Universität Ulm | 89069 Ulm | Germany

Fakultät für
Ingenieurwissenschaften
und Informatik
Institut für Neuroinformatik
Direktor: Prof. Dr. Günther Palm

Contextual influences in hierarchical motion and form processing - a modeling study in bio-mimetic architecture

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

vorgelegt von
Stephan Johannes Tschechne
geb. 20.10.1978 in Dachau

Amtierende Dekanin der Fakultät für Ingenieurwissenschaften und Informatik:
Prof. Dr. Tina Seufert

Gutachter: Prof. Dr. Heiko Neumann
Gutachter: Prof. Dr. Günther Palm

Tag der Promotion: 24. März 2016

Abstract

Processing of visual sensory input is mainly determined by filtering mechanisms to detect the presence of certain features and generate proper representations. Such features can be derived over several stages of hierarchical processing to extract increasingly more complex feature compositions characteristic of the input signal. However, context information often determines the interpretation of localized feature responses, such that, e.g., an individual item presented as part of a texture appears different from a presentation in isolation or in contrast to surrounding items. In this thesis, such contextual influences in the visual processing of form and motion information are presented and their generation is explained within a model neural architecture. The thesis details how such context information changes the gain of feature representation by up- and down-modulating the activation in distributed feature representations. The modeling investigates a bio-mimetic architecture of event-based sensor input from a *Dynamic Vision Sensor (DVS)*. It is demonstrated that local filtering generates initial motion representations where competitive mechanisms of pool normalization and modulating feedback redistribute local responses to emphasize moving features, reducing the uncertainty for the motion aperture effect and improving the signal to noise ratio of sensory input data. The work shows how benefit is taken from the high temporal resolution of the sensor and how a representation of motion streaks is achieved along the ventral pathway. In a hierarchical model of form processing we present mechanisms of perceptual organization that aggregate localized items into a meaningful representation of static scene content. With the use of complex hierarchical feedforward and feedback connections, contextual information about a contour is made available to lower regions. It is shown how labeling of figure-ground direction of boundaries is achieved by tagging using activities of competitive interactions that are biased by top-down enhancement from higher stages.

Contents

1 Introduction

We visually perceive our environment with ease, precision and without any noticeable delay. Without any apparent effort, objects and persons are recognized, physical structures are precisely perceived in three dimensions, scratchy handwriting deciphered or navigation tasks performed even in the presence of a large amount of occlusion, disruptive or incomplete information.

Developers of hard- and software struggle to achieve any comparable performance on an artificial system. In order to make computer vision applications useful, they have to be designed for a very limited, usually low-level task and grant little variation to the input. Small deviations from the specification usually massively deteriorate the result. This contrasts human performance on such problems, which is usually used as the benchmark for visual performance.

Over half a century, anatomical and physiological investigations, psychological experiments and improvements in image generating methods have increased our understanding of the functional principles of the brain of our and closely related species, and the body of knowledge is still rapidly growing. Current evidence strongly supports the belief that the brain is organized using a modular processing paradigm, which makes analysis of the participating components valuable and offers a suggestion about how to achieve a similar functionality and robustness in an artificial system.

Neural information processing adapts such mechanisms to perform them on artificial systems. On one hand, this leads to better algorithms and products by incorporating biologically inspired processing modules. Ideally, following the biological example leads to technical solutions that have an increased performance, higher robustness and tolerance towards input variations and better generalization to other problem domains. On the other hand, precise models of cortical mechanisms allow to draw conclusions and build hypotheses which by themselves contribute to the understanding we have about the involved mechanisms.

However, the hope that a precise analysis of participating components and a straight-forward connection in an *input-output* scheme would eventually lead to similar performance is quickly lost when the connecting mechanisms are considered as well. Neural regions respond when adequate features are presented to specific regions of the visual field (the receptive field), and the relation in the spatial and feature domain can be accomplished with a set of techniques (and usually a lot of patience). But, alas, this is only the beginning. Such isolated features virtually never occur in a realistic environment, where they are surrounded by many others. The crux of the matter is that neural responses are highly influenced by such surrounding features, even if they are completely outside the initially found receptive fields. This *contextual influence* has the potential to determine the interpretation of localized feature responses by means of re-modulation and by this to increase the robustness towards input variations and to enrich local representations to resolve ambiguities or to support responses that are near perceptual thresholds. The search for and modeling of these influencing mechanisms is very challenging given the vast amount of connections that are already known to exist between processing areas.

This thesis presents models of biologically inspired mechanisms of visual motion and form processing which make use of contextual information to reweigh the evidence of feature representation. The models build upon generic mechanisms known to exist in visual cortex and show how contextual influences can be modeled over different hierarchical processing stages to enrich initial feature representations. The thesis is organized into three main chapters:

- Chapter 2 (page 5) introduces the reader to the physiological background of vision including a short summary of the optical tract from the retina to the occipital cortex (Sec. 2.2). More highlight is put on the cortical mechanisms of visual processing and the neural representation of visual stimuli (Sec. 2.3). While keeping the mechanisms of single-neuron-models in mind, a generic model of cortical visual processing is presented that presents canonical mechanisms of whole processing areas (Sec. 2.3.2). It has been used in previous models of visual processing and contains three major stages of processing for a model visual area, namely a stage of initial filtering, a stage of contextual feedback modulation and a stage of non-linear filter operations by means of normalization mechanisms. The subsequent contributions in this thesis are based on these generic components.

- Chapter 3 (page 17) proposes a model of motion processing along the dorsal pathway using an event-based retina-like image sensor. The proposed model applies physiological findings for the generation of a model of optic flow estimation using the output of the sensor. Our model makes use of neural mechanisms that model the processing capabilities of early and intermediate stages in dorsal pathway of visual cortex (Sec. 3.4.1). The chapter elaborates on subsequent processing mechanisms that pick up physiological and psychophysical findings along the dorsal pathway. We present mechanisms of early surround inhibition that show to contribute to the solution of the aperture problem (Sec. 3.4.2). Motion integration and its effects on increasing the robustness of initial motion representation to noise and outliers is presented (Sec. 3.4.3). We also present how a spatial code of motion estimation is incorporated using the effect of motion streaks or speed lines. In this context the interaction of the dorsal and ventral stream for motion processing is shown (Sec. 3.5). The chapter closes with considerations of how a motion algorithm based on an event-based representation of visual input is designed in order to keep the unique processing advantages like instantaneous estimation of local estimates and sparse representation, where this algorithm is evaluated with respect to its computational complexity (Sec. 3.6).

- Chapter 4 (page 77) proposes a model of mechanisms in ventral pathway that segments a visual scenes into image regions and combines the initial responses into representations of surfaces and prototypical objects. Multiple mutually connected areas in the ventral cortical pathway receive visual input and extract local form features that are subsequently grouped into increasingly complex, more meaningful image elements. We propose a mechanism how such a distributed network of processing is capable of representing highly articulated changes in shape boundary as well as subtle curvature changes. We propose a recurrent computational network architecture that utilizes hierarchical distributed representations of shape features to encode surface and object boundaries over different scales of resolution. Our model makes use of neural mechanisms that model the processing capabilities of early and intermediate stages in visual cortex, namely areas V1-V4 and IT (Sec. 4.3.1). We suggest that multiple specialized component representations interact by feedforward hierarchical processing (Sec. 4.3.2) that is combined with feedback signals driven by representations generated at higher stages (Sec. 4.3.3). Based on this, global configurational as well as local information is made available to distinguish changes in the object's contour. Once the outline of a shape has been established, contextual contour configurations are used to assign border ownership directions and thus achieve segregation of figure and ground. The model, thus, proposes how separate mechanisms contribute to distributed hierarchical cortical shape representation and combine with processes of figure-ground segregation. Our model is probed with a selection of stimuli to illustrate processing results at different processing stages. We especially highlight how modulatory feedback connections contribute to the processing of visual input at various stages in the processing hierarchy.

2 Physiology and processing models of vision

Our rich eye-minded perception of the world is achieved by complex neural processes in our visual system. But how are these comprehensive representations exactly reaped from the patterns of light that enter our eyes? An understanding of this clearly requires a profound analysis of all the participating mechanisms. Sadly, our brain neither comes with a detailed system specification nor is it possible to consult someone that participated during its development. The only way to achieve knowledge about the underlying mechanisms is to reverse-engineer from what is observable with the tools that we have at our hands.

An essential tool in this endeavor is the methodology of building models of systems as abstractions of the real world system in order to investigate particular questions by testing or demonstrating hypotheses [Trappenberg, 2002]. Such models help to find the essential components that contribute to an observable effect. This reveals insights into the relevant mechanisms, their parameters and interactions and usually leads to more questions that require in-depth investigations. Over time, this iteration leads to a deep and precise understanding of previously unknown things.

This thesis is going to present models of visual processing in Chap. 3 and 4, but before, this part will quickly recap methods available for investigations on the brain and give a brief introduction to the optical pathway from the retina to the visual cortex. It will then show general principles of how to model neural activations in different levels of detail. This knowledge will be the foundation of the subsequent models.

2.1 Observation methods

The knowledge that the brain is the organ that provides us with cognitive capabilities was already known in ancient Greece. The largest part of today's knowledge about the anatomical structure was collected in the 19th century by systematic research on animals and pathological studies (like autopsies) on humans. The *neuron* was discovered as the basic element of cortical processes and the systematic research on revealed the first pioneering insights into their anatomy and functional properties[Hubel and Wiesel, 1959].

Brains consists of billions of neurons that show a tremendous degree of interconnectivity. Each neuron keeps contact with up to 10,000 other neurons in close proximity, but sometimes they also maintain long-range connections. The precise architecture of these connections is not static, but is subject to continuous modifications. An individual neuron is a type of cell that is specialized for transmitting information by means of electrical or chemical signals. Such signals are called *action potentials* and travel between neurons through their synapses. A neuron consists of a *soma* (the cell body), a number of *dendrites* (the incoming connections) and with very few exceptions precisely one *axon* (the outgoing connection). In a nutshell, each neuron acts as a signaling unit listening for incoming action potentials. When enough of them arrive through the cell's dendrites, it by itself generates an action potential that travels along the axons towards dendrites of other neurons. This can cause an avalanche of potentials among a huge set of neurons. What we experience as our ability to understand things that we see, our knowledge, our instincts and everything is coded in the activation and interaction of patterns of neurons.

This leads to the first question in neural modeling: What level of detail is adequate to analyze and model neural processes? The current knowledge allows to describe neurons and their mechanisms on a large range of different levels of detail. For an in-length discussion of different levels and the implications of each of them the reader may be pointed to [Churchland et al., 1990]. In a nutshell, neural principles can be investigated on scales spanning roughly ten orders of magnitude. This includes an inspection of processes on a molecular level in a metric scale of about one $\mathring{A}$ngstrom ($10^{-10}m$), a focus on single synapses or complete neurons using a resolution of 10^{-9} to $10^{-7}m$, or the analysis of complete networks and maps of neurons on a scale of a few milli- or centimeters. A complete nervous system might as well have a size of one meter. A high level of detail may be precise, but usually limits the number of instances that are possible to model on this level to a few neurons or small networks. A higher-order functionality is usually provided only by arrangements of at least of few thousand neurons and many connections, so high-detailed models quickly make investigations difficult if not intractable. A good model works on a scale that captures the essential principles but abstracts or simplifies everything that does not contribute in a significant way.

In the middle of the 20th century the list of invasive and non-invasive investigation methods was extended and since then, more fascinating insights in the working, living brain became possible. These techniques include electro-encephalography (EEG) that measures electric brain potentials or magneto-encephalography (MEG) which measures short-term changes in magnetization. Trans-cranial magnetic stimulations (TCMS) locally eliminates activations through short bursts of magnetic impulses and helps to reveal contributions of certain brain areas to behavioral tasks. Psychophysical experiments allow the investigation of perceptual mechanisms under controlled and reproducable laboratory conditions. Here, learning performance, decision times, reactions times and more can be measured and analyzed by means of statistical analysis. Among the most spectacular methods for making the activity of a behaving brain visible is positron-emission tomography (PET) that measure the concentration of a substrate that was marked with a radioactive tracer. Areas with high cortical activation require more energy provided by the blood through glucose and oxygen and this is reflected in a locally increased concentration of the tracer. Functional magnetic resonance tomographs (fMRI) makes use of the magnetic properties of oxygen-rich and oxygen-poor blood by three-dimensionaly measuring the blood-oxygen-level dependent (BOLD) contrast with high spatial resolution. Such scans generate colorful images of brain activations with a spatial resolution of a few millimeters and a temporal resolution of about one second. These parameters slightly change with the size of the scanned region. Whole brain scans usually provide less resolution than the focus of smaller regions.

The chapter will now introduce models of visual processing. It starts by summarizing the optical tract which takes the reader from the eyes to the brain and then it continues by focusing on single neuron models used for models of the visual cortex.

2.2 Optical pathway

The visual systems extracts a vocabulary of relevant features related to a visual scene which lead to meaningful perceptions. The visual processing is performed along the visual path that starts with photo-receptive cells on the retina and continues in a complex, recurrent, hierarchical architecture of visual areas in occipital cortex and other areas.

The retina has a three-layered structure of photo-receptive cone and rod cells that transform incoming light into nervous impulses. The first processing of those impulses for light adaptation and contrast increase already happens at the second layer that consists of horizontal cells, bipolar cells and amakrine cells. Ganglion cells transport processed information to the next step outside the eye. It may be noted here that the density of light sensitive cells is not evenly distributed across the retina. A small visual area, the *fovea centralis*, contains about eighty percent of the total number of cells. This has an effect to perceivable resolution, that drops significantly with distance to the *fovea centralis*. It forces the visual system to continuously move the eye in a way such that the region of interest projects onto the area of highest density. About 15° degrees horizontally in nasal direction the axons of ganglion cells leave the eye at the *papilla*, where no photoreceptive cells are available at all and form an effective *blind spot*. Many, including our, visual models do not include these deficits but assume homogeneous resolution.

The optic nerve is the bundle of axons leaving at the papilla. It contains about one million nerves. At *chiasma opticum* the nerves from left and right eye cross and connections are to a small degree redistributed. This early interconnection has a beneficiary effect on depth perception. The optic nerves continue through the *lateral geniculate nucleus*, which performs the first neural processing circuitry outside the retina. The axons continue from here along the optic radiation to the *primary visual cortex* (or Brodman area 17) at the occipital cortex. The occipital cortex (*lobus occipitalis*) is the back part of the brain (*cerebrum*) and the smallest of four brain lobes. It is part of the visual system and contains neural maps that process the visual impulses and thus called *visual centre* of the brain.

Each occipital lobe processes the visual impulses of the temporal ipsilateral and the nasal contralateral retina, meaning that the respective left parts of the retina are processed in the left occipital lobe, and both right parts of the retina are processing in the right occipital lobe. Each point of the retina projects to a small area in visual cortex. A peculiar feature of the visual area is an extra nerve bundle called *Gennari*- or *Vicq-d'Azyr*-stripes. It is noticeable in a macroscopic scale and therefore the region is also called *striped* region (*area striata*) or striped cortex (*striate cortex*).

Individual neurons in neocortex, where the occipital cortex is counted to, are organized into areas which consist of layered structure of neurons. The six layered-structure (counted from the outside) can be characterized by the amount cell type clustering, lateral connections and input output connections [Noback et al., 2005, Kurzweil, 2012]. Neurons in the upper layers II and III project within other areas of neocortex. Neuron in the deeper layers V and VI connect to areas out of the cortex like the thalamus, brainstem and spinal cord. Neuron from outside the cortex enter at layer IV and this layer distributes it to the other layers for further processing. Sec. 2.3.2 will show how certain aspects of this layering are used to build a canonical model of cortical processing.

With the amount of photo-receptive cells varying across the visual field, so is the size of dedicated cortical areas. The fovea centralis, where the visual resolution is highest, claims about 80 percent of the visual area. The spatial relationships on the retina are preserved at in the early visual areas. Cell afferents that stem from juxtaposed region on the retina are juxtaposed in visual cortex, effectively realizing a *retinotopic* arrangement. The projected size of inputs that a cell assembly receives input from is called *receptive field* (RF). Visual areas are tuned to specific stimuli and evoke an activation whenever such a stimulus is present in their receptive field.

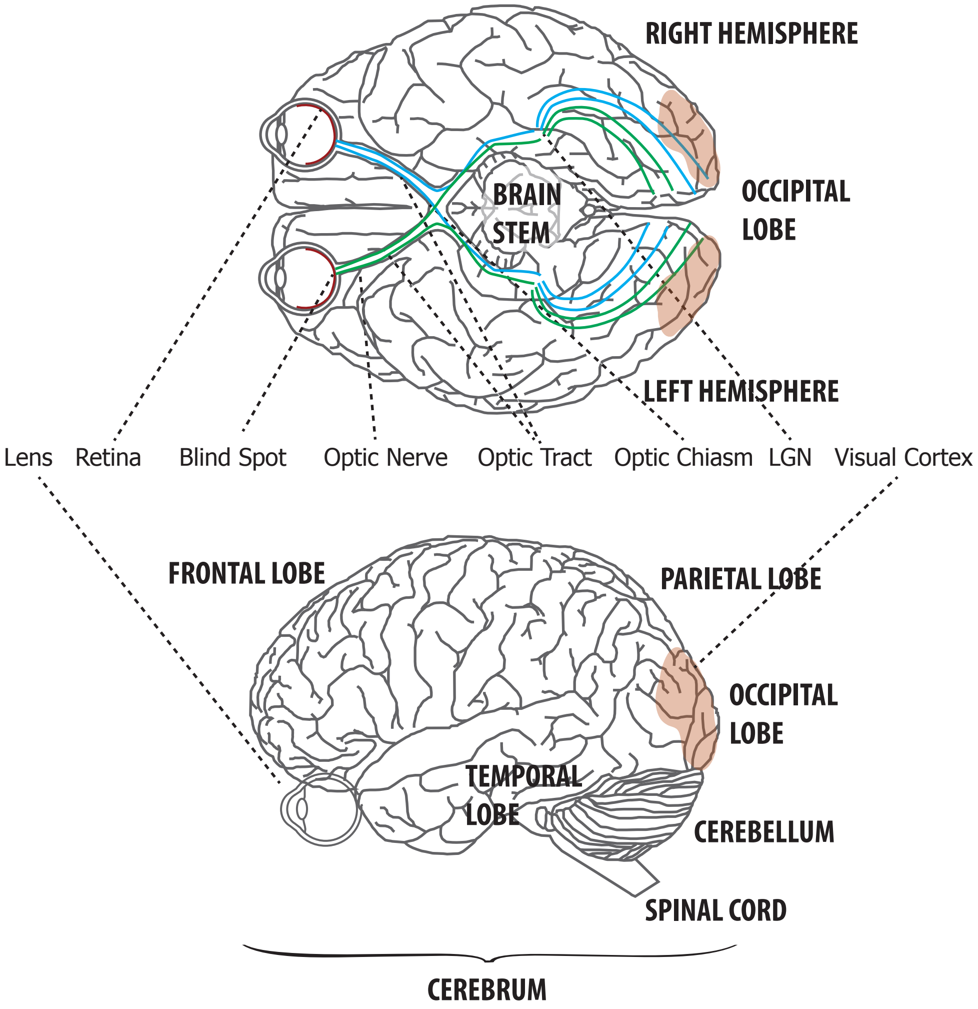


Figure 2.1: The human brain (*cerebrum*) viewed from the top and from the side with selected brain areas labeled. *Light enters the eye from the left side and evokes neural activations on the retina, which travel along the optical tract, pass through the* optic chiasm *and* LGN *and arrive at the occipital cortex, the visual processing center of the brain.*

Within the recent decades and with help of neuroimaging procedures like fMRI neurophysiologists have consolidated the opinion that the visual system of primates is an organization of many different interconnected visual areas [Ungerleider and Haxby, 1994] with have specialized functionalities. More than 30 visual areas are currently identified [Felleman and Van Essen, 1991] and it is believed that sixty percent of the cortex is dedicated to the processing and perception of visual input. That processing hierarchy can be divided into two streams that specialize in different aspects of processing. While the *dorsal stream* focuses on the processing of object locations and motion, the *ventral stream* determines the identity of objects [Ungerleider and Haxby, 1994, Mishkin et al., 1983, Felleman and Van Essen, 1991, Van Essen and Gallant, 1994]. Along these streams, the receptive fields size change from relatively small retinal areas of about one degree to larger and larger retinal areas, up to receptive fields that cover almost the complete visual field. With the increase of receptive field size increases the complexity of stimuli that yield cortical activation. Areas of early visual processing are activated by simple stimuli like contrasts or local motion into one direction. The tunings become more and more specialized and range from oriented structures [Hubel and Wiesel, 1959], to parts of objects, to objects to representations of complete scenes or sequences. The codings in dorsal stream range from initial motion detection over an integrated signal to motion patterns and to more complex motion patterns [Felleman and Van Essen, 1987].

The approaches to explain observations by means of forward-directed mechanisms like filtering and integration reach an impasse in the light of various non-linear effects of neural responses. The amount of neural activation does not linearly increase with the strength of the presented stimulus, but reach a saturation point[Heeger, 1992, Carandini and Heeger, 1994]. These normalization effects as a canonical principle do not only appear in visual models, but also in auditory processing, attention and decision making and have a positive effect on the ability to perceive contrasts and stimuli containing a high dynamic range. Our models incorporate a normalization principle in inhibitory form using lateral connections, which is elaborated in the upcoming sections. The second option of how information from outside the receptive field may influence neural activation is via downward-directed recurrent connections. Here, the source of modulation lies outside the same brain area and may represent features of another class and complexity. By this, representations from higher areas reach down and contribute to the activation found there. The top-down signals add a predictive component to feature representations and provide competitive advantage to matching activations in the modeled area[Brosch and Neumann, 2014a]. Of specific interest for vision research like in this thesis is how such lateral and backward oriented connections make contextual information available to local processing[Albright, 1984].

2.3 Models of visual neural information processing

The principles of individual neurons can be described at different levels of abstraction. Over time, five main levels have been identified to be useful.

Detailed **compartmental models** provide a high level of detail. Here, the complete dendritic tree is modeled using individual membrane and synapse conductances. Those models would allow to e.g. precisely model timings of post-synaptic propagation of potentials in dendritic integration [Herz et al., 2006]. This high level of detail is throttled in **reduced compartmental models**, where only one or few dendrite compartments are represented. The investigations in this thesis will stick to **single-compartment models** that neglect the neuron's structure and regard it as a point-like process. There exist several models of how action potentials are generated on this level, like the Izhikevich [Izhikevich, 2007], the Fitzhugh-Nagumo model [Fitzhugh, 1969, Nagumo et al., 1962], of which the most elaborated is the Hodgkin-Huxley model [Hodgkin and Huxley, 1990]. This work will be based on a systematic reduction of the mathematical descriptions necessary to describe neural activations on this level. The second important type of single-neuron models are **cascade models**, where mathematical primitives are concatenated to model the computational properties. This includes linear filters, non-linearities and random processes. These types of models have a long tradition in the investigation of the visual system. Section 2.3.2 describes the canonical model for neurons and how this internal activation is transformed to a firing rate. Further reduction of detail is achieved with **black-box models**, where the biophysical machinery is completely ignored and the signal processing capabilities are modeled by whatever might be adequate. An activation of a neuron might in this type of model as well be represented by a stochastic process.

2.3.1 Single-compartment models and membrane equation

Throughout the thesis, single-compartment models are the tool of choice for our investigations [Koch, 1998, Herz et al., 2006]. These models are motivated by electrophysiological and anatomical studies performed mostly on the brains of macaque monkeys [Van Essen and Gallant, 1994] and focus on the mechanisms of ionic currents below thresholds and the spike generation. They have helped to quantitatively understand many dynamic phenomena. In this type of model, a neuron is viewed as a point-like process, without a widespread dendritic structure. In reality an active neuron maintains a voltage gradient across its membrane to its environment by producing a different concentration of ions at the inside of the cell with respect to the outside. This causes a continuous equalization effect by

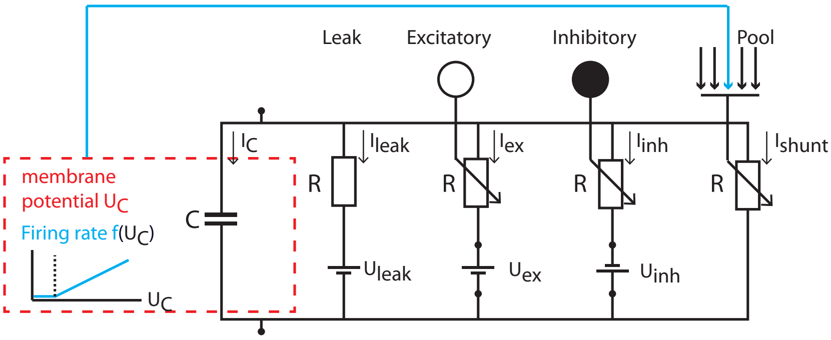


Figure 2.2: Simplified neuron model in form of a single-compartment model. *The cell's potential is represented as a capacitor. Leakage of potential as well as influences of excitatory, inhibitory and surround connections area modeled with basic electric elements. The dynamic properties of this circuit are captured in the membrane equation.*

means of a small current through the cell's membrane. The cell compensates that current with ion channels and continuously operating ion pumps. Large changes of cell voltage elicits the production of an action potential, that quickly travels along the cell's axon and elicits afferent connections at other cells.

The electrical functionality of such a simplified neuron can be modeled by a circuit containing some basic electrical components. The membrane acts as a capacitor C that separates the inside from the outside of the cell, effectively holding an electric potential. It also works as a resistor R that allows a leaking current to flow as a function of voltage, excitatory and inhibitory connections. This leakage is neutralized with ion pumps. When no input is available, currents balance each other in a dynamic equilibrium. The application of Kirchhoff's laws describe the specific dynamics of such a circuit.

This work incorporates a dynamic model of cell assemblies that represents an average spike rate code for the simulation of neural activation. We introduce the canonical form of the dynamic equations using a dynamic membrane model. We model the potential p of the cell as the electric tension with respect to its surround. The spike rate r and thus *activation* or *response* of a cell is a function of this value, so the transfer function for a cell in model area A at retinotopic location $\mathbf{x}$ is written as:

$$r^A(\mathbf{x}) = f(p^A(\mathbf{x})) \tag{2.1}$$

Common forms of this function are monotonically increasing or sigmoidal. We will without further notification assume throughout the thesis that the neuron is operating at conditions where the increase of spike rate increases linearly with potential.

The potential of a cell underlies constant change that stems from a leaking current as a function of current potential. The change rate $\dot{p}(t)$ is thus formally defined as

$$\tau\dot{p}(t) = -A \cdot p(t). \tag{2.2}$$

In a full dynamic formulation, the dynamic voltage equation is extended by excitatory and inhibitory connections to end with a generalized notation of the membrane equation [Grossberg, 1988]:

$$\tau\dot{p}(t) = -A \cdot p(t) + net_{ex} \cdot (B - C \cdot p(t)) - net_{inh} \cdot (D + E \cdot p(t)) \tag{2.3}$$

Where A is the strength of the leakage, B, C, D and E transforming the additive and subtractive components. In a dynamic equilibrium the change rates cancel each other out, resulting in $\dot{p}(t) = 0$. The effect is the same as in a car traveling at full speed where the force of the motor brought onto the road and the wind resistance cancel each other out and result in zero acceleration. Depending on motor strength and aerodynamics, this occurs at different top speeds. In our model approaches we are interested in the potential and thus, cell activation, when the balancing happens. Instead of performing a dynamic simulation (which is time-consuming) we thus often calculate the steady-state equation by setting $\dot{p}(t) = 0$ and solving for p:

$$p_{\infty} \propto \frac{B \cdot net_{ex} - D \cdot net_{inh}}{A + C \cdot net_{ex} + E \cdot net_{inh}}. \tag{2.4}$$

This equation will form the basis for the notational format used in this thesis.

2.3.2 Cascade model

In the neural architecture of neocortex six layers can be identified for each individual neuron. As a simplification, we use a cascade model with three model stages that roughly reflect contributions of layer IV (incoming connections and short-range connections), II and III (intra-cortical connections). A reduction to those layers represent computational properties that are common in biological neural architectures and allows modeling of feedforward and feedback connections. The functional effects of this columnar cascade can roughly be mapped onto compartments of cortical area subdivisions (as suggested in [Self et al., 2012]) and are namely linear filtering (layer IV), feedback (layers II and III) and response normalization (layer IV). We capture the contributions of the individual layers with a set of dynamic equations based on the canonical form presented in Eq. 2.3.

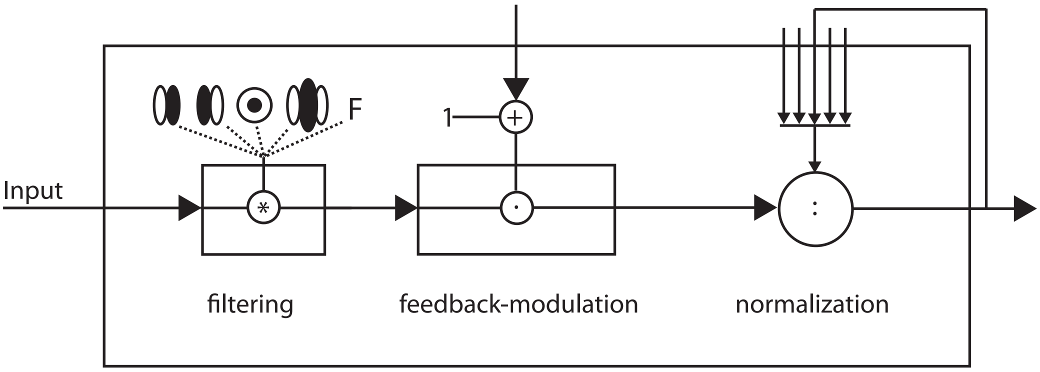


Figure 2.3: Building block of the three-stage cascade model. *Model areas are built following this generic architecture. Input enters the model area from the left and it is initially filtered using layouts of receptive fields that are adapted to the desired functionality of the model area. The responses of the filtering stage are then modulated by input from higher cortical area. At this point, a nonlinear operation is performed on the signal strength. Center-surround interaction leads to a normalization effect in the final stage.*

Filtering

The first component is a model of how cells respond to visual input. Single cells as a first approximation compute the weighted sum of light intensity distribution presented to their receptive field. Two-dimensional visual input is represented as a luminance function of spatial position. The preferred stimulus of a cell is characterized by their impulse response function. Thus, the response of a cell assembly to visual input in conveniently calculated by means of a convolution operation. The response of model cells at position $\mathbf{x}$ is thus

$$r(\mathbf{x}) = I(\mathbf{x}) * K_{pref} \tag{2.5}$$

or

$$r(\mathbf{x}) = r\begin{pmatrix} x \\ y \end{pmatrix} = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} K_{pref}\begin{pmatrix} u \\ v \end{pmatrix} \cdot I\begin{pmatrix} x-u \\ y-v \end{pmatrix} \tag{2.6}$$

Feedback

This purely linear formulation does not account for various non-linear effects that have been reported. Most of the qualities that the visual system provides, like the robustness to disturbances and the high adaptability and generalization capabilities, most probably stem from connections that connect higher cortical areas with lower ones [Hupé et al., 1998, Markov et al., 2013]. Very recently

[De Pasquale and Murray Sherman, 2013] found evidence for the modulatory properties of feedback in the visual cortices of mice. Those feedback connections are believed to play an important role in visual processing, as they enrich local representations with contextual information that is provided by higher visual areas.

