the recognition of biological movements. *Nature reviews neuroscience*, **4**, 179–192.

Gilbert, C. and Wiesel, T. N. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *Journal of Neuroscience*, **9**, 2432–2442.

Goddard, N. (1992). *The perception of an articulated motion: recognizing moving light displays*. Ph.D. thesis, Department of Computer Science, Univ. of Rochester.

Gooch, A., Gooch, B., Shirley, P., and Cohen, E. (1998). A non-photorealistic lighting model for automatic technical illustration. *Computer Graphics (SIGGRAPH)*, pages 447–452.

Gooch, B., Reinhard, E., and Gooch, A. (2004). Human facial illustrations: Creation and psychophysical evaluation. *ACM Transactions on Graphics*, **23**(1), 27–44.

Goodale, M. and Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neuroscience*, **15**, 20–25.

Gove, A., Grossberg, S., and Mingolla, E. (1995). Brightness perception, illusory contours, and corticogeniculate feedback. *Visual Neuroscience*, **12**, 1027–1052.

Graziano, M., Andersen, R., and Snowden, R. (1994). Tuning of mst neurons to spiral motions. *Journal of Neuroscience*, **14**, 54–67.

Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, **1**, 17–61.

Grossberg, S. and Mingolla, E. (1985). Neural dynamics of form perception: Boundary completion, illusory figures and neon color spreading. *Psychological Review*, **92**, 173–211.

Guo, K., Robertson, R., Nevado, A., Pulgarin, M., Mahmoodi, S., and Young, M. P. (2006). Primary visual cortex neurons that contribute to resolve the apperture problem. *Neuroscience*, **138**, 1397–1406.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52.

Hamker, F. H. (2005). The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Compter Vision and Image Understanding*, **100**, 64–106.

Hansen, T. and Neumann, H. (2004). Neural mechanisms for the robust representation of junctions. *Neural Computation*, **16**, 1013–1037.

Harris, C. J. and Stephens, M. (1988). A combined corner and edge detector. *Proc. of the 4th Alvey Vision Conference*, pages 147–151.

Hartung, B. and Kersten, D. (2003). How does the perception of shape interact with the perception of shiny material? *Journal of Vision*, **3**(9), 59–59.

Hausner, A. (2001). Simulating decorative mosaics. In *Computer Graphics (SIGGRAPH)*, pages 573–580.

Hays, J. and Essa, I. (2004). Image and video based painterly animation. In *Proceedings of the 3rd international symposium on Non-photorealistic animation and rendering*, pages 113–120, New York, NY, USA. ACM Press.

Hebb, D. (1949). *The organization of behavior*. Wiley: New York.

Hedgé, J. and van Essen, D. C. (2000). Selectivity for complex shapes in primate visual cortex. *Journal of Neuroscience*, **20**, RC61.

Heeger, D. J. (1992). Normalisation of cell responses in cat striate cortex. *Visual Neuroscience*, **9**, 181–198.

Heitger, F., Heydt, R. v. d., Peterhans, E., Köbler, L. R., and O. (1998). Simulation of neural contour mechanisms: Representing anomalous contours. *Image and Vision Computing*, **16**, 407–421.

Hertz, J., Krogh, A., and Palmer, R. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley: Redwood City, California.

Hertzmann, A. (1998). Painterly rendering with curved brush strokes of multiple sizes. In *Computer Graphics (SIGGRAPH)*, pages 453–460.

Hirsch, J. A. and Gibert, C. D. (1991). Synaptic physiology of horizontal connections in the cat's visual cortex. *Journal of Neuroscience*, **11**, 1800–1809.

Horn, B. and Brooks, M. (1985). Shape and source from shading. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 932–936.

Horn, B. and Brooks, M. (1989). *Shape from Shading*. Book, MIT Press.

Hubel, D. and Wiesel, T. (1965). Receptive fields and functional architecture of two nonstriate areas (18 and 19) of the cat. *Journal of Neurophysiology*, **28**, 229–289.

Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, **195**, 215–243.

Hubel, D. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in cat's visual cortex. *Journal of Physiology*, **160**, 106–154.

Hudgkin, A. and Huxley, A. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, **117**, 500–544.

Hummel, J. and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, **99**, 480–517.

Hupé, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P., and Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by v1, v2 and v3 neurons. *Nature*, **394**, 784–787.

Hupé, J. M., James, A. C., Girard, P., Lomber, S. G., Payne, B. R., and Bullier, J. (2001). Feedback connections act on the early part of the responses in monkey visual cortex. *Journal of Neurophysiology*, **85**, 134–145.

Ikeuchi, K. (1981). Recognition of 3-d objects using the extended gaussian image. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 595–600.

Interrante, V. and Kim, S. (2001). Investigating the effect of texture orientation on shape perception. *Human Vision and Electronic Imaging*, **6**, 330–339.

Interrante, V., Kim, S., and Hagh-Shenas, H. (2002). Conveying 3d shape with texture: Recent advances and experimental findings. *Human Vision and Electronic Imaging*, **7**, 197–206.

Ito, M. and Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area v2 of macaque monkeys. *Journal of Neuroscience*, **24**(13), 3313–3324.

Izhikevich, E. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, **14**, 1569–1572.

Jellema, T. and Perrett, D. (2003a). Cells in monkey sts responsive to articulated body motion and consequent static postures: a case of implied motion? *Neuropsychologia*, **41**, 1728–1737.

Jellema, T. and Perrett, D. (2003b). Perceptual history influences neural responses to face and body postures. *Journal of Cognitive Neuroscience*, **15**, 961–971.

Jellema, T. and Perrett, D. (2007). *Neural pathways of social cognition*, chapter 13, pages 163–177. Oxford University Press.

Jähne, B., Scharr, H., and Körkel, S. (1999). *Handbook of Computer Vision and Applications*, chapter Principles of filter design. Academic Press.

Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). A biologically inspired system for action recognition. In *Proc. of the Eleventh IEEE International Conference on Computer Vision (ICCV)*.

Jin, W., Hujun, B., Weihua, Z., Qunsheng, P., and Yingqing, X. (2002). Automatic image-based pencil sketch rendering. *Journal of Computer Science and Technology*, **17**(3), 347–355.

Jones, H. E., Grieve, L. K., Wang, W., and Sillito, A. (2001). Surround suppresion in primate v1. *Journal of Neurophysiology*, **86**, 2011–2028.

Kang, H., He, W., Chui, C., and Chakraborty, U. (2005). Interactive sketch generation. *The Visual Computer*, **21**(9), 821–830.

Kanizsa, G. (1955). Margini quasi-percettivi in campi con stimolazione omogenea. *Rivista di Psicologia*, **49**, 7–30.

Kapadia, M. K., Ito, M., Gibert, C. D., and Westheimer, G. (1995). Improvement in visual sensitivity by changes in local context: Parallel studies in human observers and in v1 of alert monkeys. *Neuron*, **15**, 843–856.

Kapadia, M. K., Westheimer, G., and Gibert, C. D. (1999). Dynamics of spatial summation in primary visual cortex of alert monkeys. *Proc. Natl. Acad. Sci. USA*, **96**, 12073–12078.

Kersten, D. (1991). Transparency and the cooperative computation of scene attributes. In M. S. Landy and J. A. Movshon, editors, *Computational Models of visual processing*, pages 209–228. MIT Press, Cambridge.

Kim, S., Hagh-Shenas, H., and Interrante, V. (2003). Showing shape with texture: Two directions seem better than one. *Proceedings on Human Vision and Electronic Imaging*, **8**, 332–339.

Knierim, J. J. and van Essen, D. C. (1992). Neuronal responses to static texture patterns in area v1 of alert macaque monkey. *Journal of Neurophysiology*, **67**, 961–980.

Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, **71**, 856–867.

Koch, C. (1999). *Biophysics of Computation: Information processing in single neurons*. Oxford University Press, New York.

Koenderink, J. (1984). What does the occluding contour tell us about solid shape. *Perception*, **13**, 312–330.

Koenderink, J. and van Doorn, A. (1980). Photometric invariants related to solid shape. *Optica Acta*, **27**(7), 981–996.

Koffka, K. (1935). *Principles of gestalt psychology*. New York: Harcourt, Brace & World.

Kohonen, T. (2001). *Self-Organizing Maps*. Springer.

Kourtzi, Z., Krekelberg, B., and van Wezel, R. (2008). Linking from and motion in the primate brain. *Trends in Cognitive Sciences*, **12**, 230–236.

Krikwood, A. and Bear, M. (1994). Hebbian synapses in viusual cortex. *Journal of Neuroscience*, **14**, 1634–1645.

Lamme, V. A. F. and Rolfsemma, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, **23**, 571–579.

Lampl, I., Ferster, D., Poggio, T., and Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. *Journal of Neurophysiology*, **92**, 2704–2713.

Lange, J. and Lappe, M. (2006). A model of biological motion perception from configural from cues. *Journal of Neuroscience*, **26**, 2894–2906.

Lesher, G. W. and Mingolla, E. (1993). The role of edges and line-ends in illusory contour formation. *Vision Research*, **33**, 2253–2270.

Li, Z. (1999). Pre-attentive segmentation in the primary visual cortex. *Spatial Vision 13*, **13**, 25–50.

Little, J. and Boyd, J. (1998). Recognizing people by their gait: the shape of motion. *Videre: Journal of Computer Vision Research*, **1**, 1–32.

Litwinowicz, P. (1997). Processing images and video for an impressionist effect. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 407–414, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Logothetis, N., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex. *Current Biology*, **5**, 552–563.

Longuet-Higgins, M. S. (1960). Reflection and refraction at a random moving surface. i. pattern and paths of specular points. *Journal of the Optical Society of America*, **50**(9), 838–844.

Lorenceau, J. and Shiffrar, M. (1992). The influence of terminators in motion integration across space. *Vision Research*, **32**, 263–273.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *Journal of Computer Vision*, **60**, 91–110.

Maffai, L. and Fiorentini, A. (1976). The unresponsive regions of visual cortical receptive fields. *Vision Research*, **16**, 1131–1139.

Markram, H., Lubke, J., Frotscher, M., and Sakman, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, **275**, 213–215.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

McDermott, J. and Adelson, E. H. (2004). The geometry of the occluding contour and its effect on motion interpretation. *Journal of Vision*, **4**, 944–954.

Metelli, F. (1974). The perception of transparency. *Scientific American*, **230**(4), 90–98.

Mikami, A., Newsome, W. T., and Wurtz, R. H. (1986). Motion selectivity in macaque visual cortex. ii. spatiotemporal range of directional interactions in mt and v1. *Journal of neurophysiology*, **55**(6), 1328–39.

Milner, A. and Goodale, M. (1995). *The visual brain*. Oxford University Press.

Mishkin, M., Ungerleider, L., and Macko, K. (1983). Object vision and spatial vision: two central pathways. *Trends in Neuroscience*, **6**, 414–417.

Moody, J. and Darken, C. (1989). Fast learning in networks of locally tuned processing units. *Neural Computation*, **1**, 281–294.

Movshon, J., Adelson, E., M.S., G., and Newsome, W. (1985). *The analysis of moving visual patterns*, pages (Reprinted in Experimental Brain Research, Supplementum 11, 117–151, 1986). Vatican Press.

Mundhenk, T. N. and Itti, L. (2003). A new computational algorithm for the modeling of early visual contour integration in humans. *Neurocomputing*, **52-54**, 599–604.