The effects of such modulatory feedback are subject of research and mechanisms are not yet understood to a full extent [Hupé et al., 1998, Markov et al., 2013]. This has led to hypotheses about how recurrent modulatory signals might be incorporated into the processing cascade. No feedback input should leave the signal unchanged. In contrast, the sole presence of a modulatory signal should not evoke any response if the preferred stimulus is absent. A simultaneous activation of both feedforward input and modulatory signal should lead to highest overall response of cell. These restrictions lead to a multiplicative model of feedback enhancement. With p_{fb} the modulated cell activation, p its original input from the filtering stage and net_{fb} the strength of the feedback signal, the overall response is modeled by

$$p_{fb} \propto p \cdot (1 + net_{fb}) \tag{2.7}$$

or in a dynamic formulation

$$\dot{p}(t) = -A \cdot p(t) + net_{ex} \cdot \lambda(1 + \beta \cdot net_{fb})(B - C \cdot p(t)) \tag{2.8}$$

The multiplicative term $1 + \beta \cdot net_{fb}$ denotes the response gain modulation generated by top-down connections. Most of the time net_{fb} stems from an integration of contextual surround of the cell from areas higher in the hierarchy. In case of a pure feed-forward signal processing sweep, the modulation is switched off by setting $\beta = 0$.

Normalization

Various non-linear effects that are observable are not captured by the first two stages of processing. Simple cells exhibit non-linearities in their responses to stimuli [Carandini et al., 1997]. We incorporate a stage of divisive normalization in our model, which extends the linear model to include mutual shunting inhibition among a large number of cortical cells. It acts divisively upon the activation level of target neurons and its effect in the model is to normalize the linear responses by a measure of stimulus energy over a population of cells in the neighborhood of a neuron.

Such kind of mutually competitive interaction has been observed in extracellular recordings of simple cells in the primary visual cortex of anesthetized macaques [DeValois et al., 1982a, Carandini and Heeger, 1994]. It accounts for both the linear and the nonlinear properties of the cells and allows to process a high dynamic range of response activations [Heeger, 1992].

$$\dot{p}_i = -\alpha_p p_i + (\beta - p_i) \cdot I_i - p_i \cdot q_i \tag{2.9}$$

$$\dot{q}_i = -\alpha_q q_i + \sum_{j \in \mathcal{N}} c_j g(p_j) \tag{2.10}$$

with I_i the input and $\mathcal{N}$ the neighborhood integration of a neuron i, where c_j denote spatial weighting coefficients of the local neighborhood. At equilibrium, the following steady state equations can be derived:

$$p_{i,\infty} = \frac{\beta I_i}{\alpha_p + I_i + \frac{1}{\alpha_q} \sum_{j \in \mathcal{N}} c_j g(p_j)} \tag{2.11}$$

$$q_{i,\infty} = \frac{1}{\alpha_q} \sum_{j \in \mathcal{N}} c_j g(p_j) \tag{2.12}$$

This application of normalization has previously been investigated [Grossberg, 1980, Heeger, 1992, Carandini and Heeger, 2012, Carandini et al., 1997] and used in various approaches touching different domains, such as the disambiguation of local motion [Bayerl and Neumann, 2004, Beck and Neumann, 2011] the processing of transparent motion [Raudies and Neumann, 2010, Raudies et al., 2011] the detection of texture boundaries [Thielscher and Neumann, 2003], the extraction of object boundaries using texture compression [Weidenbacher and Neumann, 2009], and the analysis and representation of biological motion sequences [Layher et al., 2014]. [Brosch and Neumann, 2014a] provides a detailed analysis of the dynamic equations the model is comprised of.

The thesis now continues with instances of these principles for motion and form processing. The upcoming Chap. 3 will incorporate the principles in a model of motion estimation using event-based input from a neuromorphic sensor. Chap. 4 applies the generic mechanisms in a model of form processing, where contextual influence and normalization mechanisms contribute to a segregation of figure and ground.

3 Event-based motion processing

3.1 Introduction

Evolutionary success is about optimizing chances of survival, food acquisition and reproduction. For this reason, many species have developed high mobility and clear perception of their environment. With the ability to move and others moving in the vicinity, the brain has developed according sensory mechanisms for motion perception based on the visual input that enters the brain through light-sensitive neurons on the retina. In higher mammals, the perceptual capabilities concerning motion allow the extraction of an immense amount of information including a motion-based scene segregation into figure and ground, movement direction, collision avoidance, recognition of individuals, perception of transparent motion, and more.

To be able to extract that information, the cortex has developed structures and mechanisms to detect, represent and process visual motion patterns. But how do these mechanisms work? As seen in Chap. 2, the visual perception starts with the projection of light onto the retina. Moving individuals or objects moving in the vicinity cause the projections on the retina and thus, neural activations, to shift. The descriptive field of movement vectors is called *optic flow field.* These temporally variant activations are processed by visual mechanisms in occipital cortex and extracted motions are represented by neural activations in various dedicated areas. [Hubel and Wiesel, 1962] found directionally selective (DS) cells in the visual cortices of cats soon after their work on fundamental neural processing principles [Hubel and Wiesel, 1959]. Latest after [Barlow and Levick, 1965] discovered the mechanisms of directionally tuned cells in rabbit retina, many researchers and laboratories started to intensively investigate the mechanisms of motion processing. [Borst and Euler, 2011] provide a good review and summary on early stages of motion sensing.

The investigations have led to a number of models that consider the biological estimation of visual motion. The first popular model was presented by [Hassenstein and Reichard, 1956]. It performs a spatio-temporal correlation of visual impulses that is inspired by the visual system of the fly. Two spatially separated receptor cells are connected, whereas one of them adds a temporal delay to the output of its signal. Whenever a visual stimulus first elicits a response at the delayed receptor and the right time later at the second receptor, the detector elicits a response and indicates local motion. With only a few extensions like a mechanism to reduce the probability to respond to flickering input this rather simple concept at an initial stage suffices to allow flying insects like the common housefly to become oriented and navigate through their environment. Any reader who has ever been disturbed by such an insect might agree that despite the simplicity of the underlying mechanisms its perception of the environment is accurate and response times are remarkable.

The model of [Hassenstein and Reichard, 1956] was later adapted and simplified by [Barlow and Levick, 1965]. Another model that has reached a high level of publicity is presented by [Adelson and Bergen, 1985]. With it, many explanations of effects in visual motion perception were qualitatively explained. The approach is based on the detection of oriented spatio-temporal energy. It consists of linear filters that are oriented in space-time and tuned in spatial frequency. The authors of [vanSanten and Sperling, 1985] showed that the model of [Hassenstein and Reichard, 1956] and the model by [Adelson and Bergen, 1985] are mathematically equivalent.

How is an initial motion estimation potentially performed and what are the mechanisms that lead to directionally selective cells? In a nutshell, motion sensing requires neural mechanisms that are sensitive to objects passing through their receptive field. In order to be selective to motion instead of static contrast, the spatio-temporal response function of those cells needs to be non-separable. In primates, these receptive field types have been discovered [DeAngelis et al., 1995], and they nicely fit into the concept of [Adelson and Bergen, 1985]. The precise source of how these inseparable fields are generated from separable components remained unclear until [DeValois et al., 1982a] who analyzed spatio-temporal receptive fields in visual cortices of macaque monkeys. These insights showed that the spatio-temporal characteristic can be achieved by a combination of separable subfields.

In this work, we present a mechanism of motion estimation based on those physiological findings of macaque visual cortex, see Sec. 3.4. The model uses the output of a neuromorphic silicon retina [Lichtsteiner et al., 2008] and incorporates physiological findings of receptive field characteristics to generate a motion sensitive model. The initial estimation is followed by subsequent processing steps, as they are found in the structures along the dorsal path of the primate brain.

In the following sections, we will first introduce the principles of a new neuromorphic sensor that resembles the functionality of retinal photoreceptors (Sec. 3.2).

This sensor type mimics sensory processing of the retina in that it provides a continuous stream of visual events that indicate local luminance changes. We demonstrate how physiological findings about the motion selective cells are applied to the stream of events to yield a representation of initial motion by means of directionally selective cells (Sec. 3.4.1). Subsequent processing steps are motivated by investigations about mechanisms available along the dorsal processing pathway. Here, canonical elements from Chap. 2.3 are employed. This includes a recurrent architecture for motion integration, normalization and contextual biasing found in area MT. Effect of such a contextual interaction is an increase of tolerance towards noise and the effect of generating more homogeneous early motion representations at V1 (3.4.3). In the proposed model, we also investigate the influence of contextual surround inhibition within model V1 (Sec. 3.4.2). Such a normalization reduces influence of spurious motion estimates at elongated 1-dimensional image features which positively affects the consequences of the *aperture problem*. Physiological as well as psychophysical effects of motion and form interactions are presented explained using the concept of *motion streaks* or *speedlines* in Sec. 3.5. An example of a possible algorithmic implementation of the mechanism is shown in Sec. 3.6.

3.2 DVS - A neuromorphic retina-like sensor

Classically, visual input is acquired using a camera that captures full frames at one point in time. This approach stems from the time when still projections of scenes were captured on a chemical photo-reactive surface. This concept has been successfully ported to the recording of scenes in motion as well. Eadweard Muybridge was one of the pioneers of *chronophotography* and captured the movement of a running horse with a rapid exposure of multiple photographic plates. When these images are presented successively in rapid order, our visual systems integrates the individual frames into a moving percept. This striking effect is known since the late 19th century and our cinematographic technology is still based on this illusion. Video cameras capture full frames at a fixed frame rate at distinct points in time, and the concept has survived the transition from photosensitive material to digital sensors.

When a visual scene is captured at a frame rate R of about $25Hz$ or higher, the luminance level of one individual position is very probable to be constant over multiple frames. Nonetheless, the full set of pixels is transmitted, processed and stored in modern camera systems. Over the years, resolution and frame rates have been increased, but the concept of capturing full frames remained. The thirst for computational power, storage and energy consumption of such camera systems increased accordingly. On the other hand fast image changes, like flickering or movements of a rapidly moving object, are still integrated into one frame and cause motion blur. One of the basics in signal theory, the Nyquist-theorem, states that with a given frame rate R (respective sample frequency), only signals with a frequency or less than $R/2$ can be reconstructed.

A lot of research and money has been put into the development of full-frame camera sensors, and their low prices, high quality and ubiquitous availability have attracted many vision science labs to use them for their research. However, the immanent concept of a *frame* as being the input for vision processing has blurred the fact that at a retinal level, vision works differently and independent of frames (see also Sec. 2). Here, luminance changes are detected and transmitted asynchronously between local retinal positions. For precise models of visual perception, a synchronous sampling of all retinal positions does not capture this fundamental difference.

Therefore, in 1992, [Mahowald, 1994] introduced the first *silicon retina* that breaks with the classic paradigm of full frames. It introduces a neuromorphic concept of building a vision sensor that models individual photoreceptors on a pixel-level. A dedicated circuit for every pixel emulates the function of each individual photoreceptor, bipolar and ganglion cells in the retina, see Fig. 3.1. Such a circuit consists of a photoreceptive element, a differencing amplifier and decision or comparator units. The current produced by the photoreceptor is constantly monitored and when a change is detected, the circuit generates an *event* [Delbrück, 2012]. When an event was fired, the local reference pixel intensity level is adjusted to the new measurement, that means the pixel is reset to the new reference luminance. This takes only a few microseconds, the effective signaling frequency of an individual pixel is measured in the order *Kilo*hertz. The output of such a sensor is a continuous stream of local events that indicate local changes in contrast and its sign.

This method of sampling provides a list of interesting properties. In the output stream, every information means a novelty in the sense that it indicates a change of luminance in the visual field. Redundant information is not transmitted. This has direct effects e.g. on energy consumption and necessary bandwidth. The automatic adaptation to local luminance levels allows useful responses over a very large range of luminance levels. A standard camera needs to adjust to luminance levels by globally changing sensor sensitivity or changing the aperture size. The DVS has a sensitivity range of $120dB$ which means a dark to light ratio $1 : 10^6$, or an equivalent of about 20 f-stops (aperture levels) in photographic measurements. For comparison, high-end photographic (still photography) sensors like the ones built in Canon's EOS 5D-Series provide an intra-frame dynamic range of about 11 stops ($66dB$ or $1 : 2000$) without mechanically changing lens apertures or sensitivity values [Koo, 2010, Adams, 2010]. The biologically inspired processing mechanisms of the DVS increase this sensitivity range by two orders of magnitude.

The authors of [Lichtsteiner et al., 2008] engineered this concept to a product con-

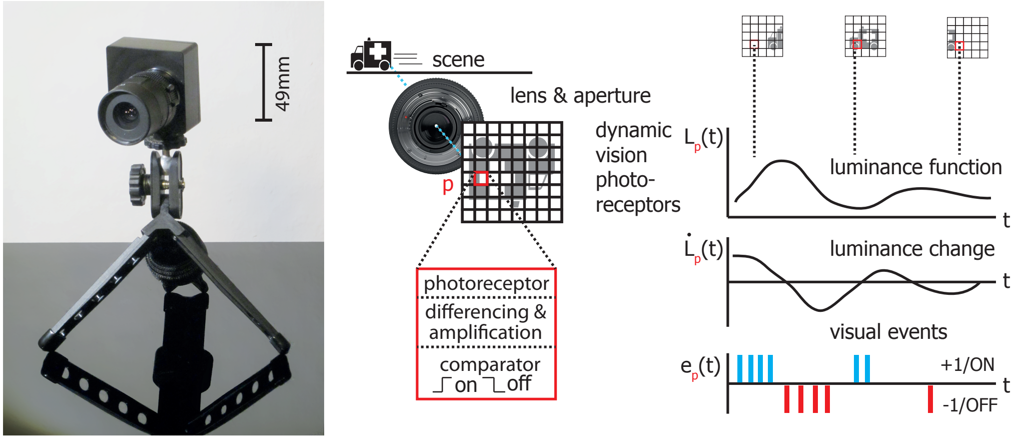


Figure 3.1: Left: ***The Dynamic Vision Sensor (DVS)***. Middle: *Functional sketch of pixel components for event generation.* Right: *Illustration of working principle for individual pixels. Events are generated when pixel circuitry detects significant changes in local luminance level.*

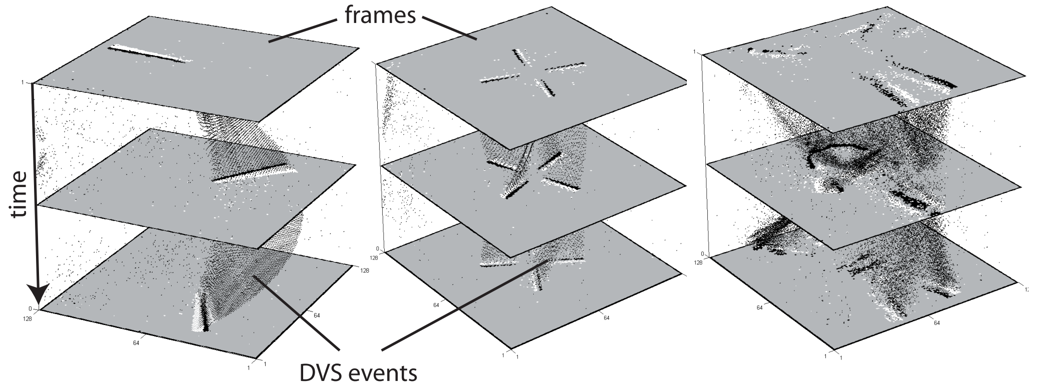


Figure 3.2: Frame- and event-based recordings. *Ordinary sensors produce frames at fixed intervals, effectively integrating visual changes in between. Events of the DVS sensor densely represent changes with much higher precision.*

taining a sensor that is small, easy to handle and affordable enough to be acquired by research facilities (see Figure 3.1, *left*). For ease of applicability, they wrapped the sensor circuitry into a framework of signal processing that allows to transmit event signals to a connected computer via USB port. A data processing layer automatically resolves event collisions, and dynamic thresholds guarantee constant performance of the sensitive electronic components independent of surrounding temperature and other influences. The authors advertise the high sensitivity and the large brightness range of the sensor. They also provide an API programmed in *Java* that supports developments with a large collection of software modules.

On the downside, the spatial resolution of the sensor is lean with only 128×128 on a 1/2" sensor, because the circuits necessary to model such a processing requires about 10 times the size of standard pixels in a CMOS approach. Very recently, [Brandli et al., 2014], from the same lab, published a new version of the sensor that combines the benefits of a standard frame-based sensor with the event-based mechanisms. It allows to simultaneously readout the asynchronous detection of brightness changes and the synchronous readout of linear intensities. Data from these two different channels can easily be superimposed because of one common optical system. The resolution is also increased to 240×180 pixels. The new sensor is called *Dynamic and-Active Pixel Vision Sensor* (*DAVIS*) and also available from *Inilabs*.

3.2.1 Previous DVS applications

The availability of such a sensor at an affordable price has motivated many labs to focus on the challenges and chances that come with the new concept. At the time of writing, the list of applications ranges from tracking [Drazen et al., 2011] and detection of gestures or actions [Fu et al., 2008, Lee et al., 2012] to embedded event-based sensory-motor systems [Delbrück and Lang, 2013, Conradt et al., 2009, Inilabs, 2014] and stereo applications [Benosman et al., 2011, Camunas-Mesa et al., 2014]. The resemblance of actual neural mechanisms has motivated to embed the sensor into larger architectures like CAVIAR [Serrano-Gotorredona et al., 2009] as well or to provide data for spiking neural network [Gibson et al., 2014]. The problem of the estimation of optic flow using the sensor has been approached by [Clady et al., 2014, Benosman et al., 2012, Benosman et al., 2014, Barranco et al., 2014, Abdul-Kreem and Neumann, 2014]. A novel method is presented in this thesis, see Sec. 3.3.

3.2.2 Asynchronous visual events

On the DVS, the circuitry of individual pixels on the sensor surface measures the local luminance level and detects relative intensity changes. If such a change occurs, the circuitry evokes one of two possible signals, depending on the sign of the change. In case of an *increase* of luminance, a *ON* signal is produced, and in the opposite case, when the local intensity *decreased*, an *OFF* signal is generated. The sensor architecture contains a 16-bit counter that generates timestamps with microsecond precision. Each *ON* or *OFF* signal is labeled with the time of occurrence, thus producing a corresponding *event*, and forwarded to the data processing layer, that provides the event to a connected computer via buffering structures and USB interface. The output of such a sensor is a stream of visual events with arbitrary spatial order and high temporal precision. We define the output of the event based sensor as the function

$$e : \mathbb{R}^2 \times \mathbb{R} \to \{-1, 0, 1\}. \tag{3.1}$$

No events are produced unless changes in the luminance function occur, so $e = 0$, except at positions and times $(\mathbf{p}_k; t_k)$ where

$$e(\mathbf{p}_k; t_k) = \begin{cases} -1 & \text{if } darker/OFF\ Event \\ 1 & \text{if } brighter/ON\ Event \end{cases} \tag{3.2}$$

For applications that require the synchronous operation of multiple DVS cameras, the timestamps can be synchronized using a 1-pin trigger cable between sensors. Such a synchronization is necessary for stereo applications or general recordings with more than one DVS camera.

3.2.3 Adress-event-representation (AER)

The data processing layer of the sensor provides an asynchronous stream of visual events that consists of spatial and temporal information where events occurred. This information is provided with data packets that each contain two integer values T and P. T directly codes the timestamp of the event in microsecond precision. The second integer value P contains the spatial position and event type. A few bit transforming operations are required to make it usable. The event type t is coded within the last bit of P and extracted with a modulo operation $t = P\,mod\,2$, where 0 indicates an ON and 1 indicates an OFF-Event. In the following, $\hat{P}$ only contains the remaining bits of P without the last one. The spatial position is extracted as follows:

$$x = 128 - \hat{P}\,mod\,128 \tag{3.3}$$

$$y = 128 - (\hat{P} - x)/128 \tag{3.4}$$

3.2.4 Visualization methods

The DVS produces an asynchronous stream of visual events that indicates local changes in contrast at a retinal position, the sign of the change, and the precise time of change. There are different ways of visualizing such an event stream. Figure 3.3 A-F shows a selection used in this thesis. The actual output of the sensor is a list of integer 2-tupels that contain the position, type and time of individual events (B). Such an individual event can be visualized in a two-dimensional frame using the x and y coordinates.

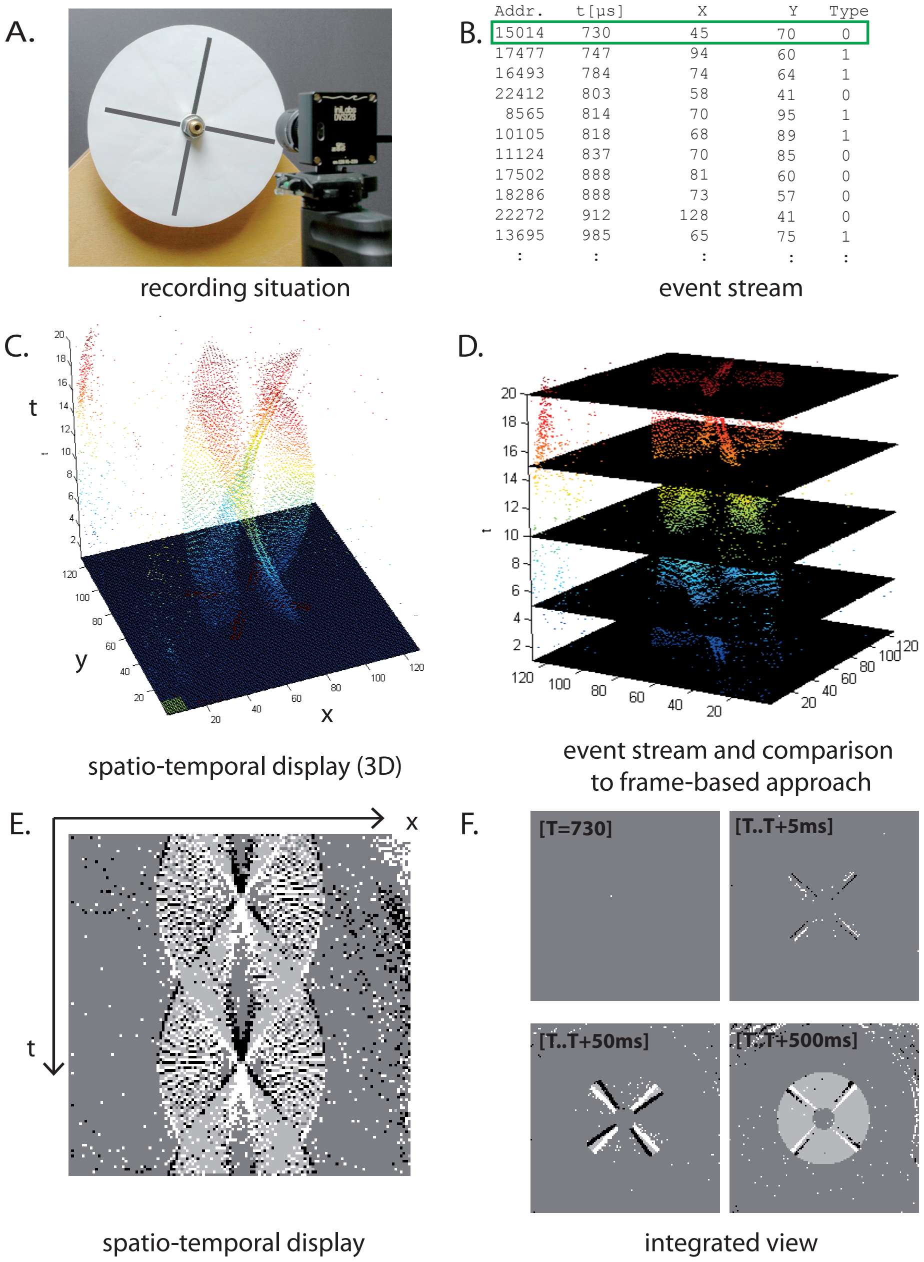
The most straight-forward way to represent an event is a 3-dimensional spatio-temporal display (C,D). Here, all events are positioned in a cube-shaped space with the axes position and time. A projection of that space can easily be accomplished and presented on screen or paper. This display is especially well suited for interactive displays, where the camera position can be changed by the viewer in real time. This type of display nicely shows the continuous character of incoming events when contrasted with the standard frame-based methods of capturing sequences. However, on a printed and thus static medium this type of visualization sometimes generates a less comprehensive display.

The visualization that is least intuitive but essential for the upcoming investigations is the 2-dimensional spatio-temporal display (E). It is repeatedly used to explain the basic concepts of the processing. In this type of display, only one spatial component (the y or x component) is represented along one axis of the display. The other axis is replaced by the temporal component of the stimulus. Throughout the progress of this thesis, the positive temporal component of such a display will be directed downwards, meaning that events farer down the illustration are younger than those located higher in the visualization. This allows to make the reader familiar with basic concepts of temporal processing and effects that occur from the specific way the sensor produces output.

To achieve a more meaningful visualization of the recorded data, multiple events are integrated and visualized in one image (F). This type of display proved to be most suitable for printed form, but it easily generates the impression that individual *frames* are processed instead of events.

3.2.5 Challenges for real recordings with the DVS

In recording situations, the DVS can be operated like a normal camera. When the lens focus and aperture are set, the high dynamic range allow recording under a large range of illumination conditions. However, achieving real recordings with high precision requires a well-planned setup. The apparently most convenient way to control and record stimuli would be to generate them synthetically and display them on a LCD screen. The results of this approach seem satisfying at



recording situation

Addr.	t[µs]	X	Y	Type
15014	730	45	70	0
17477	747	94	60	1
16493	784	74	64	1
22412	803	58	41	0
8565	814	70	95	1
10105	818	68	89	1
11124	837	70	85	0
17502	888	81	60	0
18286	888	73	57	0
22272	912	128	41	0
13695	985	65	75	1
:	:	:	:	:

event stream

spatio-temporal display (3D)

event stream and comparison to frame-based approach

spatio-temporal display

integrated view

Figure 3.3: Different visualization techniques for asynchronous DVS data. A: *Setup of recording situation.* B: *Snippet from actual data stream that arrives from the interface.* C,D: *Projection of a three-dimensional display technique. Here, the dense sampling of the temporal function is visible in contrast to the synchronous samples as would happen in a standard camera.* E: *Spatio-temporal view in y/t system where slanted structures from moving patterns become visible.* F: *Display of an individual event and events integrated over different time periods.*

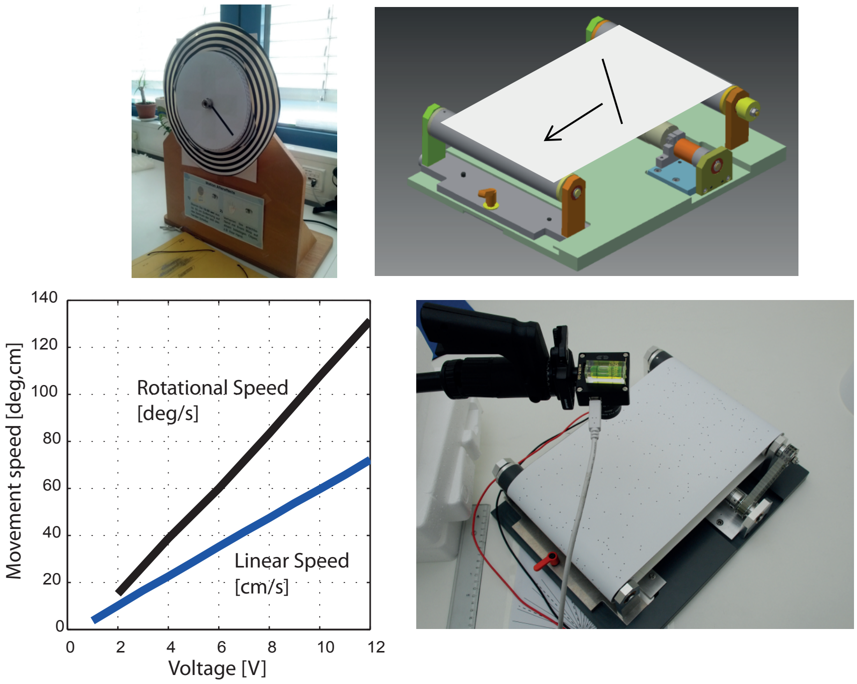


Figure 3.4: Devices used for recordings of stimuli. Top: *Rotational device and construction sketch of linear motion device.* Bottom,left: *Controllable speeds for both devices are an approximate linear function of applied voltage.* Right: *Real recording situation with random dot kinematogramm on linear device.*

first, but they are not. Modern LCD screens operate at a refresh frequency of 60 Hz, that means a new image is presented every 16.6 ms. Modern products operate at internal refresh rates of 200 Hz and more by repetition of the current frame. The high temporal resolution of the sensor easily samples this process and the synchronization provided by the screen is contained in the recorded data.

Recordings for this thesis were performed using two mechanical devices that *physically* move the desired stimulus on a rotational or linear path. The first device performs rotational movements. Stimuli are printed on a $200mm$ disk and centrally attached to the axis. A geared DC motor with operates at 2 to 12 Volts generates a rotational motion at speeds between 20 and $130°/s$.

The second device is able to perform linear motion. A $205mm \times 733mm$ sheet is looped and strained into the device. Again, a DC motor operates at 1 to 12 Volts and generates a linear motion with a speed of 0 to $700mm/s$. Figure 3.4 shows the two devices and respective speed over voltage settings.

3.2.6 Processing limitations

The event generation on the DVS is limited by the ability of the data processing layer to forward events to the USB port, and the maximum data throughput of the USB port itself. The approximate throughput of the USB 1.0 port is about $1 * 10^6$ bytes per second. One event is represented by a 16-Bit-integer value for the timestamp and another 16-Bit-integer value for the position, which in sum equals 32 Bits or 4 Bytes. The maximum throughput of the USB port is thus approximately 250.000 events with no data overhead considered. The resolution of the sensor if $128 \times 128 = 16.384$ pixels, so the data bus is maxed out when there is more activity than about 15 events per pixel per second. The reader might recall that the potential refresh rate of individual pixels is about $25kHz$. With complex stimuli covering the complete visual field of the sensor and many contrasts that are quickly moving, this limitation is easily reached. As a result, the event stream is interrupted and timestamps of individual events are spurious. In [Lichtsteiner et al., 2008] the authors specify that the device makes use of the USB 2.0 standard, which would increase tolerance towards bursts of many events, but still does not provide enough capacity to completely remove the bottleneck of event transmission. In the recordings used in this thesis, we paid attention to not overload the transmission capacities of the sensor.

3.2.7 Synthetic stimuli

With the recording of proper visual stimuli for the DVS being cumbersome, an option is to create synthetic streams of visual events. Here, direction and speed of a motion can be controlled precisely. The goal is to create an event stream as it would be produced by a given shape. In a first step, likelihoods of positions where events are generated are created using this shape. Events are then statistically spread over a definable time to yield an effective speed. This method generates a good approximation of actual sensor output with high flexibility and precision.

The model of movement is an affine transformation of an initial shape or image. The shape can be defined as a black and white or as gray scale image of the desired resolution. The resolution may not match the actual resolution of the sensor, but if we chose to simulate synthetic stimuli in the resolution of the sensor (128 × 128 pixels). The position of the artificial shape is calculated along its desired trajectory but with constant speed of *one pixel at a time*. The movement is performed by generating transformed versions of the shape using an affine transformation of each target pixel

$$\mathbf{p}' = k\,T \cdot \mathbf{p} \tag{3.5}$$

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = k \cdot \begin{pmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \tag{3.6}$$

With a to f the affine parameters including translation, scaling, shearing and rotation and k the sequence parameter which is slowly increased to generate only small movements at a time. The difference between the current transformed shape and the previous image indicates regions where local luminance levels are raised or decreased. Events will be generated at these positions with their respective type (ON or OFF). When the shape image contains gray scales between 0 and 1, so does the difference image. These may be interpreted as probability levels for event generation.

The final step is to perform the transition from the step-based transformation of motion to a continuous event stream of defined speed. The model performs transformations at distinct points in time. Generated events are thus scattered across the range between two transformations. To simulate the noise level of the sensor, each pixel is additionally activated with a probability of 0.003. Figure 3.5 illustrated the mechanism of artificial event generation.

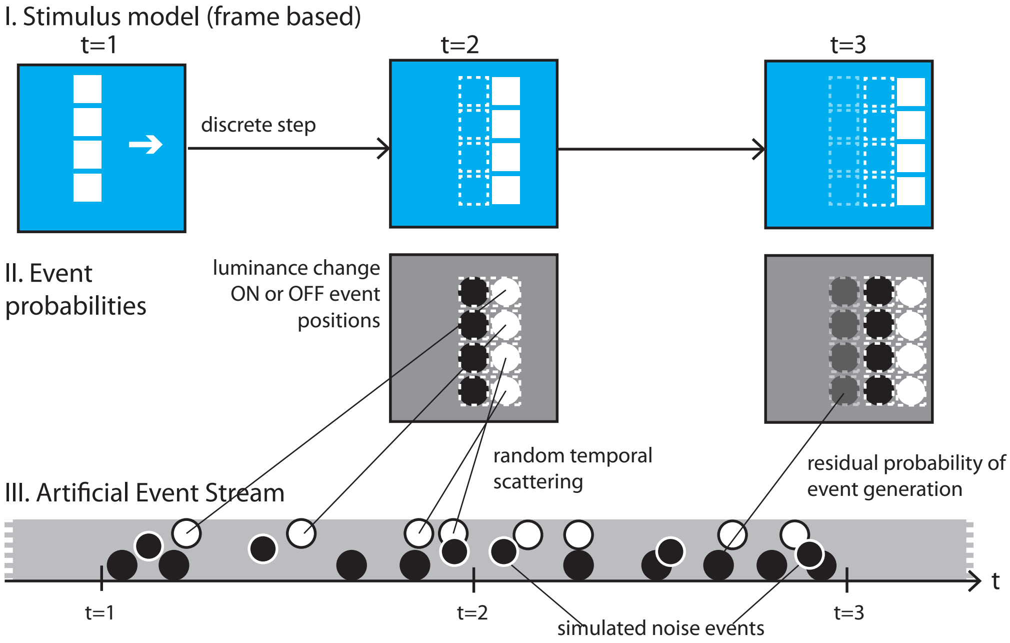

I. Stimulus model (frame based)
t=1
t=2
t=3
discrete step
II. Event probabilities
luminance change ON or OFF event positions
random temporal scattering
residual probability of event generation
III. Artificial Event Stream
t
t=1
t=2
t=3
simulated noise events

Figure 3.5:* *Artificial event generation. *Using a stimulus model, the generation mechanism keeps track of locations where ON or OFF events need to be generated with high probability. If events are locally generated according to the likelihood, they are scattered over the timespan.*

3.3 On event based motion estimation

Now that the reader is familiar with the sensor we will approach motion estimation. Intuitively, the sensor seems very appropriate for a high-quality estimation of optic flow, in that it provides changes of visual input with high temporal precision. As motion estimation is essentially estimating change over time, estimation proper motion direction seems straight-forward. Actually, other labs have already presented algorithms of motion estimation with AER input. There is a number of approaches from previous labs dealing with motion estimation on a dynamic vision sensor. [Clady et al., 2014] explored mechanisms to estimate time-to-contact, and there are works that contribute either adapted [Benosman et al., 2012, Benosman et al., 2014] or original [Barranco et al., 2014] methods for motion estimation on the sensor. However, their approaches show deficits under close investigation. In a nutshell the approaches rely on the estimation of temporal and spatial derivatives that cannot be estimated reliably using the sensor input. We will analyze this in the subsequent sections using a formal definition of the sensor input, its implications on the assumptions used for motion estimation. The core problem is basically that gradients are not easily estimated with the sensor. We then present a new approach using spatio-temporal filters that are motivated by physiological findings in Sec. 3.4. Our contribution that is presented in the following sections contrasts to approaches like [Benosman et al., 2012, Benosman et al., 2014] in that our mechanism relies on spatio-temporal filters that resemble neural mechanisms found along early stages in visual cortex.

3.3.1 Nomenclature and principal problem

When motions occur in the visual field of the sensor, it generates a series of ON and OFF-events. The sampled events are generated at the contrast edges of the moving object. In *x*-*y*-*t*-space these events cluster into a region where events occur with a high probability and this region is slanted with respect to the t axis, where the angle indicates the speed of the motion (see Fig. 3.6). Estimation of motion and speed is about estimating the slant of the region.

To describe the output of the event-based sensor, we define the function

$$e : \mathbb{R}^2 \times \mathbb{R} \rightarrow \{-1, 0, 1\} \tag{3.7}$$

which is always zero except for tuples $(x_k, y_k; t_k) = (\boldsymbol{p}_k; t_k)$ which define the location and time of an event k generated when the luminance function increases or decreases by a significant amount. In other words, the function that defines the event generation $e(\boldsymbol{p}_k; t_k) = e_k$, generates 1 if the log-luminance changed more than a threshold ϑ, i.e. an ON event, and -1 if it changed more than $-\vartheta$, i.e. an OFF event. This sampling of the lightfield essentially represents the temporal

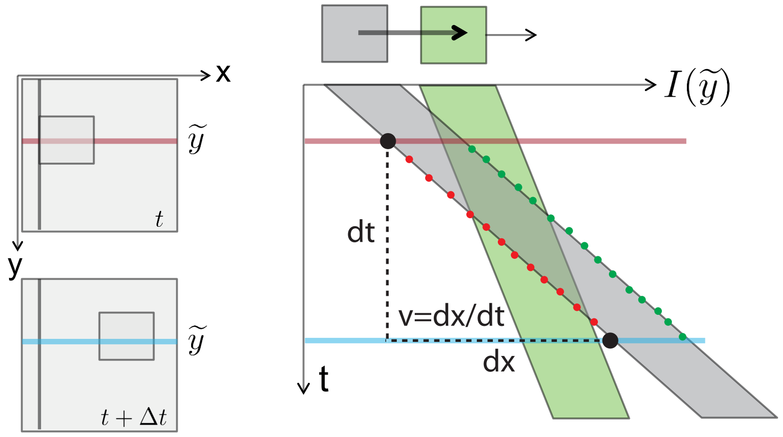


Figure 3.6: Principle of motion estimation: Left: *Moving object in spatial coordinate system at two discrete points in time.* Right: *In spatio-temporal coordinates, the object path appears as slanted region with angle to t axis a function of speed. Events are produced at the proximal and distal edge of the object.*

derivative of the luminance function g

$$\frac{d}{dt}g(\boldsymbol{p};t) = g_t(\boldsymbol{p};t) \approx \frac{\vartheta}{\Delta t} \sum_{k:\, t_k \in (t-\Delta t, t]} e_k\,, \tag{3.8}$$

with ϑ the sensitivity threshold of the event-based sensor.

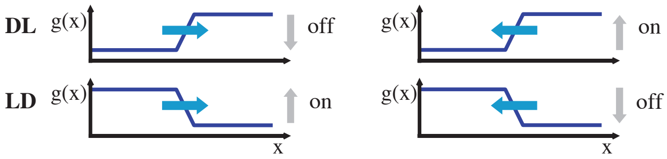


Figure 3.7: Moving DL (dark-light) and LD (light-dark) edge, *either to the left or to the right (denoted by blue arrows), have an associated temporal on/off signature. Note that without knowledge about the edge type (DL vs. LD), an on/off event alone is insufficient to determine the motion direction.*

3.3.2 Luminance constancy assumption

To estimate local translatory motion we assume throughout the paper that the gray level function remains constant within a small neighborhood in space and time, i.e. $g(x, y; t) = g(x + \Delta x, y + \Delta y; t + \Delta t)$ (gray level constancy; c.f. [Horn and Schunck, 1981]). Note that due to the low latency of $15\mu s$ of the event-based sensor [Lichtsteiner et al., 2008], this assumption is more accurate than for conventional frame based sensors. Local expansion up to the second order yields the constraint $\boldsymbol{\Delta x}^T \nabla_3 g + 1/2 \boldsymbol{\Delta x}^T \boldsymbol{H}_3 \boldsymbol{\Delta x} = 0$. Here, $\boldsymbol{\Delta x} = (\Delta x, \Delta y, \Delta t)^T$, $\nabla_3 g = (g_x, g_y; g_t)^T$ is the gradient with the 1st order partial derivatives of the continuous gray-level function, and $\boldsymbol{H}_3$ denotes the Hessian with the 2nd order partial derivatives of the continuous gray-level function that is defined in the x–y–t-domain. If we further assume that the 2nd order derivative terms are negligible (linear terms dominate) we arrive at the spatio-temporal constraint equation that has been used for least-squares motion estimation. The least-squares formulation is based on a set of local constraint measures over a small neighborhood under the assumption of locally constant translations [Lucas and Kanade, 1981], i.e. $\boldsymbol{g}_x u + \boldsymbol{g}_y v + \boldsymbol{g}_t = \boldsymbol{0}$ given that $\Delta t \to 0$ and $\boldsymbol{u}^T = (u, v) = (\Delta x / \Delta t, \Delta y / \Delta t)$. The local image motion $\boldsymbol{u}$ of an extended contrast can only be measured orthogonal to the contrast (normal flow, [Wallach, 1935, Wuerger et al., 1996, Barron et al., 1994, Fermüller and Aloimonos, 1995]). For simplicity, we assume a vertically oriented gray level edge ($g_y = 0$). Then the motion can be estimated along the horizontal directions (left or right with respect to the tangent orientation of the contrast edge). When the edge contrast polarity is known (light-dark, LD, $g_x < 0$ or dark-light, DL, $g_x > 0$) the spatio-temporal movements can be estimated without ambiguity. For an DL edge if $g_t < 0$ the edge moves to the right, while for $g_t > 0$ the edge moves to the left (c.f. Fig. 3.7).

For an LD edge the sign of the temporal derivatives g_t changes for both respective movement directions, i.e. only the ratio of gray-level derivatives yields a unique direction selector orthogonal to the contrast contour. This means that $\mathrm{sgn}(g_x/g_t) = -1$ implies rightward motion while $\mathrm{sgn}(g_x/g_t) = 1$ implies leftward motion, irrespective of the contrast polarity. Note, however, that an estimate of g_x is not easily accessible from the stream of events. Thus, a key question is to what extend the required spatio-temporal derivative information is available and can be estimated from the plenoptic function $P(\cdot)$ that is sampled by the asynchronous event sensor.

3.3.3 Moving gray-level edges and the spatio-temporal contrast model

We describe the luminance function g for a stationary DL transition by convolving a step edge $\mathcal{H}(\cdot)$ with a parameterized Gaussian,

$$g_\sigma(x) = \frac{c}{\sqrt{2\pi}\sigma} \cdot \mathcal{H}(x) * \exp\left(-\frac{x^2}{2\sigma^2}\right) + g_0 = c \cdot \mathrm{erf}_\sigma(x) + g_0, \tag{3.9}$$

with c denoting the luminance step height, g_0 the basic luminance level, and "$*$" denoting the convolution operator (since we only study the derivatives, we assume $g_0 = 0$ without affecting generality). The parameter σ controls the spatial blur of

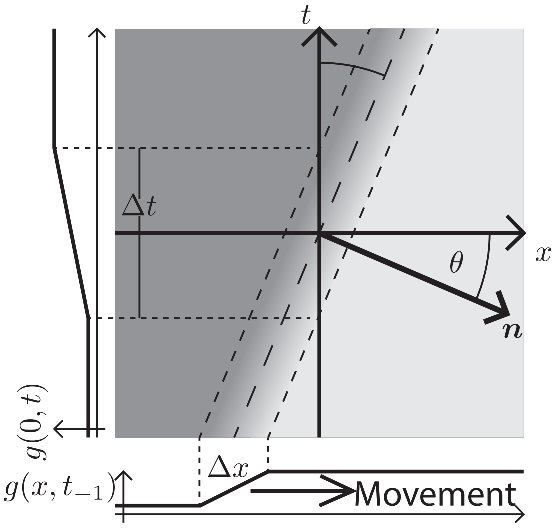


Figure 3.8: Rightward moving 1D edge illustrated in the x–t-domain. *The velocity is defined by the direction and the speed of the spatio-temporal change. In the case depicted here, the direction is to the right and the speed is encoded by the angle θ between the x-axis and the normal vector $\boldsymbol{n}$ along the spatio-temporal gradient direction (measured in counter-clockwise rotation). Alternatively, for a contrast edge of known finite length Δx, the speed can be inferred from the time Δt, it takes the contrast edge to pass a specific location on the x–axis.*

the luminance edge with $\sigma \to 0$ resulting in the step-function. Different contrast polarities are defined by $g_\sigma^{DL}(x) = c \cdot \text{erf}_\sigma(x)$ and $g_\sigma^{LD}(x) = c \cdot (1 - \text{erf}_\sigma(x))$, respectively [Neumann and Ottenberg, 1992].

When this gray-level transition moves through the origin at time $t = 0$ it generates a slanted line with normal $\boldsymbol{n}$ in the x-t-space (c.f. Fig. 3.8). The speed s of the moving contrast edge is given by $s = \sin(\theta)/\cos(\theta)$, where θ is the angle between $\boldsymbol{n}$ and the x-axis (this is identical to the angle between the edge tangent and the t–axis). For a stationary gray-level edge (zero speed) we get $\theta = 0$ (i.e. the edge generated by the moving DL transition in the x–t-domain is located on the t-axis). Positive angles $\theta \in (0°, 90°)$ (measured in counterclockwise direction) define leftward motion, while negative angles define rightward motion. For illustrative purposes, we consider an DL contrast that is moving to the right (c.f. Fig. 3.8). The spatio-temporal gradient is maximal along the normal direction $\boldsymbol{n} = (\cos\theta, \sin\theta)^T$. The function $g(x;t)$ describing the resulting space-time picture of the movement in the x-t-space is thus given as

$$g_{\sigma\theta}(x;t) = \frac{c}{\sqrt{2\pi}\sigma}\mathcal{H}(x_\perp) * \exp\left(-\frac{x_\perp^2}{2\sigma^2}\right) , \tag{3.10}$$

with $x_\perp = x \cdot \cos\theta - t \cdot \sin\theta$. The respective partial temporal and spatial derivatives are given as

$$\frac{\partial}{\partial t} g_{\sigma\theta}(x;t) = \frac{-c}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x_\perp^2}{2\sigma^2}\right) \cdot \sin\theta\,, \tag{3.11}$$

$$\frac{\partial}{\partial x} g_{\sigma\theta}(x;t) = \frac{c}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x_\perp^2}{2\sigma^2}\right) \cdot \cos\theta\,. \tag{3.12}$$

Now, recall that the event-based DVS sensor provides an estimate of g_t at a specific location (c.f. eqn. (3.8)). For a moving contrast profile this leads to a changing luminance function along the t-axis (side graph $g(0,t)$ in Fig. 3.8). The temporal derivative of this profile is formally denoted in eqn. (3.11). Given a *known velocity* specified by θ, we can combine equations (3.11) and (3.12) to determine g_x as

$$\frac{\partial}{\partial x} g_{\sigma\theta}(x;t) = -\frac{\partial}{\partial t} g_{\sigma\theta}(x;t) \cdot \tan\theta\,. \tag{3.13}$$

In sum, the temporal edge transition can be reconstructed in principle from a (uniform) event sequence at the edge location for a specific motion direction, given that

- a reliable speed estimate is available to infer a robust value for θ, and
- reliable estimates of temporal changes have been generated as an event cloud over an appropriately scaled temporal integration window Δw_t.

Note that both parameters, θ and Δw_t, need to be precisely estimated to accomplish robust estimates of contrast information of the luminance edge. Alternatively, one can try to directly estimate the partial derivatives used in the motion constraint equation from the stream of events. The construction of this approach and its related problems are described in Sec. 3.3.4 in the following.

3.3.4 Estimating spatio-temporal continuity using event-sequences

The local spatio-temporal movement of a gray-level function can be estimated by least-squares optimization from a set of local contrast measurements which define intersecting motion constraint lines in velocity space [Lucas and Kanade, 1981]. Given a dense temporal sampling the spatio-temporal gray-level function can be reasonably well captured by a first-order approximation (as summarized in sect. 3.3.1). The key question remains how one could estimate the spatial and temporal derivatives in the constraint equations, $g_x u + g_y v + g_t = 0$ from event sequences generated by the DVS. Events only encode information about the temporal derivative g_t (c.f. eqn. (3.8)). Thus, without additional information it is impossible to reliably estimate g_x or g_y as outlined in the previous Sec. 3.3.3. The derivative of a translatory moving gray level patch, however, generates a unique response in $h := g_t$. Thus, we can apply the motion constraint equation to the function h and solve $h_x u + h_y v + h_t = 0$, instead.

Using two temporal windows $\mathcal{T}_{-2} = (t - 2\Delta t, t - \Delta t]$ and $\mathcal{T}_{-1} = (t - \Delta t, t]$, we can approximate h_t, for example, by a backward *temporal* difference

$$h_t(\boldsymbol{p}; t) = g_{tt}(\boldsymbol{p}; t) \approx \frac{\vartheta}{\Delta t^2} \left(\sum_{t' \in \mathcal{T}_{-1}} e(\boldsymbol{p}; t') - \sum_{t' \in \mathcal{T}_{-2}} e(\boldsymbol{p}; t') \right) , \tag{3.14}$$

with $\boldsymbol{p} = (x, y)^T$ and ϑ denoting the event-generation threshold. The *spatial* derivatives h_x and h_y can be approximated by central difference kernels $[-1, 0, 1]$ and $[-1, 0, 1]^T$, respectively. These can be applied to the function h estimated by integrating over the temporal window $\mathcal{T}$ (e.g. $\mathcal{T} = \mathcal{T}_{-2} \cup \mathcal{T}_{-1}$)

$$h_x(\boldsymbol{p}; t) = g_{tx}(\boldsymbol{p}; t) \approx \sum_{t' \in \mathcal{T}} e(x+1, y; t') - \sum_{t' \in \mathcal{T}} e(x-1, y; t') , \tag{3.15}$$

$$h_y(\boldsymbol{p}; t) = g_{ty}(\boldsymbol{p}; t) \approx \sum_{t' \in \mathcal{T}} e(x+1, y; t') - \sum_{t' \in \mathcal{T}} e(x, y-1; t') . \tag{3.16}$$

Consequently, the resulting flow computation results in a sparsification of responses since stationary edges will not be represented in h. This approach is similar to that of [Benosman et al., 2012] but consistently employs the second derivative instead of mixing the first and second derivatives which cannot work in general.

Note, however, that this approach has multiple issues regarding any real implementation. The most important observation is that when a luminance edge passes a pixel's receptive field of the DVS sensor, the amount of events is in the range of about 10 events (often even smaller, depending on the contrast, speed and luminance conditions; c.f. the event cloud in Fig. 3.3). Thus, huge approximation errors occur for h_x, h_y and especially in h_t (since this is now the second derivative of the original gray-level function g). Furthermore, we can only estimate h_t accurately, if the temporal windows are small enough such that the gray-level edge has not already passed through the receptive field of a target cell at position $\boldsymbol{p}$. This limits the number of events to even less and leads to magnifying the outlined problems even further. Alternatively, one could try to directly approximate the temporal derivative for each event by incorporating the time-span since the last event, i.e.

$$\frac{d}{dt} g(\boldsymbol{p}; t) = g_t(\boldsymbol{p}; t) \approx \frac{\vartheta}{\Delta_W t} e(\boldsymbol{p}, t) \tag{3.17}$$

with $\Delta_W t$ representing the time since the last event generated at $\boldsymbol{p}$. This, however, assumes a constant intensity change since the last event. This, however is certainly not true for the first event because first nothing happens for a long period (i.e. $\Delta_W t$ is too big, because it incorporates a long time in which nothing changes) and then occasionally some change occurs that causes the event, i.e. the estimate will be too small, because $\Delta_W t$ is too big.

3.3.5 Motion estimation using filter banks

As an alternative to considering the LS regression in estimating the velocity tangent plane from the cloud of events, the uncertainty of the event detection might be incorporated directly. At each location, detected events define likelihood distributions $p(e|\boldsymbol{u})$ given certain velocities of the visual scene (estimated by a filter bank, for example). Using Bayes' theorem, we know that for each event $p(\boldsymbol{u}|e) \propto p(e|\boldsymbol{u}) \cdot p(\boldsymbol{u})$. If each velocity is equally likely to be observed without a priori knowledge, i.e. $p(\boldsymbol{u}_1) = p(\boldsymbol{u}_2)$, it holds $p(\boldsymbol{u}|e) \propto p(e|\boldsymbol{u})$ and thus, the velocity $\boldsymbol{u}_{est}$ of the movement that caused event e can be estimated as

$$\boldsymbol{u}_{est} = \mathrm{argmax}_{\boldsymbol{u}} p(\boldsymbol{u}|e) = \mathrm{argmax}_{\boldsymbol{u}} p(e|\boldsymbol{u})\,. \tag{3.18}$$

Thus, we can estimate the velocity from the responses $p(e|\boldsymbol{u_i})$, $i = 1\ldots$ of a filter bank, for example. In addition, a priori knowledge could be incorporated to reduce noise and to increase coherency.

3.4 Event-based motion estimation using spatio-temporal filters

Current knowledge suggests that such distributions are represented by the filter characteristics of the spatio-temporal receptive fields of V1 cells. Thus, we develop a related scheme using filter mechanisms in this section. In addition, we propose to incorporate a stage of subsequent competitive interaction between filter responses as well. This additional stage is suggested to take into account the response non-linearities of initial contrast and motion responses of cortical V1 cells as reported in, e.g., [Carandini et al., 1997, Sceniak et al., 1999] for contrast detection and [Tsui et al., 2010] for spatio-temporal motion detection. The interesting observation is that response modulations occur by features which alone would not elicit a response at the target cell [Carandini and Heeger, 2012]. One proposal is that cell responses are integrated over a neighborhood in the space-feature domain which defines a large-scale pool of cells. This provides a context to modulate the localized feature response of a target cell. The response characteristics of such non-linear normalization mechanisms have been studied in detail in, e.g., [Brosch and Neumann, 2014b]. In addition, as outlined in [Tschechne et al., 2014a], the further incorporation of modulating feedback signals enhances the noisy feature responses and suppresses spurious initial signal detections [McClurkin et al., 1994, Cerda and Girau, 2008, Raudies et al., 2011, Brosch and Neumann, 2014a]. The contribution of such feedback signals might be interpreted as a priori information that serves as instantaneous prior information that helps to maximize the a posteriori estimate of the response likelihood [McClurkin et al., 1994].

This chapter focuses on a model of motion estimation using mechanisms and methodologies introduced in Chap. 2.3. Fig. 3.9 shows an overview of the proposed model. First, physiological foundations are summarized and used to define building blocks in order to generate spatio-temporally tuned filters (Sec. 3.4.1) in model V1 with application of the generic mechanisms introduced in Chap. 2. This includes response normalization, motion integration and modulatory feedback. A second implication of response normalization is presented that shows how early modulation supports the selectiveness to two-dimensional features and thus helps to solve the aperture problem (Sec. 3.4.2). Model MT achieves an integration over an extended spatial extent and thus provides contextual modulation for initial responses at V1, see Sec. 3.4.3. Finally, implications about the interaction of motion and form is elaborated in Sec. 3.5, where the representation of *Motion Streaks* using event-based data is discussed. Finally, Sec. 3.6 presents an architecture that utilizes the sparse representation of visual events for an efficient implementation of the presented algorithm and an assessment of the necessary computational effort.

3.4.1 Initial estimation stage

In this section, we define spatio-temporal filters that are fitted to the physiological findings as reported in [DeAngelis et al., 1995] and [De Valois et al., 2000]. We specifically outline in Sec. 3.4.1.1 how Gabor filters and temporal filters (Sec. 3.4.1.2) are combined to yield a separable spatio-temporal filter.

The authors of [DeAngelis et al., 1995] characterized receptive fields in the geniculostriate pathway and highlighted their resemblance to even and odd components of Gabor functions. They also described receptive fields that are spatio-temporally inseparable. Based on physiological findings, [De Valois et al., 2000] suggested that inseparable filters stem from a combination of various separable components. Cortical V1 cells were tested and strong evidence for the coexistence of populations of cells of two distinct types of receptive fields emerged: One population showed spatio-temporal separable weight functions of either even or odd spatial symmetry. Such cells further show either temporally mono- or bi-phasic response characteristics. The other population of cells was spatio-temporally in-

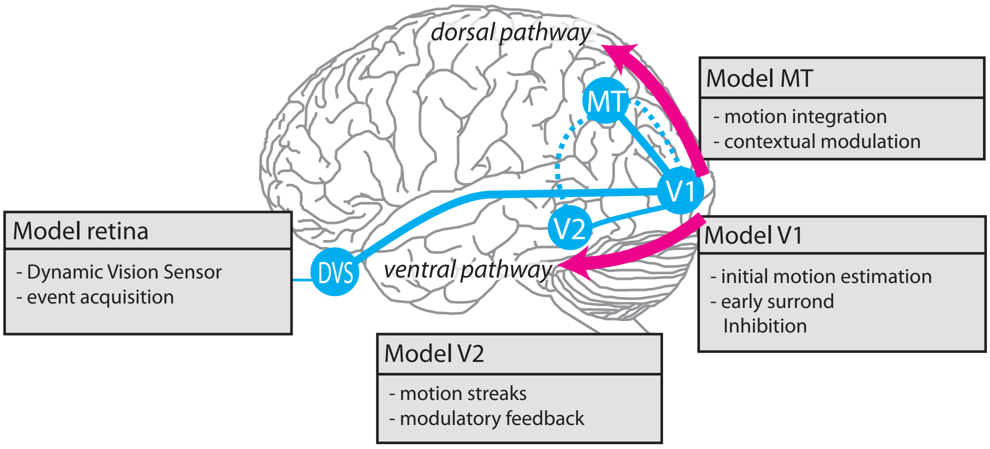


***Figure 3.9: The structure of the proposed model** is inspired by physiological and anatomical findings of cortical structures. Visual input is sampled with model retina sensor. The processing cascade starts at model V1 with initial motion estimation and includes subsequent processing steps in interacting areas V2 and MT along the dorsal and ventral pathways.*

separable showing a receptive field distribution of selectivity that were slanted with respect to the time axis. Similar as in Fig. 3.6, the slant direction corresponds to the spatio-temporal direction selectivity either leftward or rightward orthogonal to the contrast orientation, compare [De Valois and Cottaris, 1998]. De Valois and collaborators further investigated the spatio-temporal response distribution of receptive field profiles from a statistical point of view. Singular value decomposition of the response profiles revealed that spatio-temporal weightings that are separable in space-time are mainly determined by a single principal component of the 2D receptive field. The weightings that were inseparable in space-time are determined by two strong principal components in two dimensions. These components of the second group were itself spatio-temporally separable with spatially out-of-phase components and always composed of pairs of mono- and bi-phasic distributions. This main observation leads us to propose a family of spatio-temporally direction selective filters as illustrated in Fig. 3.10 that are generated by superposed separable filters with quadrature pairs of spatial weighting profiles and mono- respective bi-phasic temporal profiles. This idea of composite assembly of direction-selective cell receptive fields is in the spirit of [Adelson and Bergen, 1985]. However, the details of the model here differ as they are directly driven by experimental data.