Nakayama, K., He, Z. J., and Shimojo, S. (1995). Visual surface representation: a critical link between lower-level and higher-level vision. In S. Kosslyn and D. N. Osherson, editors, *An invitation to cognitive science*, volume 2, pages 1–70. MIT Press, Cambridge.

Neumann, H. and Sepp, W. (1999). Recurrent v1-v2 interactions in early boundary processing. *Biological Cybernetics*, **81**, 425–444.

Niebles, J. and Fei-Fei, L. (2007). A hierarchical model of shape and appearance for human action classification. In *Computer Vision and Pattern Recognition*, pages 1–8. IEEE Computer Society.

Nolte, J. (2002). *The Human Brain: An Introduction to Its Functional Anatomy*. Mosby Inc., 5th edition.

Norman, J., Todd, J., and Orban, G. A. (2004). Perception of three-dimensional shape from specular highlights, deformations of shading, and other types of visual information. *Psychological Science*, **15**, 565–570.

Nowlan, S. and Sejnowski, N. (1995). A selection model for motion processing in area mt of primates. *Journal of Neuroscience*, **15**, 1195–1214.

Oja, E. (1982). A simplified neuron model as a principal component analyser. *Journal of Mathematical Biology*, **15**, 267–73.

Oram, M. and Perrett, D. (1994a). Modeling visual recognition from neurobiological constraints. *Neural Networks*, **7**, 945–972.

Oram, M. and Perrett, D. (1994b). Responses of anterior superior temporal polysensory (stpa) neurons to 'biological motion' stimuli. *Journal of Cognitive Neuroscience*, **6**, 99–116.

Oram, M. and Perrett, D. (1996). Integration of form and motion in the anterior superior temporal polysensory area (stpa) of the macaque monkey. *Journal of Neurophysiology*, **76**, 109–129.

Oren, M. and Nayar, S. (1996). A theory of specular surface geometry. *International Journal of Computer Vision*, **24**, 105–124.

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.

Pack, C. C., Livingston, M. S., Duffy, K. R., and Born, R. T. (2003). End-stopping and the aperture problem: Two-dimensional motion signals in macaque v1. *Neuron*, **39**, 671–680.

Pasupathy, A. and Conner, C. E. (1999). Response to contour features in macaque area v4. *Journal of Neurophysiology*, **82**(5), 2490–2502.

Peelen, M. and Downing, P. (2008). The neural basis of visual body perception. *Nature Reviews*, **8**, 636–648.

Perrett, D. and Oram, M. (1993). Neurophysiology of shape processing. *Imaging and Vision Computing*, **11**, 317–333.

Perrett, D., Smith, P., Mistlin, A., Chitty, A., Head, A., Potter, D., Broennimann, R., Milner, A., and Jeeves, M. (1985a). Visual analysis of body movements by neurones in the temporal cortex in the macaque monkey: a preliminary report. *Behavioural Brain Research*, **16**, 153–170.

Perrett, D., Smith, P., D.D., P., Mistlin, A., Head, A., Millner, A., and Jeeves, M. (1985b). Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proc. Royal. Society London B*, **223**, 293–317.

Perrett, D., Harries, M., Bevan, R., Thomas, S., Benson, P., Mistlin, A., Chitty, A., Hietanen, J., and Ortega, J. (1989). Framework of the analysis for the neural representation of animat objects and actions. *Journal of Experimental Biology*, **146**, 87–113.

Peterhans, E. (1997). Functional organization of area v2 in awake monkey. *Cerebral Cortex*, **12**, 335–357.

Peterhans, E. and Heydt, R. (1989). Mechanisms of contour perception in monkey visual cortex ii. *Journal of Neuroscience*, **9**, 1749–1763.

Poggio, T. and Edelman, S. (1990). A network that learns to recognize 3d objects. *Nature*, **343**, 263–266.

Ponce, C. R. and Born, R. T. (2008). Stereopsis. *Current Biology*, **18**, R845–R850.