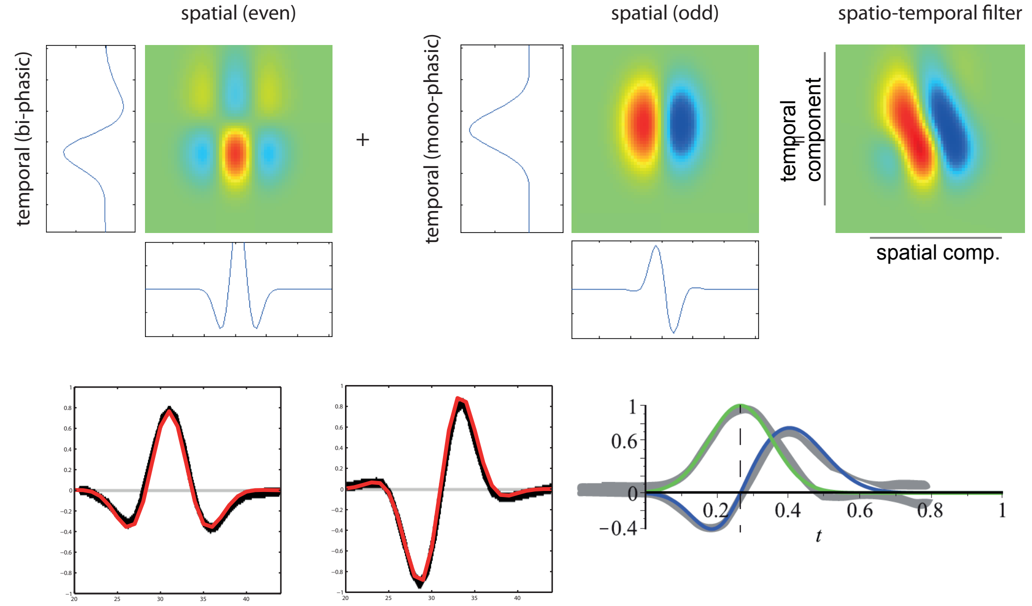


Figure 3.10: Spatio-temporal filters for motion estimation. Top: *Two principal components contribute to the generation of a spatio-temporally inseparable filter that is necessary for motion selectivity. Those components in turn are comprised of simple spatial and temporal characteristics.* Bottom: *Parameters of filters were tuned to fit physiological findings both in the spatial and the temporal domain.*

3.4.1.1 Spatial Gabor filters

To construct the spatial component of the filters illustrated in Fig. 3.10 we define Gabor filters that are fitted to the experimental findings of [De Valois et al., 2000].

$$G_{\sigma,f_x^0,f_y^0}(x,y) = \frac{2\pi}{\sigma^2} \cdot exp\left[-\frac{2\pi^2 \cdot (\hat{x}^2+\hat{y}^2)}{\sigma^2}\right] \cdot exp\left[2\pi i(f_x^0 x + f_y^0 y)\right] \tag{3.19}$$

Those filters are maximally selective for the spatial frequency (f_x^0, f_y^0) and have a standard deviation of σ in local space. This defines the two components $G_{odd} = \Im(G_{\sigma,f_x^0,f_y^0})$ and $G_{even} = \Re(G_{\sigma,f_x^0,f_y^0})$ to construct the filters as defined in 3.4.1.

To construct multiple spatio-temporally tuned filters of different spatial orientation selectivity, we employ a filter-bank of kernels as illustrated in Fig. 3.10, bottom.

3.4.1.2 Mono- and biphasic temporal filters

The temporal components distribute into those with a biphasic temporal component and those with a monophasic temporal component to replicate the experimental data of [De Valois et al., 2000]. We define temporal filter functions T_{mono} and T_{bi} that contribute the mono- and bi-phasic temporal component, respectively. To generate a fit with experimental data, we define T_{mono} as a Gaussian function

$$T_{mono}(t) = G_{\sigma_{mono},\mu_{mono}}(t) \tag{3.20}$$

and the bi-phasic component as a difference of two spatially shifted Gaussian functions:

$$T_{bi}(t) = -s_1 \cdot G_{\sigma_{bi1},\mu_{bi1}}(t) + -s_2 \cdot G_{\sigma_{bi2},\mu_{bi2}}(t) \tag{3.21}$$

with the unnormalized Gaussian function

$$G_{\sigma,\mu} = exp(-\frac{(t-\mu)^2}{2\sigma^2}). \tag{3.22}$$

When the experimental findings are incorporated, it is only necessary to choose a value for μ_{bi1}. All other parameters can be inferred according to the experimental data from [De Valois et al., 2000]:

- The bi-phasic scaling factors $s_{1/2}$ are adapted to the minimum and maximum values of the experimental data relative to the maximum value of the monophasic kernel (which is one), i.e. $s_1 = 1/2$ and $s_2 = 3/4$.
- A good fit with the experimental data reported in [De Valois et al., 2000] is achieved by setting $\mu_{bi2} = 2\mu_{bi1}$.
- The standard deviations σ_{mono} and σ_{bi1} are chosen such that the Gaussians are almost zero for $t = 0$, i.e. $\sigma_{mono} = \mu_{mono}/3$, $\sigma_{bi1} = \mu_{bi1}/3$ (3σ–rule; 99.7% of the values lie within three standard deviations of the mean in a normal distribution).
- The standard deviation of the second Gaussian of the bi-phasic kernel is about 3/2 of that of the first, i.e. $\sigma_{bi2} = \frac{3}{2} \cdot \sigma_{bi1} = \frac{1}{2} \cdot \mu_{bi1}$.
- The mean of the mono-phasic kernel μ_{mono} is given by the zero-crossing of the biphasic kernel, i.e. $\mu_{mono} = \frac{1}{5} \cdot \left(1 + \mu_{bi1} \cdot \sqrt{36 + 10 \cdot \ln(s_1/s_2)}\right)$.

Figure 3.10 illustrates that these settings result in a good fit of the temporal filters with the experimental data reported in [De Valois et al., 2000].

3.4.1.3 Combined spatio-temporal filter

The full spatio-temporal filter F is defined according to the scheme of Fig. 3.10, i.e. by the sum of two products consisting of the odd-spatial G_{odd} and the monophasic temporal T_{mono} and the even-spatial G_{even} and the bi-phasic temporal filter T_{bi}:

$$F(x, y, t) = \Im(G_{\sigma, f_x^0, f_y^0}(x, y)) \cdot T_{mono} + \Re(G_{\sigma, f_x^0, f_y^0}(x, y)) \cdot T_{bi}(t) \tag{3.23}$$

The bottom-up input filter response is generated by:

$$I(x, y, t)_{ex} = \sum_{i,j,t'} e(x, y, t) \cdot F(x, y, t) \tag{3.24}$$

3.4.1.4 Response normalization

We incorporate a stage of divisive normalization in our model, which extends the linear model to include mutual shunting inhibition among a large number of cortical cells. It acts divisively upon the activation level of target neurons and its effect in the model is to normalize the linear responses by a measure of stimulus energy over a population of cells in the neighborhood of a neuron. The normalization step is part of the processing cascade as introduced in Sec. 2.3.2.

$$\dot{p}_i = -\alpha_p p_i + (\beta - p_i) \cdot I_i - p_i \cdot q_i \tag{3.25}$$

$$\dot{q}_i = -\alpha_q q_i + \sum_{j \in \mathcal{N}} c_j g(p_j) \tag{3.26}$$

3.4.1.5 Results

In order to investigate the computational mechanisms of the initial filter-based motion detection mechanism we have conducted a number of experiments which were designed to demonstrate the desired functionalities. The model architecture consists of a layer of cortical model columns which realize an input filtering that is followed by divisive normalization. The initial estimation stage is probed with artificial and real stimuli that contain rotational and translational motion. We conducted a series of experiments to validate the modeling approach and its theoretical properties. The parameters of the spatio-temporal filters were chosen such that they fit the experimental data as reported in [De Valois et al., 2000] (up to scaling), namely $\mu_{bi} = 0.2$ for the temporal filter components, and $\sigma = 25, f_0 = 0.08$ for the spatial filter components. The parameters of the normalization mechanism were set to $\beta = 1, \alpha_p = 0.1, \alpha_q = 0.002$,c_j resemble to coefficients of a Gaussian kernel with $\sigma = 3.6$, and $\Psi_I(x) = \Psi_q(x) = max(0, x)$ denotes a rectifying transfer function. At each location the initial motion estimation creates a population code of length N with each entry corresponding to the response of a spatio-temporal filter with motion direction selectivity θ_k. For visualization purposes (in Fig. 3.12), the velocity components $u_{\mathbf{p}}$ and $v_{\mathbf{p}}$ are inferred from the initial responses $I_{\mathbf{p};k}, k \in 1, ..., N$ at each location $\mathbf{p}$ by summing them up according to

$$\begin{pmatrix} u_{\mathbf{p}} \\ v_{\mathbf{p}} \end{pmatrix} = \sum_{k=1}^{N} I_k \cdot \begin{pmatrix} cos(2\pi(k-1)/N) \\ -sin(2\pi(k-1)/N) \end{pmatrix}, \tag{3.27}$$

effectively implementing a local vector addition of component estimates in a neural population. The visualized results demonstrate that the filter based approach constitutes a robust tool to compute estimates of contour motion, i.e. locations of apparently moving contrasts and object boundaries [Barranco et al., 2014].

For an intuitive visualization of results, we use a color coding for estimated motion directions, see Fig. 3.11. Correct motion direction is estimated for translational (see Fig. 3.12, A-C) and rotational (see Fig. 3.12, D;E) motions. Along elongated

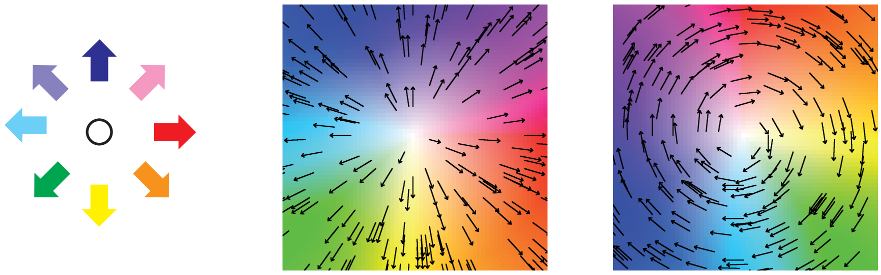


Figure 3.11: Motion color code for visualization*. Motion directions are indicated using a color code.* Left: *Arrows indicating different motion directions are colored to allow an easy perceptual distinction between different motion directions.* Middle and right*: Coloring examples for expanding and rotational motion, respectively.*

structures, the algorithm only estimates normal flow that points perpendicular from the orientation of the structure. This effect is known as the aperture problem and the following section will highlight detailed aspects of this effect.

3.4.2 Response normalization and 2-d-feature selectivity

The preceding chapter has introduced a mechanism of motion estimation using visual events from a neuromorphic sensor architecture, and showed how estimates are normalized to increase their selectivity and dynamic range.

These estimates represent a local estimation of motion from events that lie withing their receptive fields. It is long known that such local estimates of motion are ambiguous when there is only a straight contour inside the receptive field [Wallach, 1935, Wuerger et al., 1996, Nakayama and Silverman, 1988]. In this case, only the motion perpendicular to the contour orientation can be measured, because the component of parallel motion is unknown. This *aperture problem* of motion estimation has challenged neurophysiologists, psychophysisists and modelers for decades, because the precise neural mechanisms that solve the aperture problem in visual cortex are still largely unknown.

Ongoing debates still consider mainly two different concepts of how the visual system solves the problem. Some experimental results led to the idea of the *selectionist* concept. This concept integrates local measurements of motion only at positions where they are estimated reliably, like at intrinstic two-dimensional features like corners, junctions or line endings, thus *selecting* correct estimates. There have been various suggestions about how the visual system achieves this task. [Loeffler and Orbach, 1999] proposed separate streams for complex cells and those tuned to end-stop positions. On the other hand, the *integrationist* concept proposes a nonlinear integration of local components to achieve a reliable measurement [Simoncelli and Heeger, 1998, Rust et al., 2006]. Recent experimen-

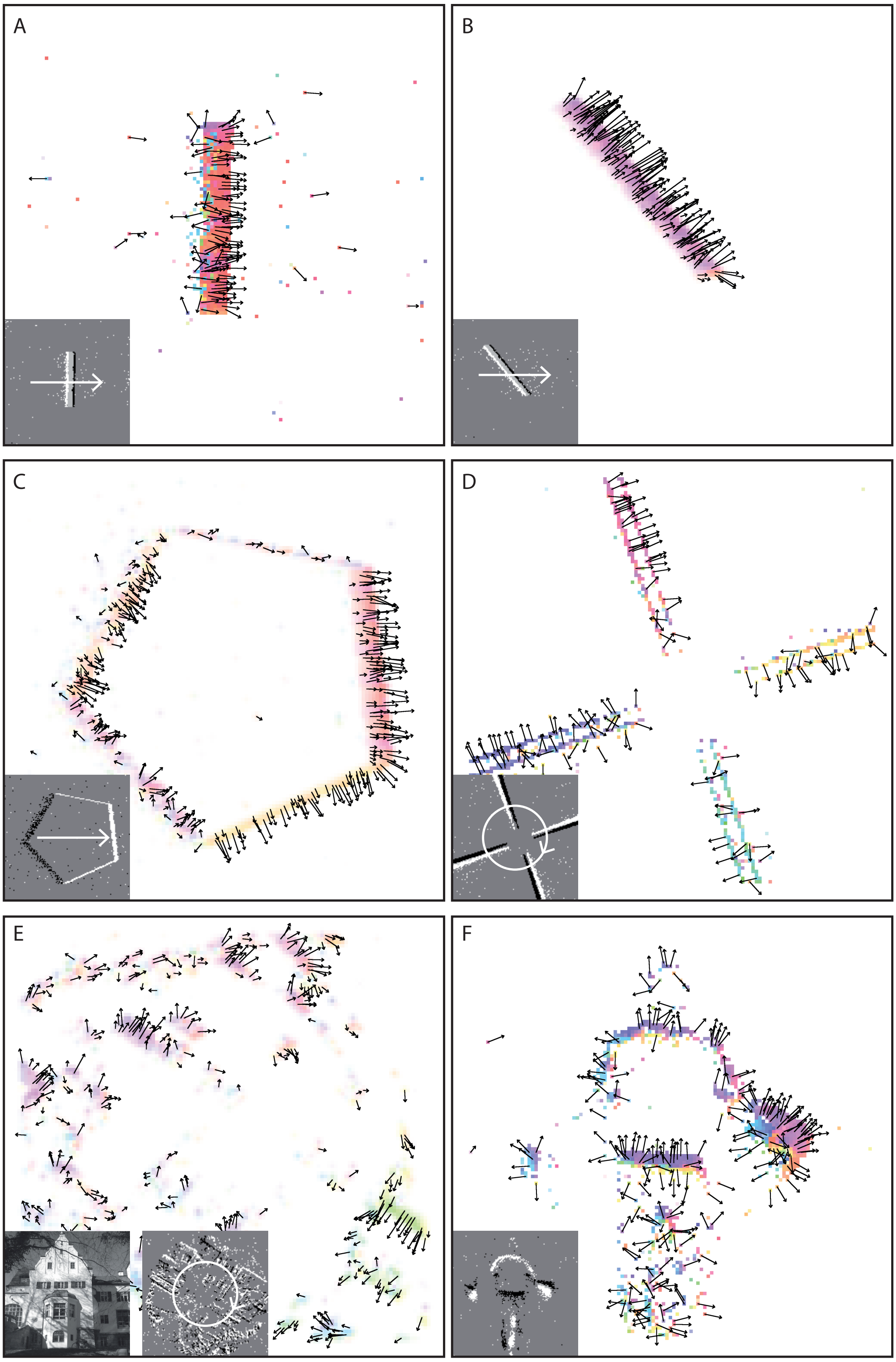


Figure 3.12: Results of initial estimation stage. *Recorded sequences contained different motions, in reading order: Linear motion of a vertical bar, an oblique bar and a pentagram, rotational motion of two orthogonal axes and a photography, articulated motion of a jumping person. The response strengths (and thus vector lengths) are normalized.*

tal data suggests that probably both concepts have their contribution to explain physiological data [Beck and Neumann, 2010].

Neural studies revealed that the aperture problem is solved at visual area MT. Here, neurons change their response from a representation of normal flow to a representation of correct flow at elongated contours after time. The observation that the time they need to change is a function of contour length[Pack and Born, 2001, Born, 2001, Born et al., 2010], which has already motivated functional models based on the mechanisms that were presented earlier [Beck and Neumann, 2011].

The authors of [Tsui et al., 2010] discuss a possible mechanism of interaction at area V1 that might already contribute to the solution of the aperture problem (The aperture problem and its impact on local motion estimation is shown in Fig. 3.13). With reference to [Heeger, 1992, Simoncelli and Heeger, 1998] and [Rust et al., 2006] they propose a mechanism to realize the necessary non-linear interactions and normalization steps in the early area V1. This idea is adapted by [Born et al., 2010] who argued that the responses can be explained by a single input stream if the responses are modulated by surround mechanisms. In both cases the integrated responses at MT level show a reduced influence of ambiguous normal flow estimates. As suggested in [Tsui et al., 2010], normalization can help to suppress responses at ambiguous parts of a contour (i.e. the inner parts of an extended contrast or line) and to enhance responses at line ends or sharp corners (c.f. Fig. 3.13).

The key question of how responses of V1 cells - that, due to their small receptive fields - are subject to suffer from the aperture problem, can be modulated within V1 in a way such that they show increased response toward intrinsically 2-dimensional features. The aim is to investigate if the surround modulation can suppress the influence of spurious estimates of normal flow at the line endings, and to what level the estimates of normal flow can be reduced in the extent of an moving elongated contrast.

Our proposed model is inspired by earlier works of [Born et al., 2010] and [Tsui et al., 2010], who extended the model of [Simoncelli and Heeger, 1998]. Instead of the motion energy detector that [Adelson and Bergen, 1985] used in the earlier investigations we employ the output of the initial stage of event-based motion processing and apply a nonlinear normalization mechanism [Carandini et al., 1999] to local V1 estimates. The impact on motion representation is empirically demonstrated by replicating experimental settings.

The normalization step follows the same canonical structure as presented in Sec. 2.3.2. Activations representing estimates to motion p are modulated by activations integrated from contextual surround:

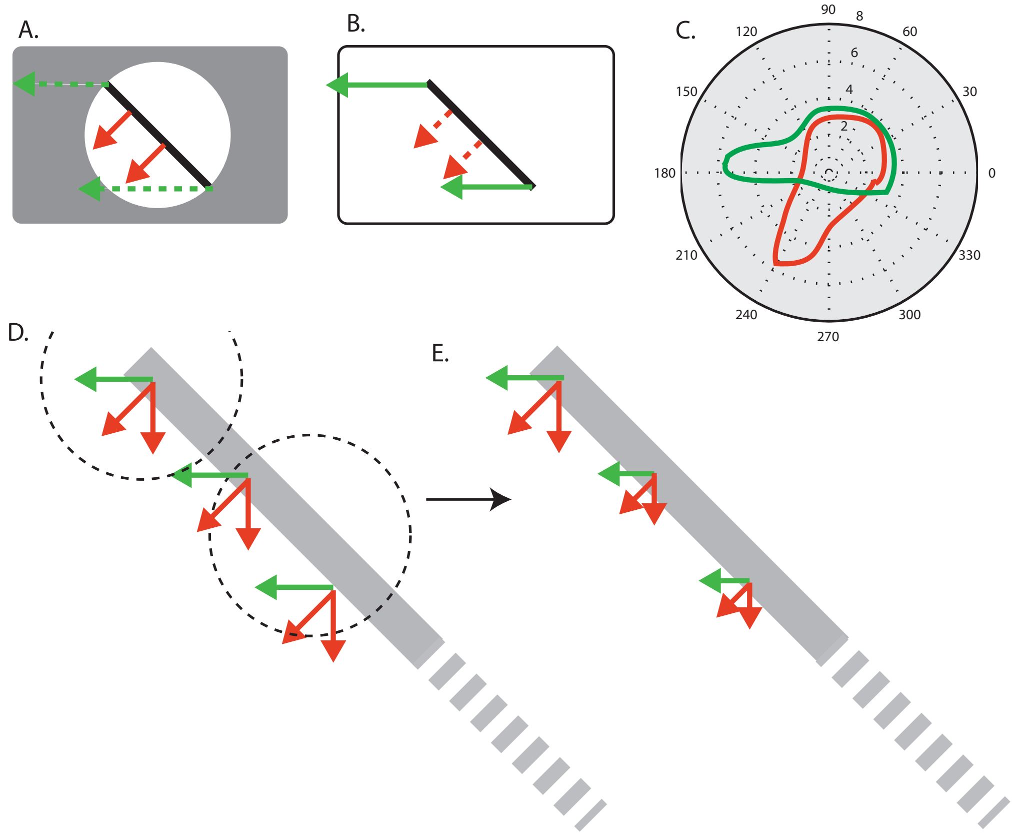


Figure 3.13: The aperture problem in vision. Top: *For movements of oblique elongated edges, only normal flow can be estimated (A) unless two-dimensional features like line ending are available (B). For an elongated edge that is slanted with respect to its motion directions, this yields spurious estimates of motion direction, as illustrated on the polar plot (C).* Bottom: *V1 surround inhibition reduces the influence of spurious estimates along an elongated contrast (D and E).*

$$\dot{p}_i = -\alpha_p p_i + (\beta - p_i) \cdot I_i - p_i \cdot q_i \tag{3.28}$$

$$\dot{q}_i = -\alpha_q q_i + \sum_{j \in \mathcal{N}} c_j g(p_j) \tag{3.29}$$

To demonstrate the effects of this mechanism, we generate a test stimulus as proposed in [Tsui et al., 2010]. The test stimulus is a bar that is moving into one direction, but the orientation of the bar is rotated 45° counter-clockwise against the motion direction. Along the contour of the bar local receptive fields of the V1 motion estimation will represent normal motion, while the true motion direction has 45° difference. In the simulations we use eight discrete motion directions evenly distributed between 0..360°.

The effects of the normalization mechanism are first demonstrated at a V1 level. Initially, without the modulation, V1 cells respond at line endings as well as at the elongated contour. The direction at the contour is the normal flow. When the normalization and surround modulation is activated, the response for one-dimensional features along the extent of the bar is highly reduced, while at the same time the sensitivity towards the real motion direction if increased at the line endings. See Fig. 3.14 for an illustration of the results.

We stated earlier that the aperture problem is solved at area MT. We will now show how the early modulation already contributes to a reduction of spurious measurements at MT level without a dedicated mechanism that is located at MT. The size of an integrating MT neuron is chosen so that it integrates the complete stimulus [Born, 2001]. The distribution of estimated motion hypotheses is presented using circular plots that indicate the relative occurrence of motion directions. A model MT cells integrates the resulting responses with a temporal response function. The MT cell used in these experiments is tuned to *leftward* motion, and all moving oblique bars are presented to that cell. The dominance of the normal flow direction along the contour causes the response maximum of that leftward tuned indicating the normal flow direction, which is 45° different from the true motion direction. The results with early inhibition indicate that the tuning is now more biased in direction of the true motion direction. The mean values of integrated motion direction in the example change from 224° to 201° (D,E) and 270° to 233° (F,G). These results indicate that normalization significantly improves the histograms to point more into the true motion direction (Fig. 3.14, blue lines).

While the initial motion estimates are highly biased by the normal flow estimates at the longer side of the bar, a surround modulation helps to reduce the response strengths at positions along the bar where only normal flow can be estimated. Simultaneously, the tuning towards the real motion direction is improved. The proposed mechanism is inspired by findings of [Tsui et al., 2010] and reveals a potential mechanism of how selectivity towards 2-dimensional features can be achieved at a level where only local estimates are available. Our model does not require explicit feature tracking performed outside of the area (say, in the form channel) and relies only on the local estimates. This contributes to a solution of the aperture problem at MT level, even if a full solution cannot be achieved with only this mechanism. Earlier work has already demonstrated that a combination of mechanism from the selectionist and the integrationist concept are likely to contribute to the solution of the aperture problem [Beck and Neumann, 2010]. Our proposed early mechanism contributes another computational component to the solution of this perceptual phenomenon.

3.4.3 Model area MT

In the previous section we demonstrated how direction-selective neurons in model area V1 may encode spatio-temporal changes of visual patterns. Such cells respond coarsely to movement of gray-level structures given a direction ϕ orthogonal to their orientation selectivity and over a broad range of speeds. It has already been demonstrated how local normalization effects at V1 contribute to the solution of the aperture problem at an MT level. In this section, we show contributions of the MT stage to the representation of motion. First, contextual influences on the initial motion estimation at an V1 level is demonstrated by incorporating modulatory feedback connections from MT to V1.

3.4.3.1 Motion integration and re-entry

Supported by physiological investigations, initial motion responses of V1 are integrated by cells in area MT which obey an increased selectivity to direction and speed [Born and Bradley, 2005]. The estimations performed initially provide an initial representation of motion, but these estimates suffer from locally spurious influences like noise or outliers. This motivates that these initial estimates are integrated from a larger surrounding context to yield a better robustness towards these influences. Such an integration of initial estimates is available at dorsal area MT. Here, cells show an increased selectivity towards motion direction and speed. Higher in the visual pathway, tunings appear to be more and more specialized to increasingly complex motion patterns. Model MT with its increased spatial extent provides contextual input to lower visual areas [Ungerleider and Haxby, 1994].

In the following we further develop the feed-forward sweep of the motion processing by incorporating the model area MT, where initial estimates become integrated to build feature representations of higher complexity along the dorsal pathway [Ungerleider and Haxby, 1994, Born and Bradley, 2005]. We incorpo-

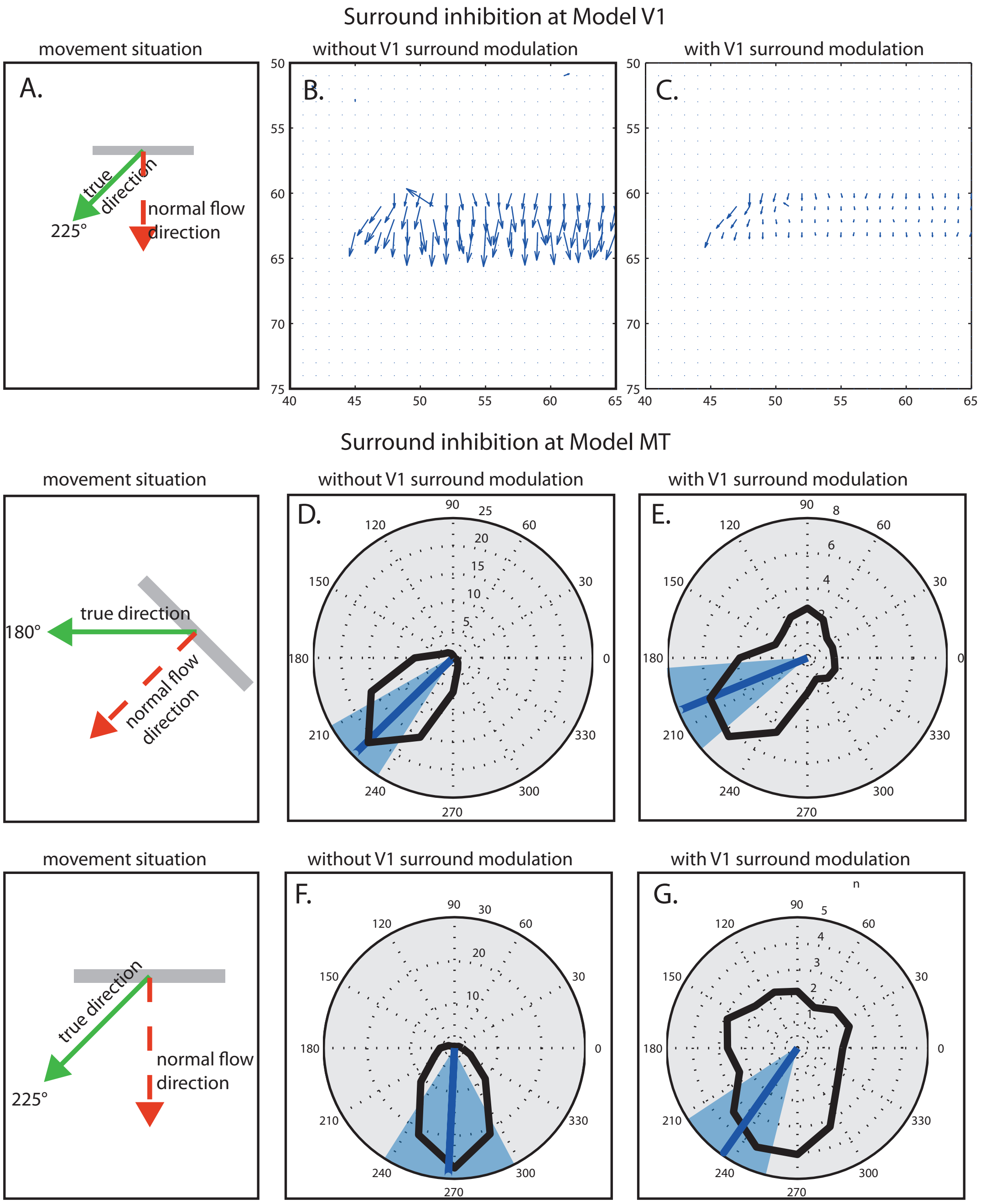


Figure 3.14: Effect of surround inhibition at V1 and MT. Top: *An oblique bar moved with 45 difference to its orientation (A). Local estimates at V1 elicit maximum response for normal flow direction (B.). With local inhibition, the responses along the line are reduced (C).* Bottom: *Integrated responses at model MT level for two examples. Without surround inhibition, integrated responses represent a motion into direction of normal flow, statistically visualized in polar plots. (D,F). Mean (thick line) and standard deviation (transparent overlay) are indicated. When surround inhibition is incorporated, the representation of motion is biased towards the true motion direction, and the standard deviation of motion hypotheses is increased (E,G).*

rate a stage of integrating early motion responses from model V1 using circular receptive field weighting functions Λ with Gaussian profile but larger spatial integration size over a neighborhood approximately five times the size of V1 filters. A pool normalization is incorporated that acts divisively upon the activities. The responses are formally calculated by the following mechanism:

$$\dot{r}^{MT}_{\phi,s} = -\alpha r^{MT}_{\phi,s} + \sum_{i,j,\theta} r^{V1}_{\theta\phi}(i,j) \cdot \Lambda_{xy,ij} \Phi_s - r^{MT}_{\phi,s} \cdot q^{MT} \tag{3.30}$$

$$\dot{q}^{MT} = -q^{MT} + \beta \sum_{i,j,(\phi,s)} r^{MT}_{\phi,s}(i,j) \cdot \Lambda^{pool}_{xy,ij}. \tag{3.31}$$

We demonstrate the effects of motion integration and re-entry of modulatory signals. The upcoming mechanisms inherit results from [Bayerl and Neumann, 2007] and [Beck and Neumann, 2010] who have investigated in motion processing and shown how feedback from higher connections can modulate and provide input to earlier stages. To demonstrate the effect of modulatory feedback, be probe our MT model with stimuli from the intial motion estimation stage. The hypotheses are integrated by applying the integration principle of Eq. 3.30 and 3.31. The effect of the contextual, modulatory feedback are presented in Figs. 3.16 to 3.19. The initial representation of motion hypotheses with a resolution of 128 × 128 is spatially sub-sampled in MT with a factor of 0.25, resulting in an MT resolution of (32 × 32). The contribution of V1 positions to adjacent MT cells is weighted with a Gaussian profile using the distance between the center of the position at V1, the centers of surrounding MT cells and a deviation of $\sigma^{MT} = 2.5$.

The integrated activations at the model MT stage now serve as modulatory feedback for the preceding visual area, namely V1, where the initial motion estimation takes place. By the modulator application, the initial estimates profit from the contextual information gathered by larger receptive field sizes. As a result, the modulated representation at V1 is spatially smoothed and the amount of outliers is reduced, leading to a much clearer representation of the initial estimates. Again, this methodology is founded on on the generic mechanisms introduced in Chap. 2.3.

At MT, the apparent spatial resolution is highly reduced and at the same time the smoothness is increased. This already removes spurious estimates, outliers and noise and maintains only the dominant motion of a local surround.

While the methodology is already presented in earlier works, this work contributes the implementation of motion MT for event-based input. Implementation aspects are elaborated in Sec. 3.6.

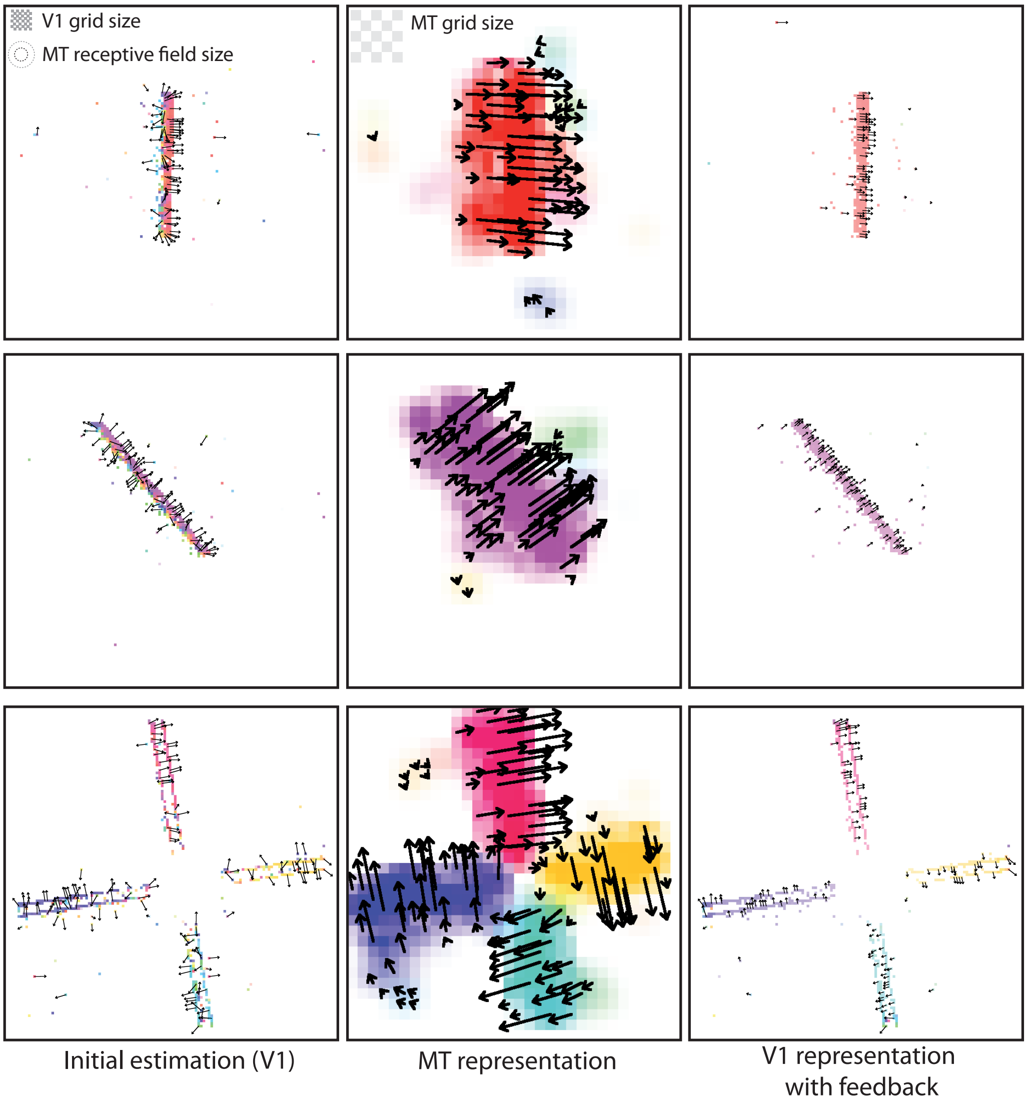


Figure 3.15: Effects of contextual motion integration and re-entry. Top: *Comparison of V1 and MT representation resolution and MT integration field size.* Depictions, left column: *Initial estimates at model V1.* Middle: *Spatially sub-sampled responses at model MT.* Right: *With contextual feedback from MT to V1, the contribution of outliers at V1 is highly reduced.*

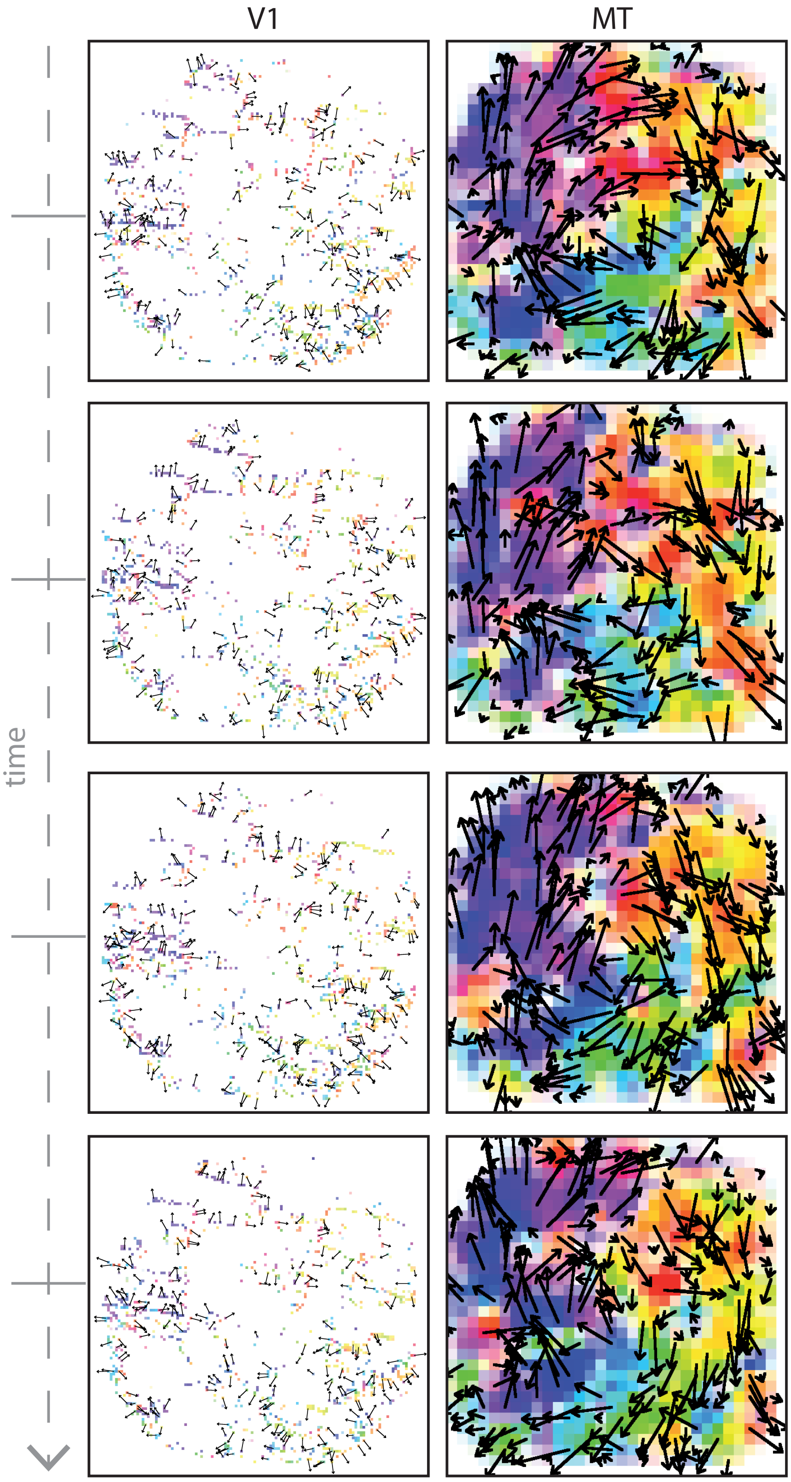


Figure 3.16: Model V1 (left) and MT (right) representations (without contextual modulation). *Snapshots from a sequence with rotating motion with time increasing downwards. Without a modulatory connection between MT and V1, representations are cluttered with large amounts of noise (compare succeeding Fig. 3.17).*

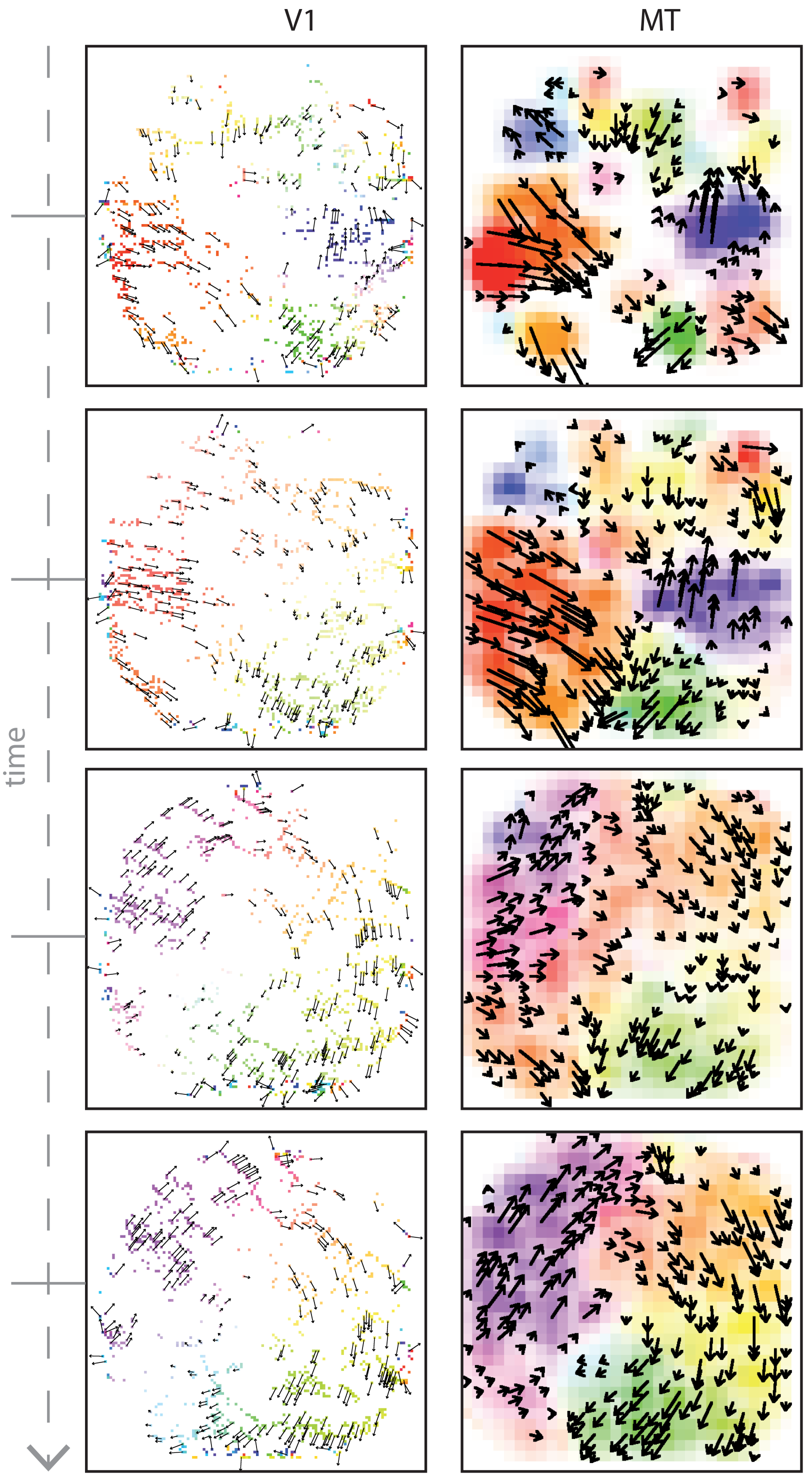


Figure 3.17: Model V1 (left) and MT (right) representation with contextual modulation. *With a modulatory interaction between V1 and MT the representation in both areas show less noise and more homogeneous regions that reflect the true motion (rotation) more adequately (compare previous Fig. 3.16).*

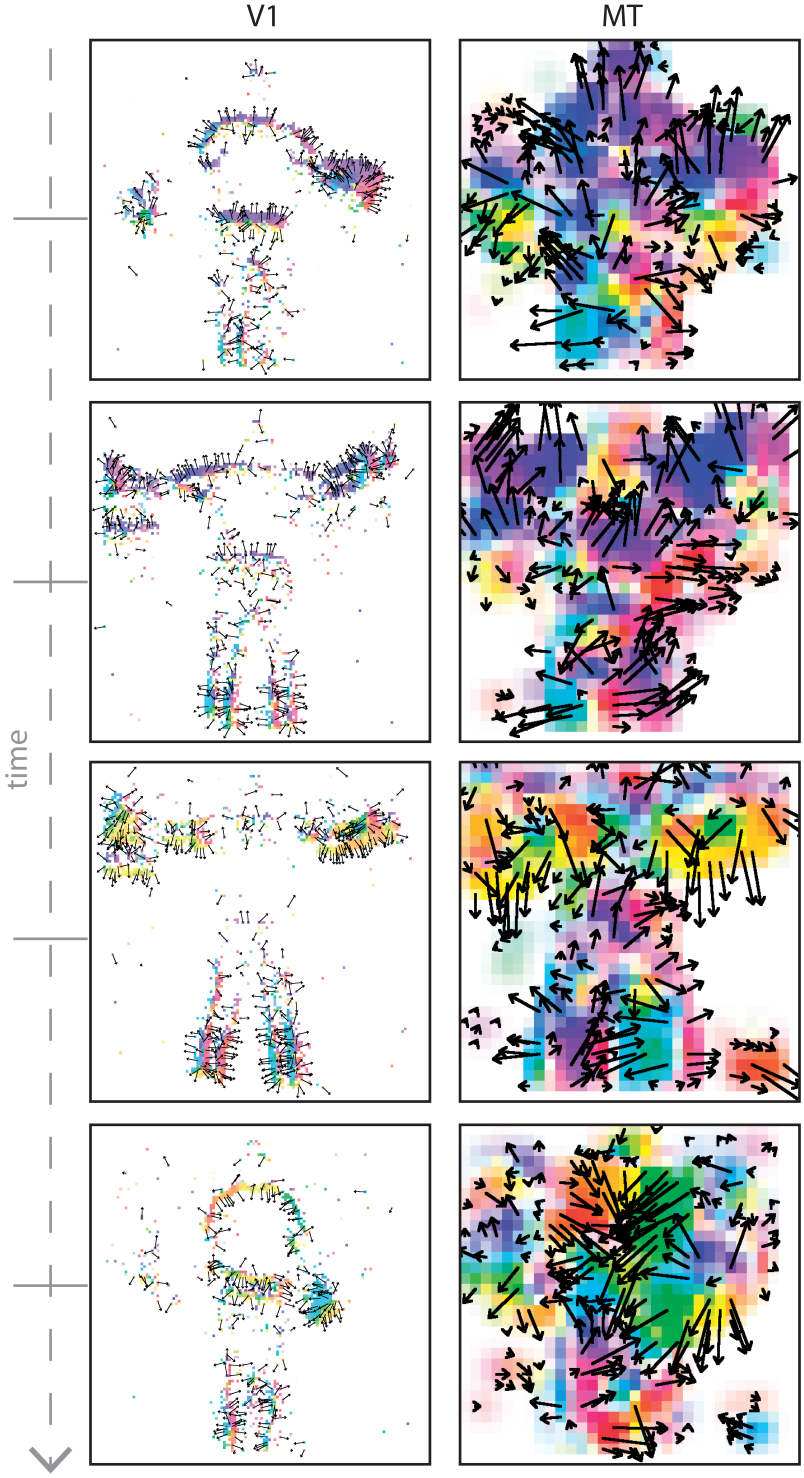


Figure 3.18: Model V1 (left) and MT (right) representations (without contextual modulation). *Snapshots from a sequence with articulated motion with time increasing downwards. Without a modulatory connection between MT and V1, representations are cluttered with large amounts of noise (compare succeeding Fig. 3.19).*

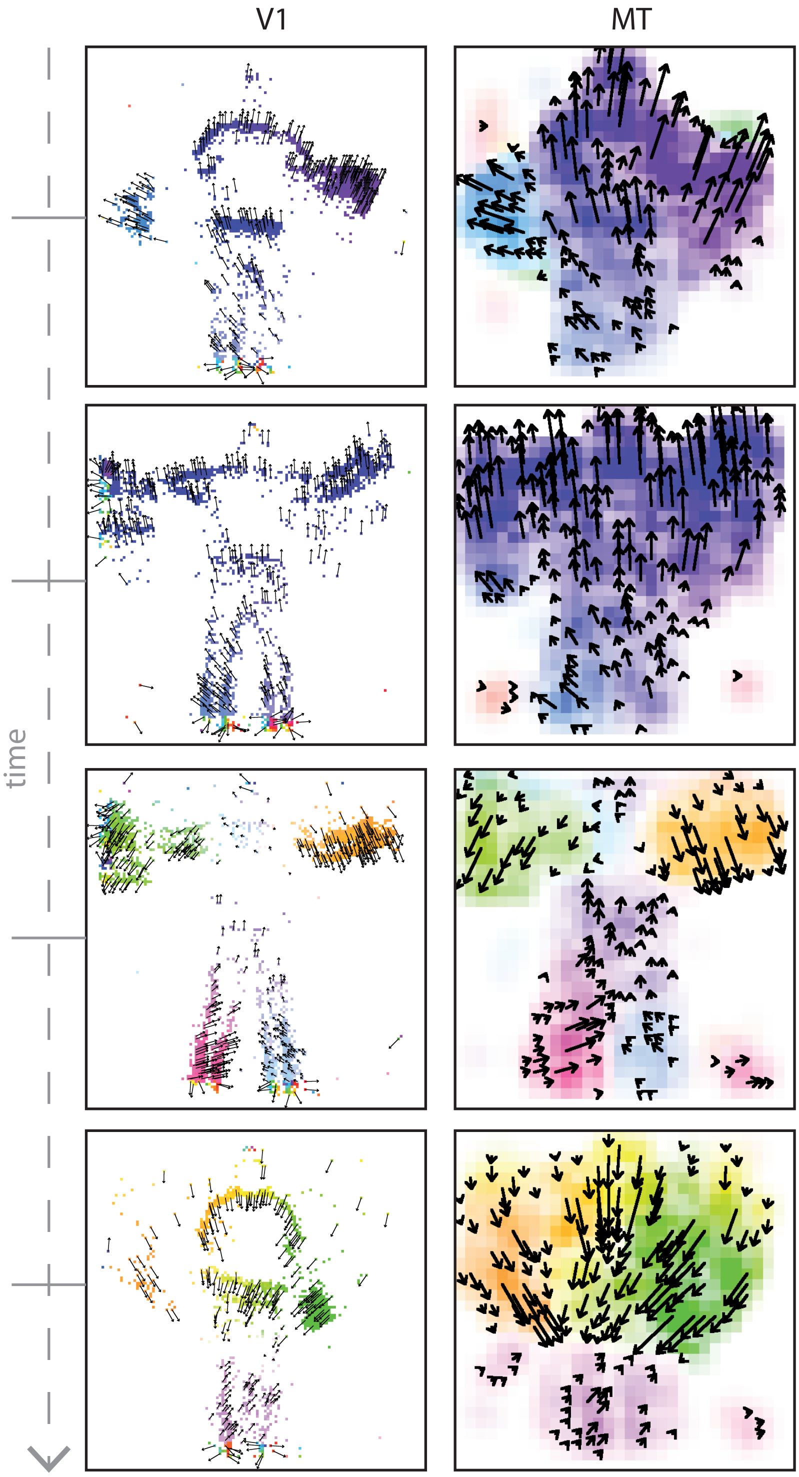


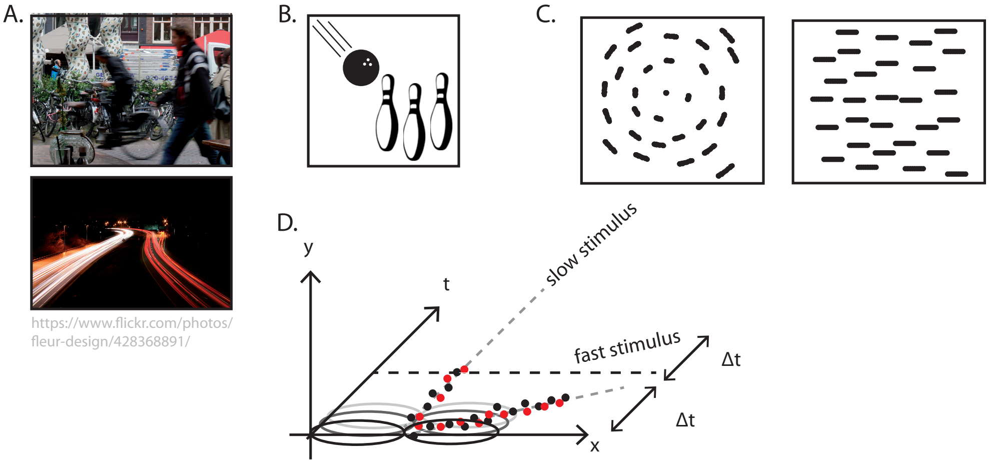
Figure 3.19: Model V1 (left) and MT (right) representation with contextual modulation. *With a modulatory interaction between V1 and MT the representation in both areas show less noise and more homogeneous regions that reflect the true motion (rotation) more adequately (compare previous Fig. 3.18).*

3.5 A spatial code for motion - *motion streaks*

In the previous Sec. 3.4.1 we demonstrated how direction-selective neurons in model area V1 may encode spatio-temporal changes of visual patterns and thus allow an initial estimation of motion. This sensitivity is provided by cells that have a tuning function which is spatio-temporally inseparable and that are thus assigned to the class of *directionally selective* cells. Besides these motion-tuned cells found in visual cortex, a significant amount (about 20 percent) of cells only show a tuning to static features [DeValois et al., 1982b]. They provide input to subsequent areas along the ventral pathway and are thus usually counted to be responsible for static form processing. Can those *non*-directionally tuned cells also contribute to the estimation of motion direction?

Evidence from perceptual psychophysical investigations suggest that under some conditions the visual system utilizes responses of the form pathway as a spatial code for local motion direction [Burr, 2000, Geisler, 1999]. The authors of [Geisler, 1999] introduced the comprehensive concept of *Motion Streaks* (sometimes called *speedlines* or *motion blur*) which basically describes the fact that a quickly moving stimulus leaves a smeared trail of integrated activation along its movement path due to the temporal integration time of the photoreceptors. When visual stimuli move over the retina, they activate juxtaposed luminance receptors in quick succession. In human vision, the temporal integration time of those receptors is about 120 ms [Barlow, 1958]. For slowly moving stimuli, this prolonged integration has no or only little effect. Rapidly moving objects, however, leave a trail of activations across the visual field. Under normal viewing conditions, these activations are effectively suppressed and not visible [Wallis and Arnold, 2009, Martin and Marshall, 1993]. In [Geisler, 1999] the author suggested that such motion streaks can help the visual system during motion estimation. With V1 neurons activated by quickly moving stimuli, they effectively provide a spatial code of motion direction. The fact that this occurs only for fast speeds supports the idea that this mechanism could efficiently extend the range of detectable motion in the V1 stage, where the tuning to motion speed of spatiotemporally tuned receptive fields is limited to a maximum detectable speed.

The precise neural mechanisms underlying the generation of *Motion Streaks* representation are still to discover, yet there have been numerous investigations along different tracks. The integration time of photoreceptors is already described in [Barlow, 1958]. During normal vision, however, the visual perception does not appear smeared, suggesting that there are active mechanisms that sharpen the perception of moving objects [Martin and Marshall, 1993, Bedell et al., 2010]. A dependency of this effect to the speed of the presented objects has been described in [Burr and Morgan, 1997]. The insight that the mechanisms could act as an perceptual aid for perceiving motion direction was hypothesized in [Geisler, 1999], as stated earlier. Later, investigations provided strong evidence that the human



https://www.flickr.com/photos/
fleur-design/428368891/

Figure 3.20: Concept of motion streaks. A. *Fast motions cause motion blur, which is suppressed under normal viewing condition.* B. *Form features can provide a compelling impression of motion in static images.* C. *Generation of motion streaks in conceptual illustrations for rotation and translatory motion using dot patterns.* D. *Event-based processing can elicit responses for cells with receptive fields resembling complex form cells when events are temporally and spatially integrated.*

visual system exploits motion streaks for the discrimination of motion direction [Burr and Ross, 2002]. In [Wallis and Arnold, 2009] the authors reported that the psychophysical effect of motion-induced blindness (MIB) might be an effect that is driven by an active mechanism that suppresses the blurred perception of moving forms. The search for physiological evidence that supports the theory of motion streaks involved experiments using binocular rivalry and masking [Apthorp et al., 2009, Apthorp et al., 2010, Pavan et al., 2013] until finally neural evidence for the presence of motion streaks of fast moving objects has been found using fMRI [Apthorp et al., 2013].

In a nutshell, there is strong evidence that motion streaks contribute to a significant degree to the perception of motion, but the neural mechanisms how a visual motion input elicits responses in the form channel remain unclear. In the following, we will demonstrate how motion streaks occur in the form channel as a consequence of the mechanisms used for event-based processing. Event-based processing allows the representation of motion streaks in one common architecture, without the introduction of special components that are exclusively dedicated to motion streaks. We will demonstrate how motion streaks are represented and how an interaction of motion and form is realized using mechanisms of the canonical model introduced in Sec 2.3.2.

3.5.1 Methods

The sensitivity for motion streaks in our model is elicited by a separate spatial and temporal tuning function. Along the ventral path, cells with sensitivity for elongated edges can be modeled using two elongated receptive subfields with multiplicative connection. An activation of these V2 complex cells is evoked when fitting visual input is simultaneously presented in their receptive fields. In our model, these complex cells are activated by temporally integrated visual events generated by the DVS, when fast objects leave a trail of events along their motion direction. This leads to a representation of motion in the form channel by complex cells that are not tuned to visual motion *per se*.

In the following we will describe the design of spatial and temporal integration components. We will then show how the response of these filters represents motion using a set of test stimuli.

3.5.1.1 Spatial integration

The spatial integration functions resemble receptive fields of cells in visual cortex area V2. Figure-eight shaped integration fields integrate static spatial features over an extended spatial area. Due to their properties to integrate over a short period of time, activation of them is also elicited by visual events caused by fast motions. In our model the receptive fields are comprised of two elongated Gaussian-shaped integration kernels that are combined into a figure-eight shape [Peterhans and von der Heydt, 1991, Neumann et al., 2007]:

$$r^{streaks}_{\theta,\text{spatial}}(\mathbf{x}) = (I(\mathbf{x}) * \mathcal{N}_{\theta,\sigma_1,\sigma_2,+\mathbf{q}}(\mathbf{x})) \cdot (I(\mathbf{x}) * \mathcal{N}_{\theta,\sigma_1,\sigma_2,-\mathbf{q}}(\mathbf{x})) \tag{3.32}$$

The orientation selectivity is defined by the proper rotation, θ. $\mathcal{N}$ denotes a rotated Gaussian kernel:

$$\mathcal{N}_\theta(x,y) = \frac{1}{2\pi\sigma^2} exp\left(-\pi\left[\frac{(\hat{x}-x_0)^2}{\sigma_1^2} + \frac{(\hat{y}-y_0)^2}{\sigma_2^2}\right]\right) \tag{3.33}$$

3.5.1.2 Temporal integration

The modeled V2 cells do not produce a response unless enough input stimuli are available in both subfields. For the representation of motion streaks, temporal integration of events provides input for the subfields. For the design of this temporal integration, we are motivated by the integration times of receptive fields from physiological measurements. Receptive fields of the model neurons incorporate a temporal tuning function that models the behavior of photoreceptor cells in the

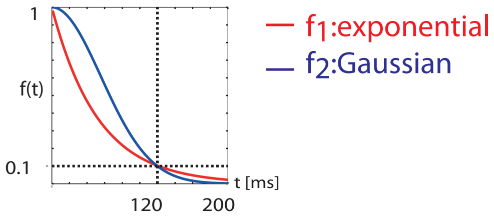


Figure 3.21: Integration strength functions for events in motion streak processing. *We employed two options for a temporal decay, and exponential and Gaussian drop of weight as a function of time.*

retina, which integrate visual events for about 100ms (comp. [Barlow, 1958, DeValois et al., 1982a, DeAngelis et al., 1995]. Events caused by quickly moving objects will provide enough activation across the spatial integration window of the cell during that period and thus yield a representation of that motion orientation. We considered two options of slightly different temporal integration functions. f_1 has the shape of an exponential decay function, while f_2 is Gaussian-shaped.

$$f_1(t) = exp(-\frac{t}{\sigma_1^2}) \tag{3.34}$$

$$f_2(t) = exp(-\frac{(\mu - t)^2}{\sigma_2^2}) \tag{3.35}$$

Recent events are weighted using one of the functions, resulting in a effective current streak snapshot at time T:

$$S_T(\mathbf{x}) = \sum_{\mathbf{p} \in \mathbf{I}} \int_0^T e(\mathbf{p}, \tau) \cdot f_n(T - \tau) d\tau, \tag{3.36}$$

with n denoting the selected temporal function and p the spatial dimension of the stimulus. The representation of motion streaks requires a precise modeling of the temporal characteristics of visual input. Both components are combined into a common representation of motion streaks:

$$r_{\theta,T}^{streaks}(\mathbf{x}) = (S_T(\mathbf{x}) * \mathcal{N}_{\theta,\sigma_1,\sigma_2,+\mathbf{q}}(\mathbf{x}) \cdot S_T(\mathbf{x}) * \mathcal{N}_{\theta,\sigma_1,\sigma_2,-\mathbf{q}}(\mathbf{x})) \tag{3.37}$$

For our results, parameters for spatial streak subfields are $\sigma_1 = 2.2, \sigma_2 = 0.44$ and a spatial shift of subfields $\mathbf{q} = (cos\theta, -sin\theta)^T \cdot 6.53$.