Raudis, F. and Neumann, H. (2010). A model of neural mechanisms in monocular transparent motion perception. *Journal of Physiology Paris*, **104**, 71–83.

Reichardt, W. (1987). Evaluation of optical information by movement detectors. *Journal of Computational Physiology A*, **161**, 533–547.

Rhodes, G., Brennan, S., and Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology*, **19**, 473–497.

Rieke, F., Warland, D., de Ruyter van Stevenick, R., and Bialek, W. (1996). *Spikes - Exploring the neural code*. MIT Press.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, **2**, 1019–1025.

Riesenhuber, M. and Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, **12**, 162–168.

Rosenblum, M., Yacoob, Y., and Davis, L. (1996). Human expression recognition from motion using radial basis function network architecture. *IEEE Transactions on Neural Networks*, **7**, 1121–1138.

Ross, W. D. (2000). Visual cortical mechanisms of perceptual grouping: Interacting layers, networks, coolumns, maps. *Neural Networks*, **13**, 571–588.

Rubin, N. (2001a). Figure and ground in the brain. *Nature Neuroscience*, **4**(9), 857–858.

Rubin, N. (2001b). The role of junctions in surface completion and contour matching. *Perception*, **30**, 339–366.

Ryan, T. and Schwartz, C. (1956). Speed of perception as a function of mode of presentation. *American Journal of Psychology*, **69**, 60–69.

Saito, H. (1993). *Brian Mechanisms of Perception and Memory*. Oxford University Press.

Salin, P. A. and Bullier, J. (1995). Corticocortical connections in the visual system: Structure and function. *Physiological Reviews*, **75**, 107–154.

Samaras, D. and Metaxas, D. (1999). Coupled lighting direction and shape estimation from single images. In *Proceedings of the International Conference on Computer Vision*, volume 2, page 868, Washington, DC, USA. IEEE Computer Society.

Sanderson, A., Weiss, L., and Nayar, S. (1988). Structured highlight inspection of specular surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **10**(1), 44–55.

Sato, T. (1989). Interactions of visual stimuli in the receptive fields of inferiortemporal neurons in awake monkeys. *Experimental Brain Research*, **77**, 23–30.

Savarese, S. and Perona, P. (2001). Local analysis for 3d reconstruction of specular surfaces - part i. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 738–745.

Savarese, S. and Perona, P. (2002). Local analysis for 3d reconstruction of specular surfaces - part ii. In *Proceedings of 7th European Conference on Computer Vision*, pages 759–774.

Savarese, S., Fei-Fei, L., and Perona, P. (2004). What do reflections tell us about the shape of a mirror? *Proceedings on Applied Perception in Graphics and Visualization*, **1**, 115–118.

Sceniak, M. P., Hawken, M. J., and Shapley, R. (2001). Visual spatial characterization of macaque v1 neurons. *Journal of Neurophysiology*, **85**, 1873–1887.

Schindler, K. and van Gool, L. (2008). Action snippets: How many frames does human action recognition require? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8. IEEE Press.

Schmid, C., Mohr, R., and Bauckhage, C. (2000). Evaluation of interest point detectors. *International journal of computer vision*, **37**(2), 151–172.

Schmidt, K. E., Goebel, R., Löwel, S., and Singer, W. (1997). The perceptual grouping criterion of colinearity is reflected by anisotropies of connections in the primary visual cortex. *European Journal of Computational Neuroscience*, **9**, 1083–1089.

Shipley, T. F. and Kellman, P. J. (1990). The role of discontinuities in the perception of subjective figures. *Perception and Psychophysics*, **48**(3), 259–270.

Shiraishi, M. and Yamaguchi, Y. (2000). An algorithm for automatic painterly rendering based on local source image approximation. In *Proceedings of the 1st international symposium on Non-photorealistic animation and rendering*, pages 53–58, New York, NY, USA. ACM Press.

Sillito, A., K.L., G., Jones, H. E., Cudeiro, J., and Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, **378**, 492–496.