3.5.2 Results

For the upcoming results of motion streak representation, the parameters of temporal integration functions are set to that the functions drop below 10% of the initial value at 120ms ($\sigma_1 = 7.25, \sigma_2 = 79$), see Fig. 3.21 and [Barlow, 1958]. The results presented on the subsequent pages are produced using the second integration function f_2, because it yields a slightly higher and clearer overall activation.

3.5.2.1 Results for translatory motions

The model is probed with stimuli containing translatory motion of random dots, see Fig. 3.22. We probed the model with 16 different motion directions of adequate speed, so that the integrated events fill the receptive fields of spatial cells. We model a population of 16 motion streak sensitive cells using the spatial integration fields and temporal integration functions described above. The integrated events of moving random dots evoke activation of the accordingly tuned integration cells. This allows differentiation of available motion up to a parallel direction component. Because the spatial integration fields are symmetric, parallel motions with same orientation but opposite direction cannot be distinguished.

3.5.2.2 Results for rotating motions

The second stimulus contains events generated from a rotating random dot kinematrogram (RDK). Rotation is artificially created by rotating a random dot display by θ degrees per simulation step and then distributing the events over a period P (see also Sec. 3.2.7). Figure 3.23 shows the stimulus and results. A motion orientation field is generated by drawing the corresponding orientation for the maximum of the population at every image position.

3.5.2.3 Dependency on motion speed

Physiological evidence shows that motion streaks only occur for high speeds [Burr and Morgan, 1997]. The last experiment tests our model with regard to this hypotheses. The model is probed with a translating RDK stimulus of different speeds, see Fig. 3.25. Fig. 3.25 shows results of the motion representation achieved with motion streaks. As can be seen, for the slow stimulus only little activation is elicited. In the fast moving condition, the responses show a clear orientation bias that matches the local motion available in the stimulus.

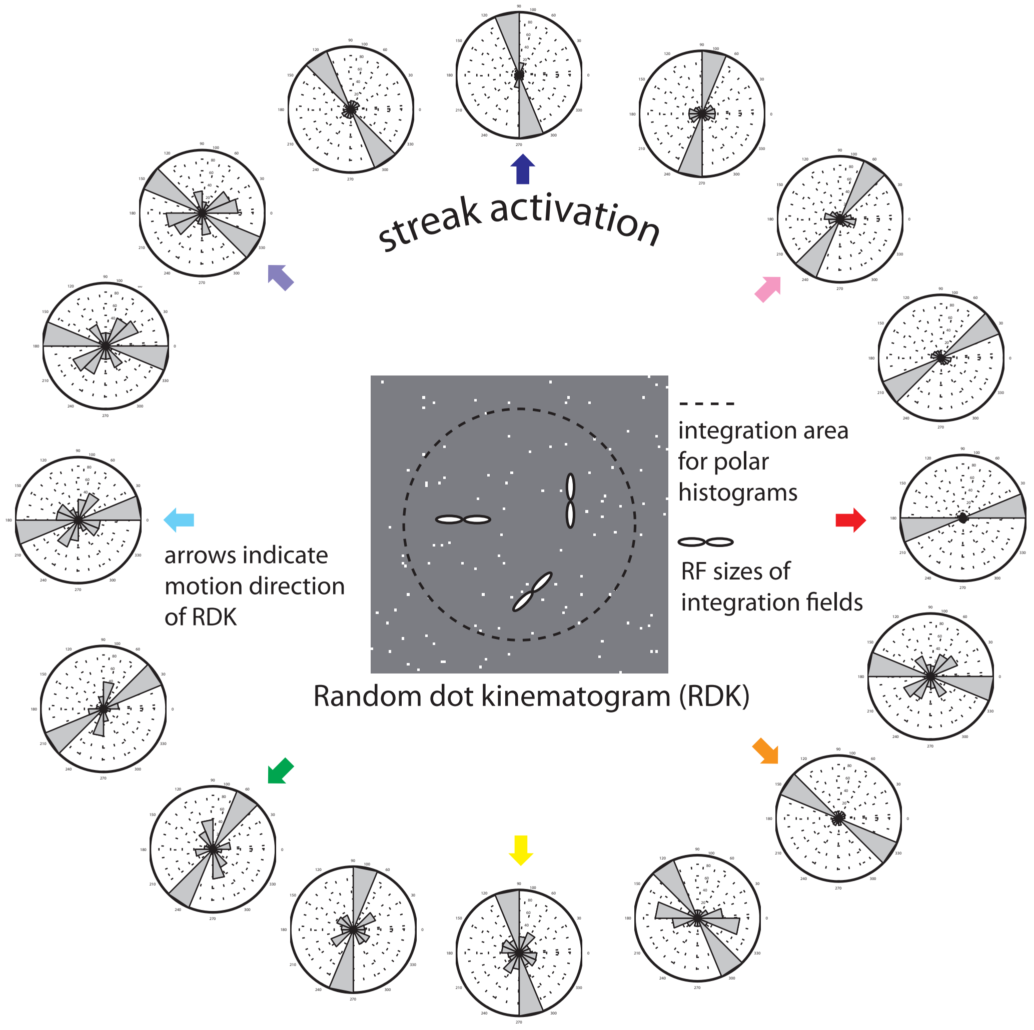


Figure 3.22: Representations of motion streaks for translatory motion. *A random dot kinematrogram (RDK)* (center) *is moving in different directions (16 steps from* 0°..360°*), which evokes a response distribution (*polar plots in outer ring*) of initially contrast-sensitive figure-eight shaped cells (RF size indicated inside RDK).*

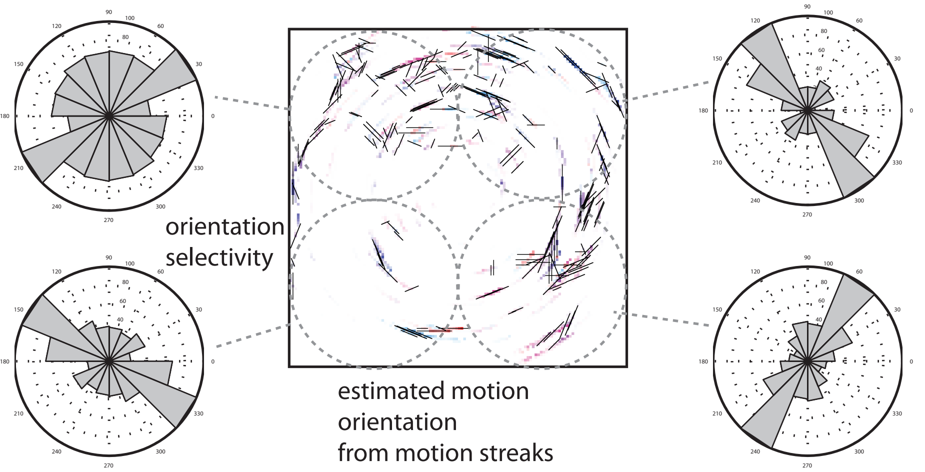


***Figure 3.23: Motion streaks for rotational motion**. A random dot pattern is rotating around its center. Motion streak activations are used to generate an orientation flow field (center). Motion streak activations are plotted for four quadrants (outer plots). The dashed circles circumscribe the regions with the input for calculating the population responses shown in the periphery.*

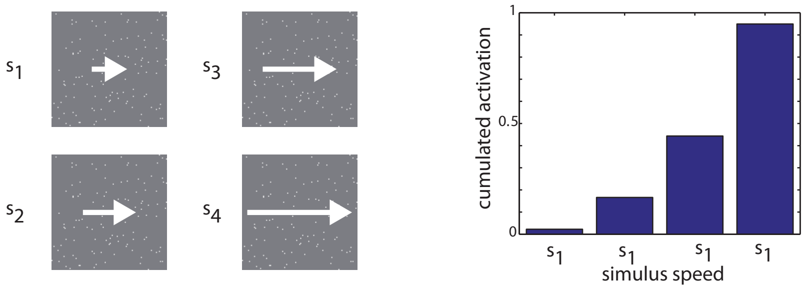


***Figure 3.24: Streak activation as a function of stimulus speed.** Representations of motion streaks only appear for rapid stimulus motion. Bars represent the activation strength for slow to fast speeds of moving RDKs.*

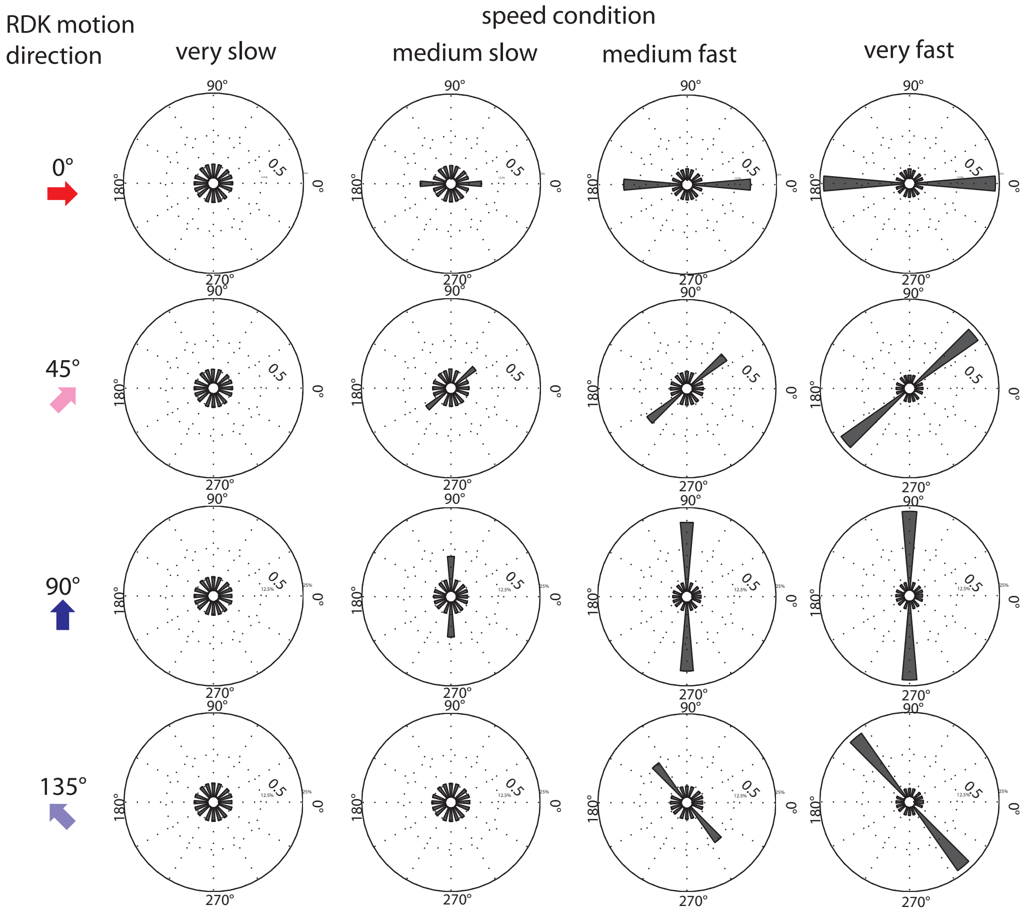


Figure 3.25: Dependency of motion streak response for different stimulus speeds with a pattern of translatory motion. *A RDK was presented to model cells. Their response strength is a function of presented speed (see conditions in columns) and the contained motion direction (rows).*

3.5.3 Form-motion interaction with motion streaks

As stated before, the mechanism of motion streaks might contribute to the estimation of motion direction. The possible advantage of a combined mechanism could lead to an increased precision of motion direction for normal speeds, because the broad tuning of spatio-temporally tuned filter functions do not allow a precise direction estimation. Second, it would increase the detectable range of motion speeds. For high speeds, the precision of motion detection drops as the speed progressively lies at the extreme end of the detectable range.

We use the initial stage of motion estimation with spatio-temporal filters (see Sec. 3.4.1) and the estimation of streak responses from the previous section to show interactions of motion streaks and signals from motion estimation. The interaction between motion direction cells tuned to estimated motion direction ϕ and form cells oriented parallel to that direction is modeled using a modulatory interaction (see Sec. 2.3.2). We employ a response modulation mechanism such that

$$r_\phi^{V1mod} \propto r_\phi^{V1} \cdot (1 + r_\theta^{V2}) \tag{3.38}$$

This modulatory interaction is followed by a pool normalization as already introduced in Eq. 2.9, which we apply in the steady state formulation:

$$r_{\phi,\infty}^{V1mod} = \frac{\beta r_\phi^{V1mod}}{\alpha_r + r_\phi^{V1mod} + \frac{1}{\alpha_q} \sum_{j \in \mathcal{N}} c_j g(p_j)} \tag{3.39}$$

$$q_{i,\infty}^{V1mod} = \frac{1}{\alpha_q} \sum_{j \in \mathcal{O}} r_j^{V1mod} \tag{3.40}$$

We probe the model with stimuli that contain linear motion with different speeds to highlight the effect of the modulatory input at different speeds, see Fig. 3.26. The sequence contains moving dots traveling at different speeds. The results indicate that a representation of quickly moving objects elicits responses in the form channel due to the integration of motion streak patterns. With modulatory interaction between model V1, V2 and model MT, the representation of motion is strengthened.

3.5.4 Summary

When an object is traveling across the visual fields, cells tuned for form show a response for sufficiently fast motions. We reproduced this behavior using event-based data from the DVS. For a quickly traveling object, the sensor produces a

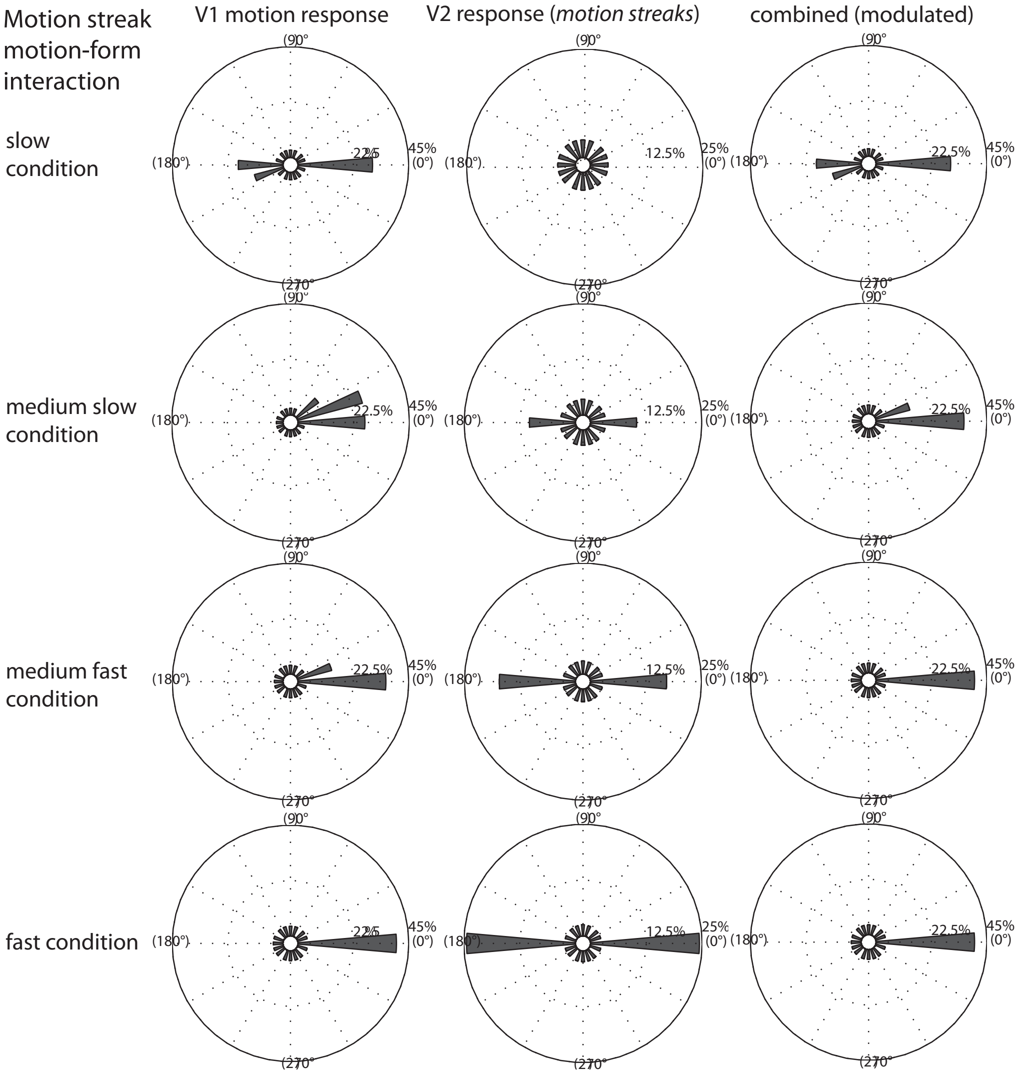


Figure 3.26: Form-motion interaction with motion streaks. *Plots indicate representations gained from RDK stimuli. For slow motions, motion streaks do not provide extra information. For high speeds, motion directions corresponding to oriented form responses generated by motion streaks modulate and enhance responses of motion selective cells.*

cloud of events along the object's trajectory. Considered an integration period of a few fractions of a second, the event cloud resembles the form of a streak that indicates the motion direction of the object. We designed cells in model area V2, which is originally sensitive to static input such that they respond to the input produces by a moving stimulus.

The results indicate that in our model likewise oriented contrast-sensitive cells are co-activated parallel to motion direction, thus representing motion streaks. Without the need to assume separate motion channel representations, these streaks occur in the form channel as a direct consequence of fast coherent motions along single directions, provided only a temporal weighting function that models temporal integration characteristic. The representation is only available for fast motions of random dot kinematograms and is thus in agreement with physiological findings. The model makes predictions concerning the strength of the streak patterns and sheds new light upon mechanisms of computing motion from form in a unified representation using event-based visual input.

3.6 An AER neuromorphic motion algorithm

The previous chapters showed mechanisms of motion estimation using event-based data generated by an asynchronously operating vision sensor. Instead of full frames at a time, this type of sensor provides a constant stream of visual events with a data rate that is a function of available image motion in the scene[Lichtsteiner et al., 2008]. This asynchronous operation provides some unique qualities that an algorithmic implementation must consider. This section will first present the requirements that an algorithmic implementation must preserve and later how this can be achieved by an adequate software design.

3.6.1 Requirements

An algorithm for the estimation of motion using event-based sensor input must fulfill the following requirements to preserve the unique features of the sensor output:

1. The algorithm has to implement the **physiologically plausible mechanism** presented in Sec. 3.4.

2. The algorithm has to preserve the **cortical organization** of the model, namely the relevant compartments V1,MT and V2. Each area contains its own representational format of events.

3. Incoming events are specified by their position, time and type, the so-called address-event-representation (AER). Any result of the algorithm should be

comprised of the same **address-event-based format**. This includes the same high temporal resolution of generated output.

4. The date structure and processing should respect the **sparse structure** of the incoming data.

5. Events exist independently and are only aggregated for visualization purposes. The algorithm has to achieve a **continuous estimation** of motion for each individual incoming event. This includes the initial motion estimation, application of feedback and normalization.

6. When many visual changes appear in the view of the sensor, many events are generated in a short period of time (see Sec. 3.2.6). When the scene is not moving, no changes appear. This variation should be reflected in a **non-constant computational load**.

7. The algorithm must be comprised of an **efficient design** with respect to the data structure and computational complexity.

3.6.2 Algorithmic design

The following paragraphs presents the software architecture that was chosen to fulfill the requirements. The processing steps of the algorithm is visualized in Fig. 3.27.

3.6.2.1 Data representation

The algorithm receives input as a list of datagrams, where each item represents an event with address, a time stamp and a type. For the calculation it uses a buffer structure and models of weight functions based on temporal differences. Incoming events and hypotheses are stored in a data structure that allows to quickly access events in local neighborhoods without traversing the event list. The events are buffered in an first-in-first-out buffer to allow quick lookup of neighboring events. The buffer is retinotopic in a way that a short list exists at each image position that holds events in AER format. This later allows to quickly lookup recent events at desired positions, without integrating events into a 'flat' datastructure.

During motion estimation, input is put into relation with previous inputs to estimate motion hypotheses. Such a motion hypothesis is also comprised of a spatial and temporal location, but instead of two types for ON- and OFF-events, the algorithm generates N different directional motion hypotheses.

The algorithm outputs results that represent motion hypotheses in a form similiar to AER, with motion events instead of events that indicate luminance change.

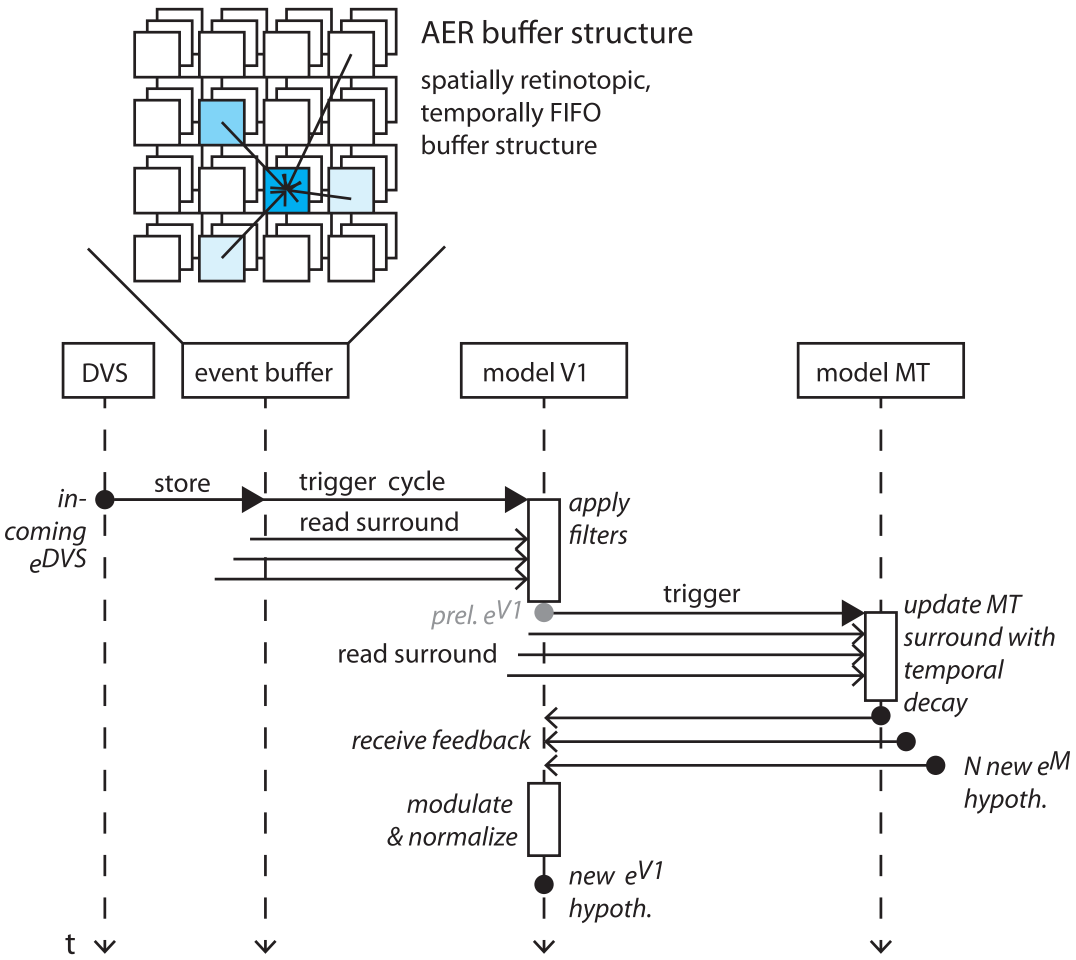


Figure 3.27: A neuromorphic algorithm for event-based motion processing. Top: *Data structure for event processing with fast access to newest event in list.* Bottom*: Sequence diagram of algorithm structure and data pathways. See text for details.*

The time between incoming events is not constant. While in scenes that contain many motions events may arrive at microsecond intervals, static scenes may contain only very few incoming events generated by sensor noise. When such an incoming event is put into relation to its neighbors to calculate filter responses, the temporal difference between the newest event and those in the vicinity is crucial to achieve correct estimates. The buffer structure holds the timestamps B_t of previous events, and the spatial position is defined by the buffer list itself. The calculation of a weight H at position $\mathbf{p}$ at time t for an incoming event is thus a combination of two components

$$H(t, \mathbf{p}) = \sum_{\mathbf{q} \in \mathcal{N}} w_s(\mathbf{p} - \mathbf{q}) \cdot \sum_{n=1}^{N} w_t(t - B_t(\mathbf{q}, n)) \tag{3.41}$$

with $\mathbf{q}$ the position of spatial neighborhood, w_s a spatial weight function and w_t a temporal weight functions in an aggregation. In case of the initial motion estimation step, the two functions are similar to those in Sec. 3.4.1.3.

In case of motion integration in model MT, the spatial representation is subsampled and w_s reduced to a Gaussian function of spatial weight, its temporal function is reduced to a temporal decaying function like

$$w_t^{MT}(\Delta t) = exp(-\frac{\Delta t}{\sigma}). \tag{3.42}$$

with σ to control the steepness of temporal decay and Δt in microseconds. Typical values for are $\sigma = 200$ to simulate a $100ms$ timespan of temporal integration (10% weight at $\Delta t = 100ms$). A data representation like this including the temporal functions is implemented for V1, MT and for a representation of motion streaks as seen in Sec. 3.5.

The separation of the temporal and spatial components here reflects the fact that the neuromorphic filter design (see Sec. 3.4) also incorporates two separable components in a linear combination to yield a spatio-temporally inseparable filter for motion detection.

3.6.3 Processing

For clarity, we will in the following specify events of different sources: Original DVS events that have been used throughout the thesis as e are now referred to as e^{DVS} to avoid confusion. Motion hypotheses (or motion events) at model level V1 which represent activation for a motion direction are referred to as e^{V1}.

This representation hold the complete population response for different motion directions, see Sec. 3.4.1.3).

The processing sequence is triggered by an arriving event from the DVS (see Fig. 3.27), which is stored in the event buffer for future reference. The model V1 stage polls the buffer entries at the same position and the the surround and calculates the contribution based on their spatial and temporal distance from the trigger-event. The final hypothesis of V1 requires feedback influences by MT. Because the influence of the stored MT hypotheses have changed with the time passing since their generation, their representation needs to be updated with the current timestamp. After the initial V1 hypothesis is generated, an update process of the MT is thus triggered. This requires to read motion representations of surrounding V1 events. The MT representations are then applied to modulate the V1 hypotheses. The final V1 is a population response containing N different motion directions. This representation is normalized with the available representations from the surround. The final (modulated) V1 response is then stored as motion event e^{V1}. After this processing sequence, the algorithm is prepared to process the next incoming event.

3.6.4 Considerations on complexity

The proposed algorithm estimates optic flow from an asynchronous representation of visual events using mechanisms found to exist in early stages along the dorsal pathway. Instead of operating at a constant frame rate, the computation time depends on the event sequence that is influenced by the visual structure of the recorded scene. In this section we derive an approximative estimate of the real computational effort necessary for optic flow computation.

In order to end up with a comparable assessment of the complexity, we make use of a simplified version of register machines where the complexity can be summarized by multiply-and accumulate operations, where the effort r increases with the cost of every processing step X times the number of its repetition:

$$r \leftarrow r + N \cdot X \tag{3.43}$$

For the evaluation we consider a time window T and regard the amount of operations necessary in a frame-based and in an event-based operating domain. The number of frames inside T is

$$N_{frame} = \frac{T}{\text{frame-rate}} \tag{3.44}$$

With the continuously firing local events in a time-window T, the number of potential events is

$$N_{event} = \frac{T}{\text{event-rate}} \tag{3.45}$$

that yields a gain-factor

$$\lambda = \frac{N_{event}}{N_{frame}} \tag{3.46}$$

We will in the following compare the computational complexity of our proposed method to a hypothetical version of the algorithm that is implemented using a conventional camera technology. The precise platform will not be of importance for our considerations, nor are we going to consider any optimization methods like operation in frequency domain, parallel processing or similar. Our hypothetical camera is operating at a constant and typical frame-rate F of 25 Hz. The resolution is set to 128 × 128, according to the DVS sensor. The hypothetical algorithm works analog to the method proposed in Sec. 3.4, in that it calculates responses of directionally tuned cells by means of spatial integration from their respective receptive fields. The hypothetical response is generated for a population of direction-sensitive cells for different motion directions θ. In addition, our hypothetical final response at V1 level is generated from a convolution operation of one frame at a a time with the filter functions, following the model of Sec. 3.4.1.3. All solutions are based on the assumption that for every incoming event a cycle of operations is triggered in the neuromorphic algorithm. The essential operation is the estimation of response weights given the surrounding events and their temporal influence function.