Sillito, A., Cudeiro, J., and Jones, H. (2006). Always returning: feedback and sensory processing in visual cortex and thalamus. *Trends in cognitive sciences*, **29**(6), 307–316.

Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, **24**, 49–65.

Smith, A. and Snowden, R. (1994). *Visual detection of motion*. Academic Press.

Smith, S. and Brady, J. (1997). Susan: A new approach to low level image processing. *International journal of computer vision*, **23**(1), 45–78.

Sperling, G. (1970). Model of visual adaptation and contrast detection. *Perception and Psychophysic*, **8**, 143–157.

Stalling, D. and Hege, H.-C. (1995). Fast and resolution independent line integral convolution. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 249–256.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Reviews Neuroscience*, **19**, 109–139.

Tanaka, K., Hikosaka, K., Saito, M., Yukie, M., Fukada, Y., and Iwai, E. (1986). Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey. *Jourjal of Neuroscience*, **6**(1), 134–144.

Tanaka, Y., Koyama, T., and Mikami, A. (1999). Neurons in the temporal cortex changed their preferred direction of motion dependent on shape. *Neuroreport*, **10**, 393–397.

Thielscher, A. and Neumann, H. (2005). Neural mechanisms of texture processing: Texture boundary detection and visual search. *Spatial Vision*, **18**, 227–257.

Thielscher, A. and Neumann, H. (2008). Globally consistent depth sorting of overlapping 2d surfaces in a model using local recurrent interactions. *Biological Cybernetics*, **98**(4), 305–337.

Thorpe, S. (1990). *Parallel processing in neural systems and computers*, chapter Spike arrival times: A highly efficient coding scheme for neural networks, pages 91–94. Elsevier.

Tinbergen, N. (1951). *The Study of Instinct*. Oxford University Press, Clarendon, Oxford.

Tinbergen, N. and Perdeck, A. C. (1950). On the stimulus situation releasing the begging response in the newly hatched herring gull chick (larus argentatus argentatus pont). *Behaviour*, **3**, 1–39.

Todd, J. and Mingolla, E. (1983). Perception of surface curvature and direction of illumination from patterns of shading. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 740–745.

Todd, J., Norman, J., Koenderink, J., and Kappers, A. (1997). Effects of texture, illumination and surface reflectance on stereoscopic shape perception. *Perception*, **26**, 806–822.

Todd, J. T. (2004). The visual perception of 3d shape. *Trends in cognitive science*, **8**(3), 115–121.

Todd, J. T., Norman, J. F., and Mingolla, E. (2004). Lightness constancy in the presence of specular highlights. *Psychological Science*, **15**(1), 33–39.

Tolhurst, D. J. and Heeger, D. J. (1997). Comparison of contrast-normalization and threshold models of the responses of simple cells in cat striate cortex. *Visual Neuroscience*, **14**, 293–309.

Ts'o, D. Y., Gibert, C. D., and Wiesel, T. N. (1986). Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. *Journal of Neuroscience*, **6**, 1160–1170.

Ungerleider, L. and Mishkin, M. (1982). *Analysis of visual behavior*, chapter Two cortical visual systems, pages 549–586. MIT Press.

Varin, D. (1971). Fenomeni di contrasto e diffusione cromatica nell' organisatione spaziale del campo percetiivo. *Rivista di Psicologia*, **65**, 101–128.

von der Heydt, R., Peterhans, E., and Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, **224**, 1260–1262.

Wallach, H. (1935). Ueber visuell wahrgenommene bewegungsrichtung. *Psychologische Forschung*, **20**, 325–380.

Watanabe, T. and Cavanagh, P. (1993). Transparent surfaces defined by implicit x junctions. *Vision Research*, **33**, 2339–2346.

Weidenbacher, U. and Neumann, H. (2009). Extraction of surface-related features in a recurrent model of v1-v2 interactions. *PloS One*, **4**(6).