Frame-based complexity for motion estimation

In a frame-based approach, motion estimation requires local weighted sums (scalar products) of surrounding pixels. Here, X_{dim} is the image size and X_F is the filter size. This needs to be performed at every image location and for every image in time window T, resulting in the operations O_{frame}

$$O_{frame} = N_{frames} \cdot X_{dim}^2 \cdot X_F^2 \tag{3.47}$$

Event-based estimation

A time window T is densely sampled by visual events. Operations are calculated at every pixel in the moment they are provided by the sensor. For each arriving event, the algorithm calculates the contribution of the event regarding the filter neighborhood X_F^2 and a temporal sub-window X_T. These operations occur at every image location X_{dim}^2 times N_{event} which ends up with a total number of operations

$$O_{event} = N_{events} \cdot X_{dim}^2 \cdot X_F^2 \cdot X_T. \tag{3.48}$$

$$= \lambda \cdot N_{frame} \cdot X_{dim}^2 \cdot X_F^2 \cdot X_T. \tag{3.49}$$

Empirical studies and results

From a theoretical viewpoint, the complexity using an event-based processing algorithm is much higher than with frame-based approach because of the immense maximum event rate individual pixels can produce. Empirical studies however revealed that the realistic amount of events in a time window is much lower which is mainly influenced by the sparseness of data and depends on the contrast and speed contained in the scene. This results in benefits for the event-based approach, depending on the scene content. Furthermore, the algorithm efficiently calculates a sparse scalar product using the buffer structure, so X_T simplifies to a constant factor. The equation which considers a sparseness factor ϕ of the data yields

$$O_{event} = \phi \cdot \lambda \cdot N_{frame} \cdot X_{dim}^2 \cdot X_F^2. \tag{3.50}$$

The fundamental and critical aspect of the event-based sensor is the density of events in the relevant time box. The number fluctuates with the temporal changes in the visual field. We will show that the typical number of events is small enough to yield a computational benefit for the event-based processing.

In the introductory part to the sensor (Sec. 3.2) we mentioned the potential refresh rate of an individual pixel is $f^p = 25kHz$, so the number of theoretically observable events per pixel is $N_{p,event}^{theo} = \frac{T}{f^p} = 25,000ev/s.$.

A far more realistic consideration is the rate that is technically plausible when dealing with the sensor and its hardware. Sec. 3.2.6 already mentioned a rate of events that is limited by the bus technology and embedded processing capabilities. This maximum integrated event-rate is about $250,000s^{-1}$, to yield N_{event}^{tech} this has to be divided by the size of the visual field, thus $N_{p,event}^{tech} = T \cdot 250,000/X_{dim}^2 \approx 15ev/s$. This marks the maximum event rate from the processing standpoint.

Under real-world conditions, this number drops further, because sequences are rarely densely filled with motion that could cause events at that rate. We assume

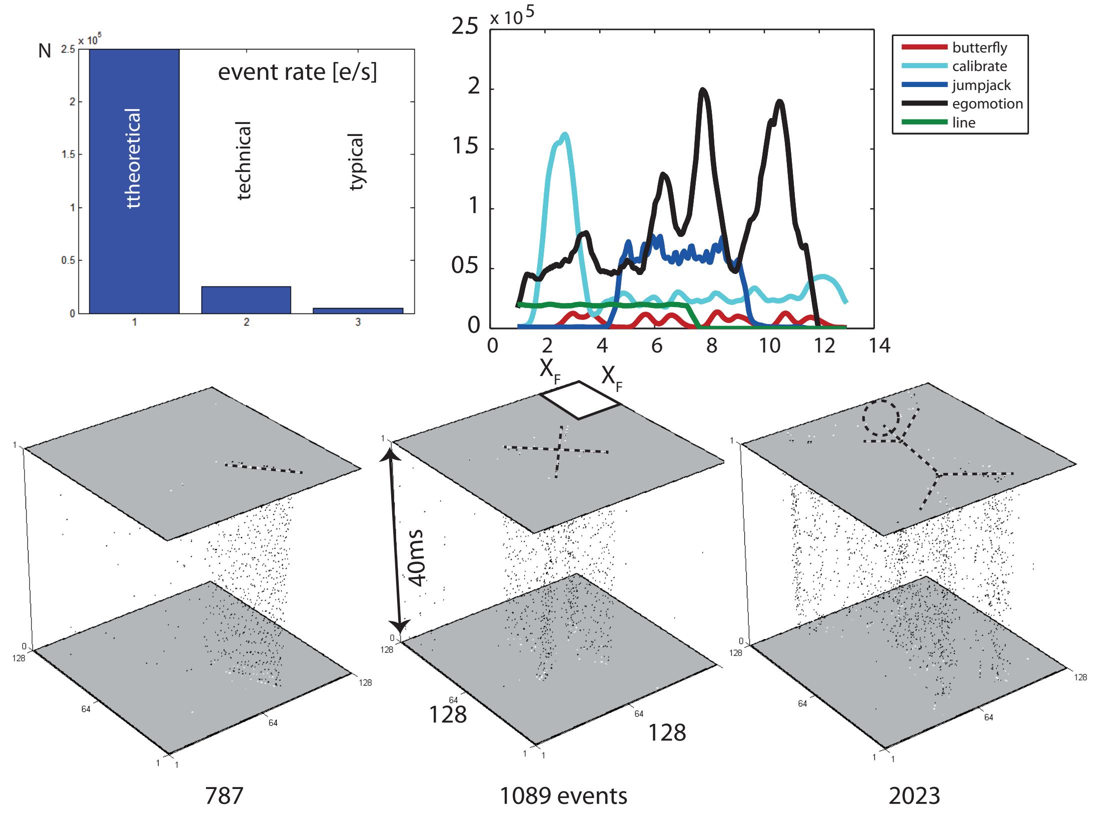


Figure 3.28: Number of events per time slot for different stimuli. Top-Left: *Theoretical maximum of possible events outnumbers the realistic value significantly.* Top-Right: *Event rates in e/s for a set of recorded scenes. Synthetic stimuli (*line, *green) exhibit a low and constant rate of about 20,000 ev/s , whereas the rate in recordings of real situations ranges from even less (*butterfly-action, *red) to 200,000 ev/s for a scene containing* ego-motion *(black).* Bottom: *Actual event cloud between two consecutive frames at 25Hz (40ms difference). The available event in that region are counted. The size X_F of used filters is indicated in the central frame.*

a typical event rate of $N^e_{typ} = 25,000$ in the complete scene, thus $N^{real}_{p,event} = T \cdot 25,000/X^2_{dim} \approx 1.5ev/s$.

Fig. 3.28 shows the typical event rate for a set of processed scenes. This stands against an effort per pixel of $25 \cdot X^{F2}$.

Table 3.1 summarizes the efforts of comparing a hypothetical algorithm of motion estimation with the proposed event-based mechanism. We set $F = 25Hz$, $S = 128$ and $N^{theo}_{event} = \frac{T}{f_p} = 250,000$ and $N^{tech}_{event} = 250,000$ and $N^{real}_{event} = 25,000$.

This means that under typical conditions, the initial stage of the proposed algorithm operates with only about 6% of and equivalent conventional frame-based algorithm. Even in the case of an event rate that is limited by the bus the computational benefit is a factor of 1.6.

class	events/pixel in T	total events	frame-based ops	ratio
theo. Max N^{theo}	25,000	409,600,000	409,600	1000:1
tech. Max N^{tech}	15,3	250,000	409,600	0.6:1
emp. N^{real}	1,5	25,000	409,600	0.06:1

Table 3.1: *Considerations on operations for event- and frame-based operation.*

This consideration does not assume memory demands of the systems. At a closer inspection, these are however of similar complexity. Both algorithms need to maintain a data structure that holds a buffer of events at each spatial position. In our proposed algorithm this buffer was of the size of 20 elements. A similar storage overhead would be necessary for the conventional algorithms as well to generate temporal integrations. These investigations are supposed to provide a first measure of the potential of event-based algorithms. The sparse representation, high temporal resolution and low energy consumption has earlier been demonstrated [Delbrück and Lang, 2013] and these calculations motivates the further investigation of transforming neuromorphic mechanisms to event-based hardware.

3.7 Conclusion and discussion

This chapter proposed an algorithm that achieves asynchronous estimation of optic flow from event-based input of a neuromorphic hardware. It applies findings about the processing structure and architecture of the dorsal pathway.

3.7.1 Summary of contributions

We introduced the DVS128 sensor and its capabilities in Sec. 3.2, along with the concept of asynchronous events, their representation in the AER format as well as their visualization for frame-based screens or printed devices. We presented hardware for the recordings of precise visual input and showed technical solutions for the generation of synthetic stimuli. Sec. 3.3 focused on theoretical implications in context with the sensor. We proposed a method of estimation of optic flow that applied knowledge from physiological findings to generate a model of initial motion estimation. The model includes subsequent processing steps along the visual dorsal pathway. We show how motion integration increases robustness by means of modulatory feedback. The liability of the initial filter stage to the aperture problem is reduced on an early level by application of inhibitory feedback. This extension of a normalization stage reduces the confidence of estimated events along elongated contrasts, where the local estimates suffer from the aperture problem. The model also proposed interaction of the motion and form pathway by means of motion streaks. The streaks or speed lines provide a spatial code for motion direction. We showed how a representation and detection of such motion streaks is available in the dense representation of visual changes of the event stream. This bridges to the ventral stream of form processing by picking up the theories of motion streaks. We demonstrated how such motion streaks or speedlines can be represented using the temporal events from the DVS128 sensor and how they might be used for an alternative representation of motion hypotheses in the form channel. A possible interaction with the available architecture of the dorsal stream is also mentioned. Sec. 3.6 shows how such an event-based estimation mechanism can be implemented while preserving the sparse representation and asynchronous estimation given the inputs from the event stream. We propose a data structure and introduce several addition events types along the processing hierarchy. The chapter concludes by a consideration of complexity of the proposed algorithm. We compared our implementation of the event-based mechanisms to a hypothetical algorithm that operates in a frame-based manner.

3.7.2 Relation to previous models

To our knowledge the proposed model is the first to apply mechanisms of motion estimation and related subsequent processing steps to asynchronous event-based data that have been acquired from a DVS128 sensor. Previous models of motion estimation on this type of data exist, but they adapted known approaches like the Lukas-Kanade-Algorithm to AER data[Benosman et al., 2012, Benosman et al., 2014]. This approach however introduce conceptual flaws due to the nature of the contained information in the AER data, which is a temporal derivative of the luminance function. Furthermore, the robust estimation of image gradients is difficult in the sparse DVS data, which further deteriorates the estimation result. We achieve robustness of the estimation by a novel initial estimation stage and a recurrent architecture inspired by previous models[Bayerl and Neumann, 2004]. Both concepts proved to be very well suited for the asynchronous data and to generate satisfying results. The models shows a how the output of a neuromorphic retina sensor (the DVS) and its unusual representation of visual events is combined with physiological evidence about response characteristics in the visual cortex to yield motion sensitive cells. Furthermore, our model is able to functionally link the ventral and dorsal pathway of visual cortex with the concept of motion streaks, with is a novel approach in modeling mechanisms of motion estimation. The high temporal resolution and characteristics of the DVS data makes it possible to model an activation of cells along the form pathway and lets them represent motion and motion orientation. Those effects have been observed and there exists evidence for an interaction between the channels. We also propose a software architecture for an implementation of the proposed functional modules that maintains the unique features of the data representation like asynchronism in the estimation, sparse representation and low computational load.

3.7.3 Limitations

The proposed model emphasized the combination of retina-like event representation with known principles of filtering and response characteristics of cells along the dorsal pathway. The functionality is demonstrated with a set of test stimuli. Limitations of this process are given by the lean resolution of the image sensor and a processing and data throughput bottleneck that is located on the hardware that we performed experiments with. Future versions will most certainly increase the related performances. We performed a theoretical analysis of complexity given a set of hypothetical and typical event rates to get a feeling for the necessary operations during motion estimation. However, the prototypical implementation never intended to achieve real-time capabilities but focused on an understandable and maintainable implementation that provided many interfaces for experiments. By the time of writing a new implementation with focus on real-time-processing is being developed.

3.7.4 Publications

The contributions contained in this chapter are the accumulated outcome of the preceding research period. Prior investigations and earlier results have been presented in individual publications. [Tschechne and Neumann, 2014b] introduced the concept of motion streaks using event-based sensor input. [Tschechne et al., 2014b] presented the first bio-inspired approach of motion estimation based on asynchronous sensor data. In [Tschechne et al., 2014a] and [Brosch et al., 2014] we have focused on more theoretical investigations concerning the utilization and implications to modeling of such a neuromorphic event-based data stream.

4 Hierarchical representation in visual cortex - from localized features to figural shape segregation

4.1 Introduction

Contextual modulation is a central part in visual processing and the underlying principles can be observed throughout the visual system [Nurminen and Angelucci, 2014] (for an extensive review see [Krause and Pack, 2014]). While the influence of contextual modulation has been extensively described in the striate area, also extrastriate areas profit from it.

The last chapter has shown how the generic mechanisms of contextual modulation have been applied in a model of motion estimation. Here, it improves initial representations and contributes to a number of effects like an influence to the aperture problem and motion streaks. The small temporal integration period of initial receptive fields can be massively extended using contextual feedback.

The mechanism that applied to the motion estimation process can also be used for other domains. In the following, we will show how it can be applied to form processing. Here, contextual modulation will influence local representations that would otherwise suffer from their small receptive fields. The three stages of the cascadic model (initial filtering, feedback, normalization) allow to introduce contextual information to other, earlier areas. This on the one hand make representation more robust (as already observed in the motion processing), but also allows some processing steps in the first place. Contextual modulation makes information from later stages available to earlier areas and thus allows a more detailed representation of visual information. The processing is comprised of the same architecture components as were introduced in Sec. 2.3.

4.2 Visual processing along the ventral pathway

We visually perceive our environment as a stable and comprehensive combination of components. This allows us to easily identify objects and persons and we efficiently analyze geometrical cues that allow a precise and robust recognition, navigation and interaction. Local visual features act as a vocabulary for such a scene description and they are integrated from simple and local features into more meaningful representations of increasingly complex scene components. The visual cortex extracts such visual features along the ventral pathway and segments the environment into image regions and combines those into a representation of surfaces and prototypical objects.

We propose that on the way from generalizing early local features to higher meaningful representations, the role of object boundaries plays an essential part. Contrasts indicate spatial changes in local illumination which might coincide with object boundaries that allow segregation from background. However, contrasts indicating a real transition from one object to another or from the object to the background must be separated from those indicating an illumination change and those caused by textured regions. This must be accomplished using contextual information. The region delimited by such a boundary is a surface with locally constant parameters, and a set of surfaces forms objects, scenes and eventually our complete visual environment.

We believe that the processing capabilities of early and intermediate stages of visual cortex are used to transform local representation into an intermediate, more meaningful representation of contours, shapes and surfaces. Following those ideas, we propose that a stable representation of shape may be established by interacting assemblies that are each devoted to specific features properties. Multiple mutually connected areas in the ventral cortical pathway receive visual input and extract local form features that are subsequently grouped into increasingly complex, more meaningful image elements. Such a distributed network of processing must be capable to make accessible highly articulated changes in shape boundary as well as very subtle curvature changes that contribute to the perception of an object.

We propose a recurrent computational network architecture that utilizes hierarchical distributed representations of shape features to encode surface and object boundary over different scales of resolution. Our model makes use of neural mechanisms that model the processing capabilities of early and intermediate stages in visual cortex, namely areas V1-V4 and IT. We suggest that multiple specialized component representations interact by feedforward hierarchical processing that is combined with feedback signals driven by representations generated at higher stages. Based on this, global configurational as well as local information is made available to distinguish changes in the object's contour. Once the outline of a shape has been established, contextual contour configurations are used to assign border ownership directions and thus achieve segregation of figure and ground. The model, thus, proposes how separate mechanisms contribute to distributed hierarchical cortical shape representation and combine with processes of figure-ground segregation.

Our model incorporates processing at low and intermediate areas of visual cortex in a hierarchical architecture of two-dimensional shape representation. Each model area consists of a three-stage processing cascade of initial filtering, application of modulatory feedback effects and center-surround interactions leading to an activity normalization[Carandini and Heeger, 1994, Carandini et al., 1999, Kouh and Poggio, 2008, Carandini and Heeger, 2012]. Our model combines the representation of visual shapes with mechanisms for figure-ground segregation on the basis of assigning border ownership and incorporates a distributed representation of local contour curvature over different cortical areas. In our model we emphasize the computational role of feed-forward and feedback mechanisms [Grossberg, 1980, Edelman, 1993] to generate a hierarchical distributed representation of shape information. The feedback amplifies the sensory signal such that the subsequent competition between neurons builds a competitive advantage [Tsotsos, 1988, Girard and Bullier, 1989, Desimone, 1998, Roelfsema et al., 2002, Reynolds and Heeger, 2009]). Boundaries and their orientation are represented after initial processing in model area V1 and a grouping stage in model area V2. Contextual boundary configurations are also represented at a coarser spatial level at model V2 and V4 to achieve selectivities towards contour curvature. With the influence of feedback, those cells are enhanced at lower stages that contribute to a matching bottom-up signal.

The output of our model is a representation of shapes and shape segments where contextually compatible boundary information benefits from recurrent feedback connections. Such a representation could provide input to subsequent processing stages for e.g. object classification tasks, which would clearly benefit from the enhanced representation.

This model extends previous own works [Neumann and Mingolla, 2001, Hansen and Neumann, 2004, Weidenbacher and Neumann, 2009] but introduces functional properties that have been inspired by the works of other groups. A model of curvature representation can also be found in [Cadieu et al., 2007]. The authors modeled physiological findings of the same group [Pasupathy and Connor, 1999, Connor et al., 2007] that has focused on the dynamics of contour processing [Yau et al., 2013b]. Cell representations from early visual areas are combined to intermediate-level shape descriptors are used in a computational model by [Rodríguez-Sánchez and Tsotsos, 2012]. [Riesenhuber and Poggio, 1999, Riesenhuber and Poggio, 2000, Mutch and Lowe, 2008] released very powerful models of object and object class categorization in a hierarchical modeling approach. The physiological [Zhou et al., 2000, O'Herron and von der Heydt, 2011] as well as the computational[Layton et al., 2012] aspects of border ownership are subject to intense research. Models of contour integration and perceptual grouping also exist from [Zhaoping, 1998] and [Jehee et al., 2006, Roelfsema, 2006].

The following section introduces the reader to our proposed model. The model incorporates generic mechanisms of cortical information processing, of which the most relevant ones are briefly summarized in Sec. 4.3.1. For a complete introduction to the mechanisms the reader may be referred to Chap. 2. Subsequently, the model areas are described in detail, starting with the forward directed mechanisms in Sec. 4.3.2 and involved areas, followed by the feedback sweep and its role into providing contextual influence to early representations in Sec. 4.3.3. Results of all participating areas are presented in a concentrated form at the end of this chapter, see Sec. 4.4.

4.3 Hierarchical form representation in visual cortex

Fig. 4.1 presents the overview of the model. Visual input enters the model at the bottom and is subsequently processed by interconnected functional areas with increasingly large receptive field sizes. Solid arrows indicate feedforward, dashed arrows indicate feedback, or modulatory, connections. Each area implements a generic architecture of building blocks that consists of (i) filtering of the input, (ii) modulation by feedback and (iii) response normalization (see Sec. 2). **Model V1** consists of image filters that resemble properties of early processing in LGN and V1, namely simple and complex cells that are tuned to circular or elongated image contrasts (see Sec. 4.3.2.1). **Model V2** integrates responses of model V1 with long-range integration cells. A multiplicative combination of sub-cells responds best to elongated contrasts of one dominant orientation. Also at V2, a population of cells represents border ownership directions. At population of long-range curved integration cells help represent different boundary curvatures. The **Model V2/V3 complex** hosts representations of corners by integrating V1 responses from orthogonal configurations over a small spatial surround (see Sec. 4.3.2.2). **Model V4** consists of cells that asymmetrically integrate responses from V1 and V2 to become curvature selective at an increased spatial scale (see Sec. 4.3.2.3). In **Model IT**, cells with large receptive fields integrate responses from V1, V2 and V4 at local figure convexities to achieve a contextual segregation into figure and ground. Area V4 allows a description of a shape by means of cues that are represented on distributed areas in the model. Those cues exist at different spatial scales and their mutual interaction generates dynamic processes in the model (see Sec. 4.3.2.4).

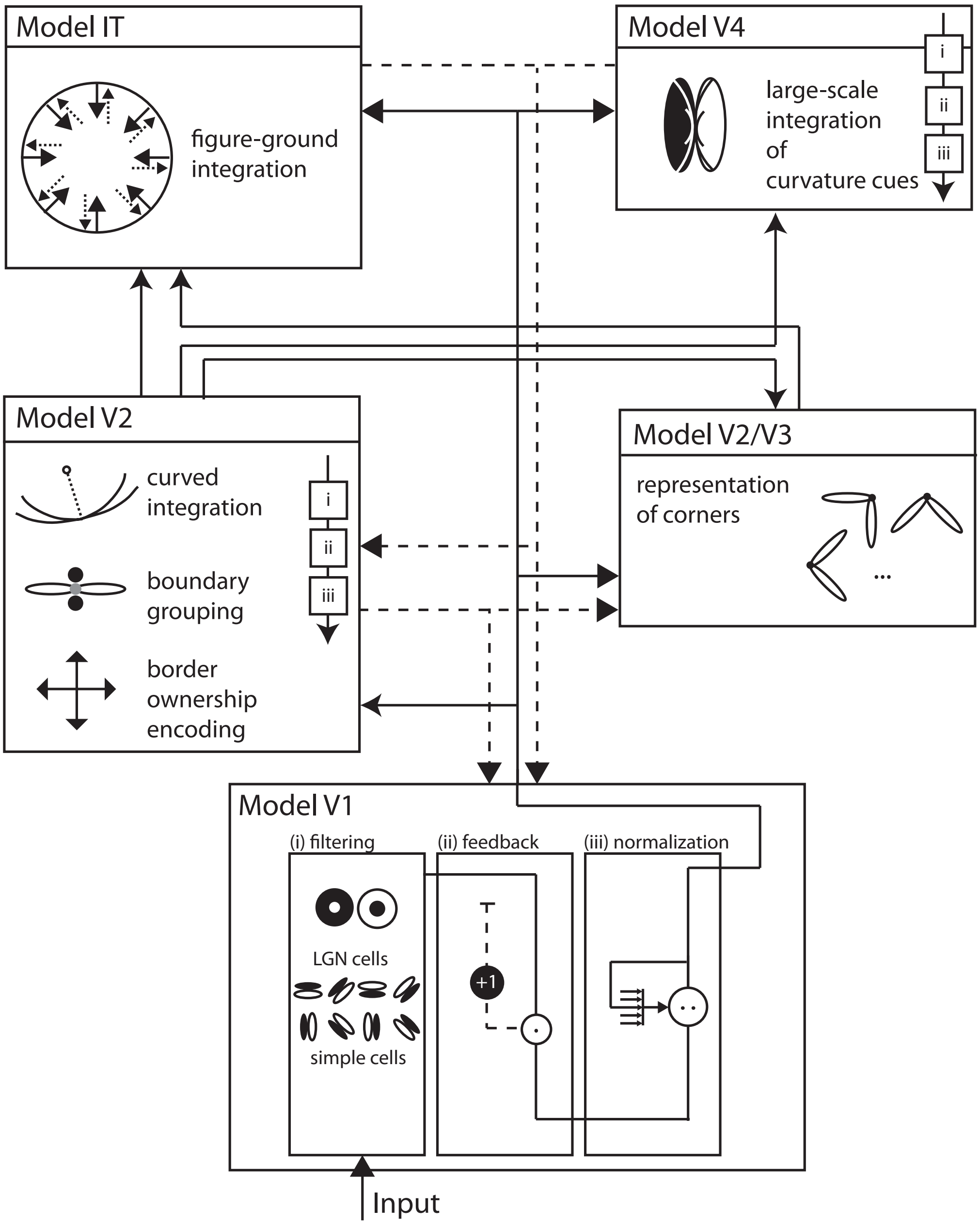


Figure 4.1: Overall model architecture. *Input enters the model from below and is hierarchically processed along different model stages. Solid lines denote feedforward connections, dashed lines denote modulating feedback connections. Each stage implements a model cascade of the components filtering, contextual modulation and normalization. See text for details.*

4.3.1 Processing mechanisms

Our model is comprised of several hierarchical building blocks. The following paragraphs recapitulate the basic mechanisms that are of special interest for form processing. The mechanisms presented in this chapter base upon the architecture introduced in Chap. 2 and make intensive use of the three-stage cascade that was presented there. The relevant aspects are briefly repeated here.

The model cascade is comprised of a stage of initial filtering, application of contextual feedback modulation, and response normalization.

The first model stage of the cascade is the initial filtering that calculated the response of a cell assembly with a given receptive field and tuning to presented input. In our model, neural activations or response levels are modeled using a scalar representation of the neural firing rate. In general, model cell responses follow first-order dynamics and represent the changes of membrane potentials. Such dynamics are influenced by excitatory and inhibitory inputs and a passive decay of activity (see Chap. 2). In order to simplify the computations in our large-scale simulations we assume that such linear feed-forward filtering operations quickly relax at their equilibrium state. We thus model the response for the preferred stimulus in the visual field with a two-dimensional convolution operation of the given input I and a model of the tuning function (or *preferred stimulus*), which acts as a convolution kernel K_{pref}. The response of model cells at position $\mathbf{x}$ is thus $r(\mathbf{x}) = I(\mathbf{x}) * K_{pref}$ or

$$r(\mathbf{x}) = r\begin{pmatrix} x \\ y \end{pmatrix} = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} K_{pref}\begin{pmatrix} u \\ v \end{pmatrix} \cdot I\begin{pmatrix} x-u \\ y-v \end{pmatrix} \tag{4.1}$$

A frequently used kernel in our model serves as elementary building block and is a two-dimensional Gaussian distribution that is elongated along one axis and rotated around its center. We refer to this distribution by $\mathcal{G}$ with parameters for orientation θ, deviation along the axes σ_1, σ_2 and the center of the distribution μ.

$$\mathcal{G}_{\theta,\sigma_{1,2},\mu}(\mathbf{x}) = \frac{1}{2\pi\sigma_1\sigma_2} exp\left(-\left(\frac{(\hat{x}-\mu_x)^2}{2\sigma_1^2} + \frac{(\hat{y}-\mu_y)^2}{2\sigma_2^2}\right)\right) \tag{4.2}$$

with a contained rotational transformation using

$$\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = \begin{pmatrix} cos\,\theta & -sin\,\theta \\ sin\,\theta & cos\,\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \tag{4.3}$$

If parameters are not specified they are considered having the following default values: $\theta = 0°, \mu = (0,0)^T, \sigma_1 = \sigma_2 = 1$. In the following, functional filter kernels will often be designed as a combination of multiple such elementary components.

The coefficients of the kernel that models the preferred stimulus might incorporate negative weights to account for the inhibitory connections a cell may receive. This could lead to overall responses that are numerically negative. We thus use a rectification operator after convolution and feedback stages to ensure that numerically the response rate r of a population is not negative:

$$|r|^{+} = max(0, r) \tag{4.4}$$

At the second stage of the cascade, response levels are modulated by recurring input from higher visual areas. We propose a feedback mechanism that exerts a purely modulatory gain control on the input. That means that feedback alone cannot generate activities without activation by the initial filtering step (compare Fig. 2.3). With r being the unmodulated driving signal and net_{FB} being the strength of the feedback, the modulated response is

$$r_{FB} \propto r \cdot (1 + net_{FB}). \tag{4.5}$$

Using this approach, given $r = 0$ no signal is generated as output irresponsible of the strength of the feedback net_{FB}. On the other hand, if no feedback signal is available, the right part of the equation leaves the input signal r unchanged [Hupé et al., 1998, Salin and Bullier, 1995, Gilbert and Li, 2013, Eckhorn, 1999]. At times we apply a non-linear transfer function to map the computed responses to a cell activation level. In our model, we use the following function with k the nonlinearity parameter.

$$\tilde{r} = r^{k} \tag{4.6}$$

At the final stage, we incorporate a mechanism that keeps the response level limited by using a *shunting inhibition* that leads to a non-linear compression of high amplitude activities resembling the *Weber-Fechner-Law* of perceptual thresholds. In its dynamic formulation, the rate of change of the signal $\dot{r}_{\theta}^{norm}$ depends on the current activation level as well as the amount of input I_{net}:

$$\dot{r}_{\theta}^{norm} = -\alpha \cdot r_{\theta}^{norm} + \beta \cdot r_{\theta} - r_{\theta}^{norm} \cdot I_{net} \tag{4.7}$$

$$I_{net} = \frac{1}{N} \sum_{i=1}^{N} r_i. \tag{4.8}$$

With N the number or angles of represented orientations. When this equation is solved at equilibrium, i.e. when $\dot{r}_\theta = 0$, the activation becomes

$$r_\theta^{norm} = \beta \frac{r_\theta}{\alpha + I_{net}} \tag{4.9}$$

The constants influence the steepness of the nonlinearity (α) and the scale of the normalized signal (β).

4.3.2 Feedforward sweep

In the following, we describe the forward sweep of our model, from early towards intermediate processing stages. After all areas have been described in detail, we will elaborate on the feedback connections (Sec. 4.3.3) that build the recurrent model structure.