Weidenbacher, U., Bayerl, P., Neumann, H., and Flemming, R. W. (2006). Sketching shiny surfaces: 3d shape extraction and depiction of specular surfaces. *ACM Tansactions on Apllied Perception*, **3**(3), 262–285.

Williams, D. and Phillips, G. (1987). Cooperative phenomena in the perception of motion direction. *Optical Societey of America A*, **4**, 878–885.

Wolfe, J. M., Randall, S. B., Kunar, A. M., and Horowitz, T. S. (2005). Visual search for transparency and opacity: Attentional guidance by cue combination? *Journal of Vison*, **5**, 257–274.

Yamamoto, S., Mao, X., Tanii, K., and Imamiya, A. (2004). Enhanced lic pencil filter. In *Proceedings International Conference on Computer Graphics, Imaging and Visualization, 2004. CGIV 2004.*, pages 251–256.

Yu, A., Giese, M., and Poggio, T. A. (2002). Biophysiologically plausible implementations of the maximum operation. *Neural Computation*, **14**(12), 2857–2881.

Zhang, R., Tsai, P.-S., Cryer, J., and Shah, M. (1999). Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**(8), 690–706.

Zhaoping, L. (2005). Border ownership from intracortical interactions in visual area v2. *Neuron*, **47**, 143–153.

Zheng, J. and Murata, A. (2000). Acquiring a complete 3d model from specular motion under the illumination of circular-shaped light sources. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 913–920.

Zheng, Q. and Chellapa, R. (1991). Estimation of illuminant direction, albedo, and shape from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, **13**(7), 680–702.

Zhou, H., Friedman, H. S., and von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *Journal of Neuroscience*, **20**(17), 6594–6611.

Zisserman, A., Giblin, P., and Blake, A. (1989). The information available to a moving observer from specularities. *Image and Vision Computing*, **7**(1), 38–42.

# Zusammenfassung

In dieser Arbeit wird ein neuronales Modell für die visuelle Wahrnehmung von Flächen, Objekten und Personen vorgestellt. Weiter werden Ergebnisse des vorgestellten neuronalen Modells verglichen mit Bildverarbeitungsverfahren des maschinellen Sehens.

Das visuelle System des Menschen segmentiert 3D Szenen in Flächen und Objekte, welche in verschiedenen Abständen vom Beobachter vorkommen können. Da bei der Aufnahme ein Ausschnitt der 3D Welt auf die Bildebene projiziert wird, entstehen in vielen Fällen Verdeckungen von Flächen und Objekten. Es gibt experimentelle Hinweise darauf, dass flächenbasierte Merkmale (Begrenzungskonturen, Kreuzungspunkte) als Indikatoren für die robuste Segmentierung von Flächen verwendet werden. Diese Flächenmerkmale zeichnen sich durch ihre Robustheit gegenüber der Umgebungsbeleuchtung aus und sind deshalb auch für das Lernen von komplexeren Formen wie Gesichtern und deren unterschiedliche Ansichten geeignet.

Es wird ein biologisch inspiriertes, rekurrentes Modell vorgestellt, welches flächenbasierte Merkmale aus einem 2D Grauwertbild extrahiert und geeignet zur Repräsentation von Flächen interpretiert. In Anlehnung an die Neurophysiologie des Gehirns von Primaten basiert das Modell auf wenigen Basismechanismen, welche in jeder Schicht unterschiedlich parametrisiert zur Anwendung kommen. Die Architektur zeichnet sich außerdem auch durch Feedback-Verbindungen aus, welche zu einer zeitlichen Dynamik der internen Aktivitäten führen. Unter Verwendung dieser Vorverarbeitungsschritte wird ein mehrschichtiges Lernverfahren zur formbasierten Wahrnehmung der Kopfbewegung vorgestellt, welches typische Kopfbewegungsmuster bei der visuellen Kommunikation (z.B. Zuwendung / Abwendung) repräsentieren und detektieren kann.