4.3.2.1 Model Area V1

The processing starts at early stages of visual cortex where we model the functionality of LGN and V1 cells where LGN cell responses provide feed-forward input to V1 cells. Here, the visual input is initially processed to generate a representation of local image contrasts and local contrast orientations [Hubel and Wiesel, 1962].

$$r^{LGN} = |I * (\mathcal{G}_{\sigma_1} - \mathcal{G}_{kappa})|^+ \tag{4.10}$$

with σ and κ denoting the width of center and surround kernel, respectively. To model cells that are tuned to oriented contrasts, we use elongated Gaussian kernels $\mathcal{G}$ that are combined into odd-symmetric simple cell profiles using anisotropic σ_1 and σ_2 and a radius ρ_1 for the spatial shift of the integration kernels. The responses of such cells are denoted by the steady-state equation

$$r_\theta^{V1}(\mathbf{x}) = \left| r^{LGN}(\mathbf{x}) * (\mathcal{G}_{\theta,\sigma_{1,2}}(\mathbf{x}+\mathbf{p}) - \mathcal{G}_{\theta,\sigma_{1,2}}(\mathbf{x}-\mathbf{p})) \right|^+ \quad \text{with} \tag{4.11}$$

$$\mathbf{p} = \rho_1 \begin{pmatrix} p_x \\ p_y \end{pmatrix} = \rho_1 \begin{pmatrix} cos(\theta+\pi) \\ sin(\theta+\pi) \end{pmatrix} \tag{4.12}$$

The filter kernel that is defined that way yields high response activations at positions with local luminance contrasts that match the layout of the filter kernel. To achieve insensitivity against the sign of contrast, pairs of equally oriented filters

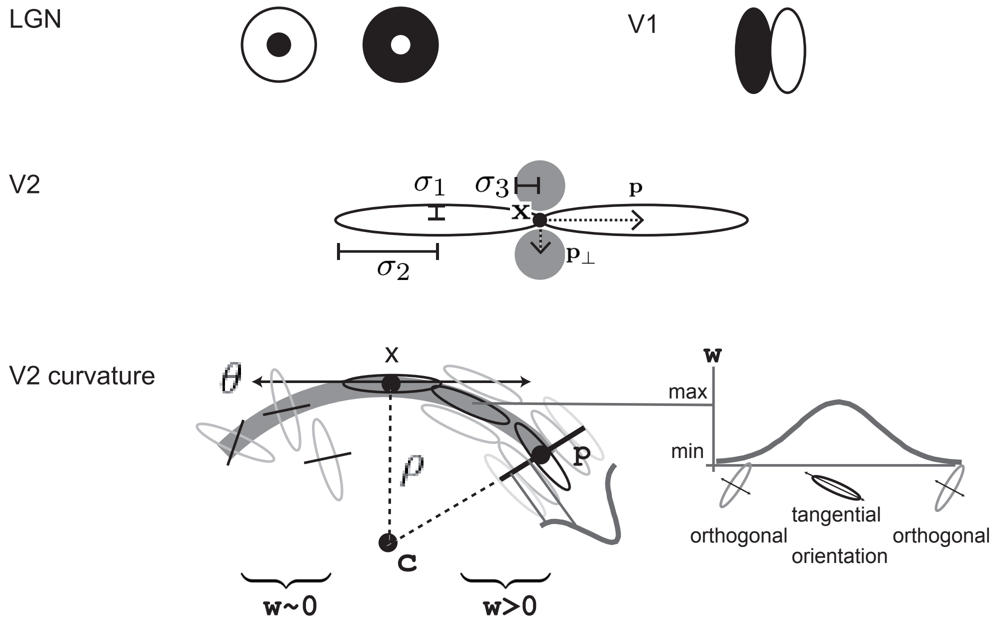


Figure 4.2: Receptive fields of model cells in a qualitative depiction. *From top to bottom: LGN cells and V1 cells for contrast representation. V2 cells integrate from two larger excitatory and two smaller inhibitory regions. V2 curvature cells integrate responses of oriented cell along a curvilinear path that forms around an imaginary central point* **c**. *Integration weights additionally depend on the distance from the cell's reference point* **x**, *angular difference to the tangential trajectory and a function of local radius, indicated here by a Gaussian profile. Correctly aligned orientations that result in a large integration weight w are shown in the right side of the arc, while some that result in weights close to zero are shown along the left arc in this illustration.*

with opposite sensitivity to contrast polarity are used. Such filters populate a set with evenly distributed orientation tunings that represent possible contrast orientations. The locally dominant orientation can be derived by selecting the orientation channel with maximum response, $\theta_{max}(\mathbf{x}) = argmax_\theta(r_\theta^{V1}(\mathbf{x}))$.

4.3.2.2 Model Area V2/V3 complex

Model area V4 contributes four components to our model. Here, we model cells sensitive to contextual influences of contour segments that are arranged in larger spatial extent compared to V1 receptive fields, namely elongated contrasts and curvatures. In addition, border ownership is represented here. In a complex of V2 and V3 this stage represents corners as well.

Elongated contrasts

The integration of elongated contours in V2 makes use of a mechanism that links cells of like orientations over larger spatial distances. The receptive fields are modeled using elongated Gaussian kernels positioned at $\mathbf{p}$ with offset ρ_{ex}^{V2} to the center of the cell. The parameters of the elongated Gaussian kernels are set to build a combined kernel of an elongated integration field, which reflects the highly significant anisotropies of long-range connections in visual cortex [Bosking et al., 1997]. The subfields sample the activations generated by V1 complex cells [Grossberg and Mingolla, 1985, Neumann and Sepp, 1999]. The subfields are combined in a multiplicative fashion. This resembles a logical *and*-operation for the individual subfield activations. Modeled V2 cells only become activated when both subfields receive input. The response is thus able to bridge local gaps in contours. This is in line with physiological findings, as V2 neurons respond to elongated luminance contrasts as well as to illusory contours [Heitger et al., 1998, von der Heydt et al., 1984] like in the Kanisza square (see Fig. 4.4). This integration mechanism is enhanced by local inhibitory effects. Smaller and isotropic integration fields are positioned along an orthogonal axis from the receptive field's center with distance ρ_{inh}^{V2}, building a cross-like zone of excitatory and inhibitory integration, compare [Piëch et al., 2013]. At those positions $\mathbf{p}_\perp$, activity from all orientations is integrated and has an inhibitory effect on the total response. This has a strong suppressive effect on contour fragments that are positioned within a cluttered surround, while isolated boundary segments are not affected. The complete response for an elongated V2 cell is calculated by the steady state equation:

$$\begin{aligned} r_\theta^{V2}(\mathbf{x}) =& r_\theta^{V1} * \mathcal{N}_{\theta,\sigma_{1,2}}(\mathbf{x}+\mathbf{p}) \cdot r_\theta^{V1} * \mathcal{N}_{\theta,\sigma_{1,2}}(\mathbf{x}-\mathbf{p}) && (4.13)\\ & -\gamma r_\theta^{V1} * \mathcal{N}_{\sigma_3}(\mathbf{x}+\mathbf{p}_\perp) - \gamma r_\theta^{V1} * \mathcal{N}_{\sigma_3}(\mathbf{x}-\mathbf{p}_\perp) && (4.14) \end{aligned}$$

with

$$\mathbf{p} = \rho_{ex}^{V2} \begin{pmatrix} cos(\theta) \\ sin(\theta) \end{pmatrix} \tag{4.15}$$

$$\mathbf{p}_\perp = \rho_{inh}^{V2} \begin{pmatrix} cos(\theta+\pi) \\ sin(\theta+\pi) \end{pmatrix} \tag{4.16}$$

Curvature

We also model V2 neurons that respond to more complex stimuli like they appear in contours that form curves or angles. We propose a population of V2 cells tuned to curved contour outlines, r^{V2c}, that allows integration of smooth and even fragmented boundary configurations [Field et al., 1993]. They resemble the functionality of elongated V2 cells but their integration fields are designed such that they are curved. A curvature direction is defined either to the left or the right of the tangent orientation at the target location. The center of curvature **c** defines an osculating circle with given curvature-radius ρ.

The model cells integrate activations from V1 neurons. The integration weight is modeled by a weight function w depending on the current position and orientation tuning where activity is integrated from. This weight function splits into contributions of a distance weight function w^{dist}, an orientation weight function w^{ori} and a function that models the widths of the integrating lobe, w^{width}. Basically, activations from V1 are integrated with maximum weight only when their orientation is tangential to the curvature trace at their relative positions. This yields a sharp tuning of the cell for a certain curvature level. The complete response for an curved V2 cell is calculated by the steady state equation:

$$r_\theta^{V2c}(\mathbf{x}) = \sum_{\phi \in \mathcal{N}(\theta)} \sum_{\mathbf{p} \in \mathcal{N}(\mathbf{x})} w(\phi, \mathbf{p}) \cdot r_\phi^{V1}(\mathbf{p}) \tag{4.17}$$

with ϕ and **p** neighborhoods of orientation and position. As described, w splits into three components. The following weight equations require the center of the osculating circle **c** and the angle ξ between the position **x**, the current integration positions **p** and that center **c**. The calculation of **c** makes use of a parameterizable curvature radius ρ.

$$\begin{align} w(\phi, \mathbf{p}) &= w^{dist}(\mathbf{p}) \cdot w^{ori}(\phi, \mathbf{p}) \cdot w^{width}(\mathbf{p}) \tag{4.18} \\ \mathbf{c} &= \mathbf{x} + \rho \cdot (cos(\theta), sin(\theta))^T \tag{4.19} \\ \xi &= \measuredangle(\overrightarrow{\mathbf{cp}}, \overrightarrow{\mathbf{xc}}) \tag{4.20} \\ & \tag{4.21} \end{align}$$

The overall weight is calculated using contributions of the other weight functions:

$$\begin{align} w^{dist}(\mathbf{p}) &= exp(-\frac{\|\overrightarrow{\mathbf{xp}}\|^2}{\sigma_1^2}) \tag{4.22} \\ w^{ori}(\phi, \mathbf{p}) &= sin(\xi - \theta) \tag{4.23} \\ w^{width}(\mathbf{p}) &= exp(-\frac{(\|\overrightarrow{\mathbf{pc}}\|^2 - \rho)^2}{\sigma_2^2}) \tag{4.24} \end{align}$$

Here, w^{dist} decreases with distance from the response center **x** using an exponential function with deviation σ_1. w^{ori} is modeled with a simple sine function, such that tangential orientations receive a higher weight. Finally, w^{width} generates highest weight when the integrating position $\mathbf{p}$ is on the curvature radius ρ with a second deviation parameter σ_2.

Border Ownership

Cells in visual cortex V2 also show selectivity to the figure-ground arrangement of the scene in the visual field [Williford and von der Heydt, 2013]. So-called border ownership cell responses are elicited when a figure of arbitrary shape is presented on their preferred side with respect to the center of their receptive field. From the same group, [O'Herron and von der Heydt, 2013] have also shown that during visual motion caused by eye motion or object motion, these border ownership signals are remapped to different neurons. The visual system uses this information to resolve depth arrangements in the stimulus [Qiu and von der Heydt, 2005]. The pointing direction of border ownership cells indicates the direction of the frontal surface at every image location. This reflects to commonly known Gestalt rule that a boundary is owned by the frontal figure.

We model border ownership cells by a retinotopically arranged population representing four potential directions where the figure can be positioned relative to the cell's center. Border ownership responses are initially isotropic and only occur together with local contrast activations. Cells indicating opponent border ownership direction are mutually rivaling in our model. The complete response for an border ownership V2 cell is calculated by the steady state equation:

$$r_{\lambda}^{V2B} = \begin{cases} f(r_{\theta}^{V2}) & when \quad \lambda \perp \theta \\ 0 & when \quad \lambda \parallel \theta \end{cases} \tag{4.25}$$

The function $f()$ only maps activations of V2 to activations of V2B. In the model, this functions is a linear transformation to account for different numerical activations levels that occur in the model.

With λ four border ownership directions $[0°, 90°, 180°, 270°]$. The mutual competition between activations indicating opposing border ownership directions r_a^{BO} and r_b^{BO} is calculated by

$$\dot{r}_a^{BO} = -\alpha \cdot r_a^{BO} + A(1 - r_a^{BO}) - \beta \cdot r_b^{BO} \tag{4.26}$$
$$\dot{r}_b^{BO} = -\alpha \cdot r_b^{BO} + A(1 - r_b^{BO}) - \beta \cdot r_a^{BO} \tag{4.27}$$
$$\tag{4.28}$$

Corners

Based on empirical evidence of neural representations generated by cells selective to multiple orientations [Felleman and Van Essen, 1987, Ito and Komatsu, 2004, Anzai et al., 2007] we incorporate model representations of corners in a dedicated model area V2/V3 complex. We build upon the proposal developed in [Weidenbacher and Neumann, 2009] that corner and junction configurations can be made explicit by specific read-out mechanisms. Here, we employ a simplified version as of [Hansen and Neumann, 2004] to generate corner representations by grouping V1 responses of orthogonal orientation fields. In a steady-state formalism the response reads

$$r_\theta^{V2/V3} = \left| r_\theta^{V1} \cdot r_{\theta+\pi}^{V1} \right|^+ . \tag{4.29}$$

4.3.2.3 Model Area V4

Inspired by experimental evidence model neurons in model V4 integrate responses of V1,V2 and V2/V3 to achieve a selectivity that considers large-scale boundary fragments as well as local variations in curvature and a selectivity for corners [Pasupathy and Connor, 1999, Yau et al., 2013a]. Curvature selective cells are modeled in a two-stage cascade of mechanisms. The first level integrates V2 contour responses and is selective to curvature directions, left or right (relative to the cell's orientation preference). The second level combines opposite curvature directions into one response, like in V1 complex cells. This model mechanism differs from the one proposed by [Rodríguez-Sánchez and Tsotsos, 2012] which utilizes single stage filter computations. In this approach specific subfield mechanisms sensitive to orientation, tangential contour outline and scale are combined in a non-linear fashion to selectively respond to contour fragments of different curvatures. We develop a mechanism that is distributed over different stages to first group responses to extended contour outlines in V1 and V2 suppressing non-contour clutter. In the case of sharply localized corners and junctions the dedicated representations of localized multi-orientation responses will be activated. Those responses of grouping cells (or the junction representations) are integrated at the subsequent stage. Here, curvature selectivity is made explicit that distinguishes left and right curvatures. Different integration scales generate selectivity to curvature. This distribution allows to associate regions of high contour curvature at an intermediate scale with localized outline details at the finer scale. The model cell responses in our model are described by the following equations:

$$\dot{r}_{\theta}^{V4left} = -\alpha_4 r_{\theta}^{V4left} + (1 - r_{\theta}^{V4left}) \cdot A_{\theta} - (1 + r_{\theta}^{V4left}) \cdot B_{\theta} \tag{4.30}$$

$$\dot{r}_{\theta}^{V4right} = -\alpha_4 r_{\theta}^{V4right} + (1 - r_{i\hat{\theta}}^{V4right}) \cdot A_{\theta} - (1 + r_{\theta}^{V4right}) \cdot B_{\theta} \tag{4.31}$$

with

$$A_{\theta}(\mathbf{x}) = r_{\theta}^{V2} * \mathcal{G}_{\theta,\sigma_{4,4b}}(\mathbf{x} + \mathbf{p}) \tag{4.32}$$

$$B_{\theta}(\mathbf{x}) = r_{\theta}^{V2} * \mathcal{G}_{\theta,\sigma_{4,4b}}(\mathbf{x} - \mathbf{p}) \tag{4.33}$$

$$\mathbf{p} = \rho^{V4} \begin{pmatrix} (cos(\theta) \\ sin(\theta) \end{pmatrix} \tag{4.34}$$

These responses are calculated at equilibrium and averaged subsequently, leading to the model V4 filter response

$$r_{\theta}^{V4} = \frac{1}{2} \frac{|(A_{\theta} - B_{\theta})|^{+} + |B_{\theta} - A_{\theta}|^{+}}{\alpha_4 + A_{\theta} + B_{\theta}}. \tag{4.35}$$

This integration mechanism yields a response for locally curved boundary segments at a larger spatial scale. For elongated contour segments that show no curvature, the response of individual cells will be equal and the combined response very low.

4.3.2.4 Model Area IT

So far, we have described how our model integrates local features from model V1 into elongated, potentially curved boundaries at model V2-V4. Model area IT performs contextual integration that allows a segregation into figure and ground and a representation of prototypical objects at a large spatial scale. As discussed above, a population of V2 cells responds selectively to the direction of figure-ground direction. The local representation of border ownership at model V2 represents a set of available local hypotheses that cannot locally be resolved, as this step requires contextual influence from a larger spatial surround. Cells in IT cortex have been shown to be shape selective with properties generalizing over contrast polarity and mirror reversal [Baylis and Driver, 2001]. The authors demonstrate that such cells do not, however, generalize over the assignments of figure-ground direction. The investigation supports the view that the population of probed IT cells is mainly driven by the sidedness of contours and less so by the contour itself. Given the rapidness of ownership selectivity observed in V2, we propose that ownership computation relies on a network of V2-V4-IT cell interaction. Our model uses local shape configuration in the outline of an

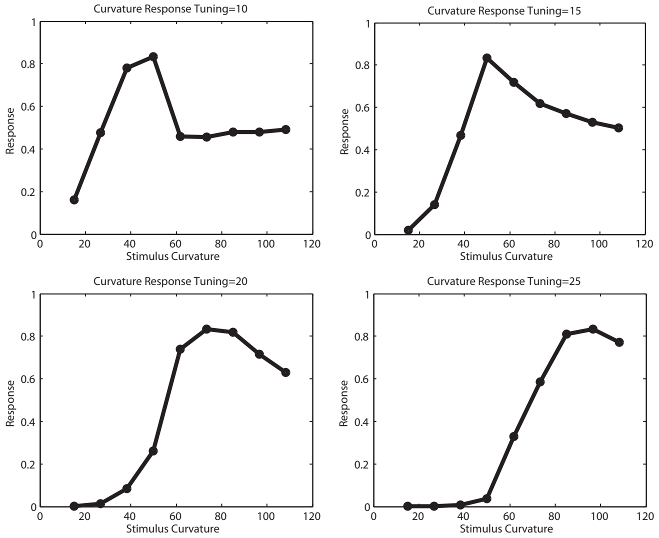


Figure 4.3: Tunings of different curvature cells. *The x-axis show the curvature of the presented stimulus, the y-axis the response strength of a curvature cell tuned for* $10, 15, 20$ *and* 25 *pixels curvature radius in model V2, which correlates for a curvature of* $40, 60, 80$ *and* 100 *pixels in model V1. For the smaller curvature radii, subsampling artefacts cause the tuning function to be less smooth.*

object to collect confidence about the direction of figure and ground. We adopt an approach of [Zhou et al., 2000] and model an integration cell at level IT that integrates border-ownership hypotheses from a larger spatial extent from model V2 input. For each location in the image, border ownership activations in a local neighborhood that point towards the inside of the respective receptive field contribute to the activation of an IT cell. This results in strong responses in model IT where local image regions are surrounded by contour convexities. Local activities of border ownership cells in model V2 then receive a positive enhancement if they contributed to such an integration process. This recurrent architecture resolves the initially ambiguous assignment of border ownership. Taken together, this makes the model belong to the class of feedback architectures according to the categorization in [Williford and von der Heydt, 2013]. The response of cells and their interaction is denoted by the following equations:

$$r_{\theta}^{IT}(\mathbf{x}) = \sum_{\phi\in\mathcal{N}(\theta)} \sum_{\mathbf{p}\in\mathcal{N}(\mathbf{x})} f(\mathbf{x},\phi,\mathbf{p}) \cdot r_{\theta}^{V2}(\mathbf{p}) \cdot exp(-\frac{(\rho^{IT} - \|\overrightarrow{\mathbf{xp}}\|)^2}{\sigma^2}) \tag{4.36}$$

with

$$f(\mathbf{x},\phi,\mathbf{p}) = cos(\measuredangle(\overrightarrow{\mathbf{xp}},\phi)) \tag{4.37}$$

Such an IT cell at position $\mathbf{x}$ integrates responses of V2 cells in its proximity $\mathbf{p}$. The integration weight function $f(\mathbf{x}, \phi, \mathbf{p})$ depends on the angle between $\mathbf{x}$ and $\mathbf{p}$ and the currently integrated orientation ϕ. This grants orientations parallel to an imaginary line toward x_0 high weights, while orthogonal orientations receive low weights.

This model area receives connections from the early as well as from the intermediate functional stages V1 and V2 where curvature is represented. This means that high-resolution local cues as well as contextual cues like corners from a larger region are available. A shape can thus be described as a set of contributing prototypical elements that contribute to the local configuration at every image location. Those elements are not solely generated through integration of lower areas, but exist as a distributed representation in all modeled areas and profit from mutual interaction through feedback and exhibit dynamic processes when a stimulus is presented.

4.3.3 Contextual influence

The following sections will explain the contribution of contextual influence on local representations of features.

4.3.3.1 Contour enhancement

The mechanisms so far presented contributed to the feed-forward sweep of the model. We stated earlier that in visual cortex (and in neural processing in general), the input of cortical areas of higher stages highly contribute to the performance of individual earlier areas. By such recurring connections, contextual information is introduced in lower regions. We are thus now going to focus on the recurrent connections that are incorporated in our model.

Let's briefly recall that we model feedback connections that have a modulatory effect [Girard and Bullier, 1989] as outlined in Sect. 2.3.2, eq. 2.7. We mentioned that a feedback signal alone cannot elicit responses as long as no input activation is present. On the other hand, feedback that matches input configurations will increase those activations. We stick to this convention throughout our following elaborations.

V2 long-range and curved cells represent continuous straight or curved contours. Their multiplicative combination of receptive field subcomponents caused the cells to elicit responses whenever a contour of matching orientation was presented in their receptive fields. Now, those cells in V1 that contributed to the integration process will receive feedback and be thus increased in activity. The following non-linear transformation stage increases the difference in response strength with respect to other oriented contour cells that did not receive feedback. At the subsequent normalization stage, local response levels are now slightly increased by the recurrent input. Now, surrounding activations without feedback have a competitive disadvantage and receive a higher divisive normalization relative to their activation due to the increase response in their neighborhood that contributed to the sum.

The dynamics of these interactions are denoted in formal terms and can be found in Chap. 2.

4.3.3.2 Curvature representation

As stated earlier, the modeled V4 cell do not at all or only marginally respond to straight elongated contours. Responses of V2 cells to curved boundaries are integrated in model V4, where integration cells sensitive to opposite sign of curvature mutually compete for equal orientations. These cells respond at positions with a local curved contour configuration, but are silent at elongated straight contours. Feedback is generated for those V2 cells that contribute to those curved boundary segments the corresponding model V4 cells respond to maximally. Regions with curved boundary segments thus elicit a strong response of V4 cells while regions with mostly straight contours do not elicit such a strong response. This signal can thus be used to differentiate regions of many straight contour segments from regions with many curved contours. In formal terms, the V2-V4 cell interactions are defined by

$$\dot{r}^{V2c} = -\alpha r_\theta^{V2c} + (1 - r_\theta^{V2c}) \cdot A(1 + r_\theta^{V4}) \tag{4.38}$$

$$A = r^{V2c} * \mathcal{G}_\sigma \tag{4.39}$$

4.3.3.3 Figure-ground segregation

The contribution of feedback to figure-ground segregation is twofold in our model. First, at a local level, hypotheses of border ownership are generated by intra-area recurrent connections from long-range grouping cells. Contextual feedback from model IT resolves the remaining ambiguities. Initially, all directions of border ownership have equal likelihood along contrasts. With increasing confidence about local contrast orientations generated by V1 and V2, two options for border ownership directions are discarded and only two orthogonal border ownership directions remain. Activations of long-range V2 cells that indicate elongated surface boundaries and their orientation locally increase activities of those border ownership cells that are directed perpendicularly to the orientation of the boundary. Activity normalization for V2 border ownership cells then leads to a suppression of activities for ownership directions orthogonal to the boundary orientation. Formally, this is accomplished by the dynamics

$$\dot{r}_{\phi}^{BO} = -r_{\phi}^{BO} + \beta r_{\theta}^{V2} - r_{\phi}^{BO} \cdot q_{\phi}^{BO} \tag{4.40}$$

$$\dot{q}_{\phi}^{BO} = -\alpha q_{\phi}^{BO} + \sum_{\gamma} r_{\gamma}^{BO} \tag{4.41}$$

$$\theta = \phi + \frac{1}{2}\, mod\, \pi. \tag{4.42}$$

Second, V2 border ownership cells receive feedback from cells in model IT. Here, border ownership as well as figural cues, e.g., from local junctions, or curvature maxima, were integrated by IT cells. For the correct inference of figure and ground, feedback from IT to V2 is essential. Figure-Ground cells at IT level integrate border ownership activations from V2 in a circular fashion to integrate the coherence of directions indicating a convex pattern of figure outline. In the feedback sweep, this contextual information is now fed back to these border ownership cells compatible with the configuration using recurrent connections. In formal terms, this extends the dynamics presented in eqn. 4.40 above by incorporating a modulating feedback signal from model IT cells, namely

$$\dot{r}_{\theta}^{BO} = -r_{\theta}^{BO} + \beta(r_{\theta}^{V2} + h_{tonic}) \cdot (1 + \lambda_2 \cdot r_{\theta}^{IT}) - r_{\theta}^{BO} \cdot \sum_{\gamma} r_{\theta}^{BO} \tag{4.43}$$

This also concludes the feedback sweep of our recurrent model. In the following section, we will show the performance of the model and its individual areas in the results section.

4.4 Results

In this section we illustrate the capabilities of our model in a number of simulations. To demonstrate how the model processes shapes, we use some artificial images to show working principles of various subcomponents of our model. These simple shapes were taken from the *Webdings* font freely available with a Microsoft$^{®}$ Windows^{TM}8.1 operating system. We also include also a depiction of a *Kanisza* square [Kanizsa, 1955]. This is a special stimulus because it elicits the perception of illusory contours at the outline of the occluding square, a sensation our model is also capable to represent.

To demonstrate the abilities of our model to process real world images we acquired the dataset of [Fowlkes et al., 2007] and selected a few examples that we included in our results sections. These images have a resolution of 321×481 pixels in landscape or portrait orientation. They were converted to gray-scale images

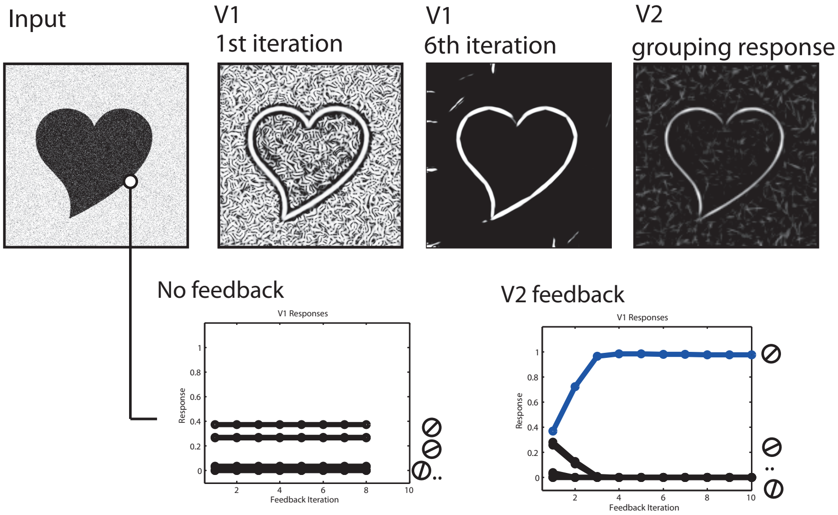


Figure 4.4: Results of early processing stages V1 and V2. *The two plots indicate time courses for V1 activations. Initially, multiple V1 neurons are activated due to a broad tuning width (first plot). Without feedback, this effect prevails through iterations. With feedback, the correct orientation (blue) receives feedback and gradually reduces activations of other orientations (second plot).* Bottom: *Example how model V2 neurons show responses at positions formed by illusory contours (in green circle) due to contextual integration.*

using the Mathworks® Matlab® `rgb2gray` function which performs a perceptionally weighted combination of the red, green and blue channel. We used 8 to 12 iteration steps to allow recurrent feedback signals to build up. The angular resolution of cell populations is defined by selecting eight $\frac{\pi}{8}$ steps to encode orientation. Border ownership is represented by a population representing 4 directions. Model V2 curvature cells also used 8 orientations for tangential orientations, but due to two possible curvature directions, our model contains a population of 16 curvature cells. A list of parameters used is given in table on page 103.

4.4.1 Early processing stages

To begin with, we show how the processing at early stages achieves a representation of the stimulus concerning contrasts and elongated contours. Local contrasts are represented in the early stages by model V1 and V2 cells. However, as can be seen in Fig. 4.4 the responses rapidly change in the first few iteration steps. The contained contour as well as the added noise signal both elicit responses at the V1 level (*second column*) and cause the shapes outline to be not clearly separated from the background. However, those responses are grouped into elongated contour representations in model V2 (*4th column*). Elongated contour segments

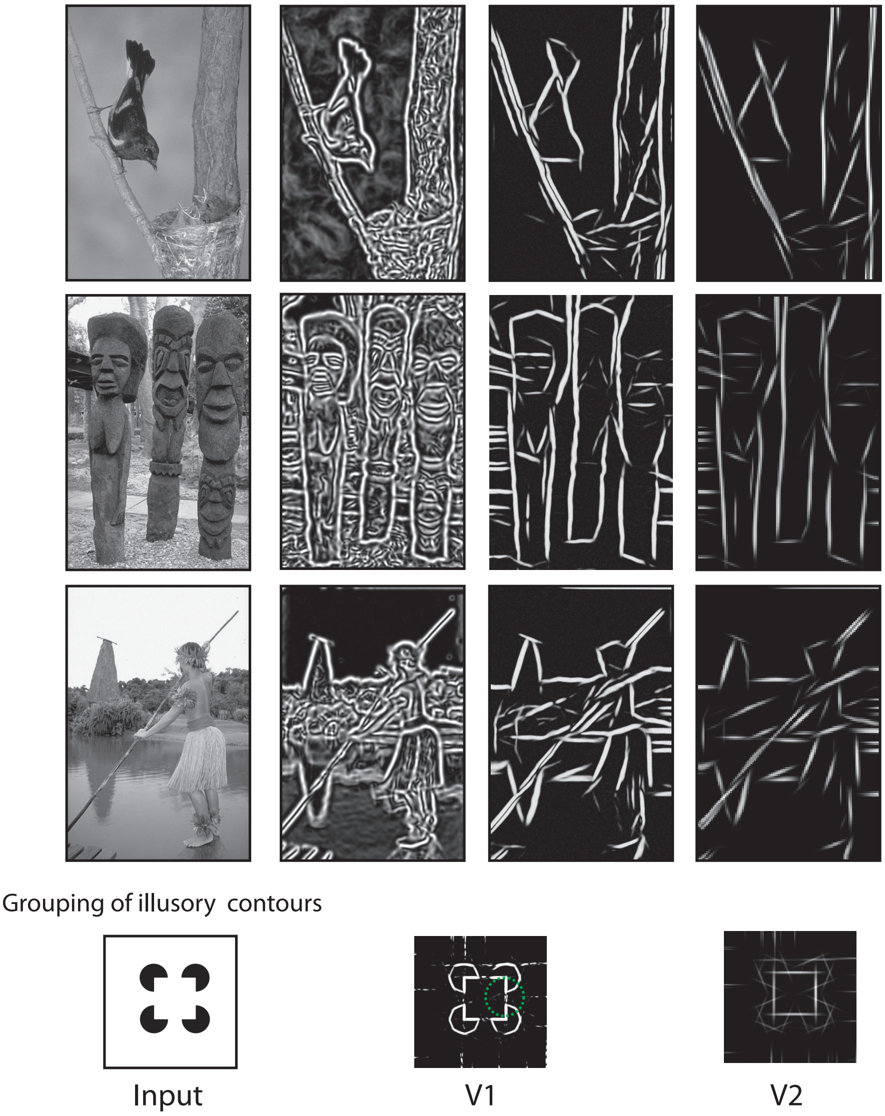


Figure 4.5: Results of early processing stages V1 and V2. Left column: *Initial input images.* Second and third column: *Cumulated responses of model V1 neurons at the initial processing iteration and a few iterations steps, respectively.*Fourth column: *Responses of model V2 neurons. Elongated edges formed by like-oriented contrasts are grouped as reflected by responses at respective locations. This stage also shows activations for illusory contours contours (* third row*) at the gaps between contrasts.*

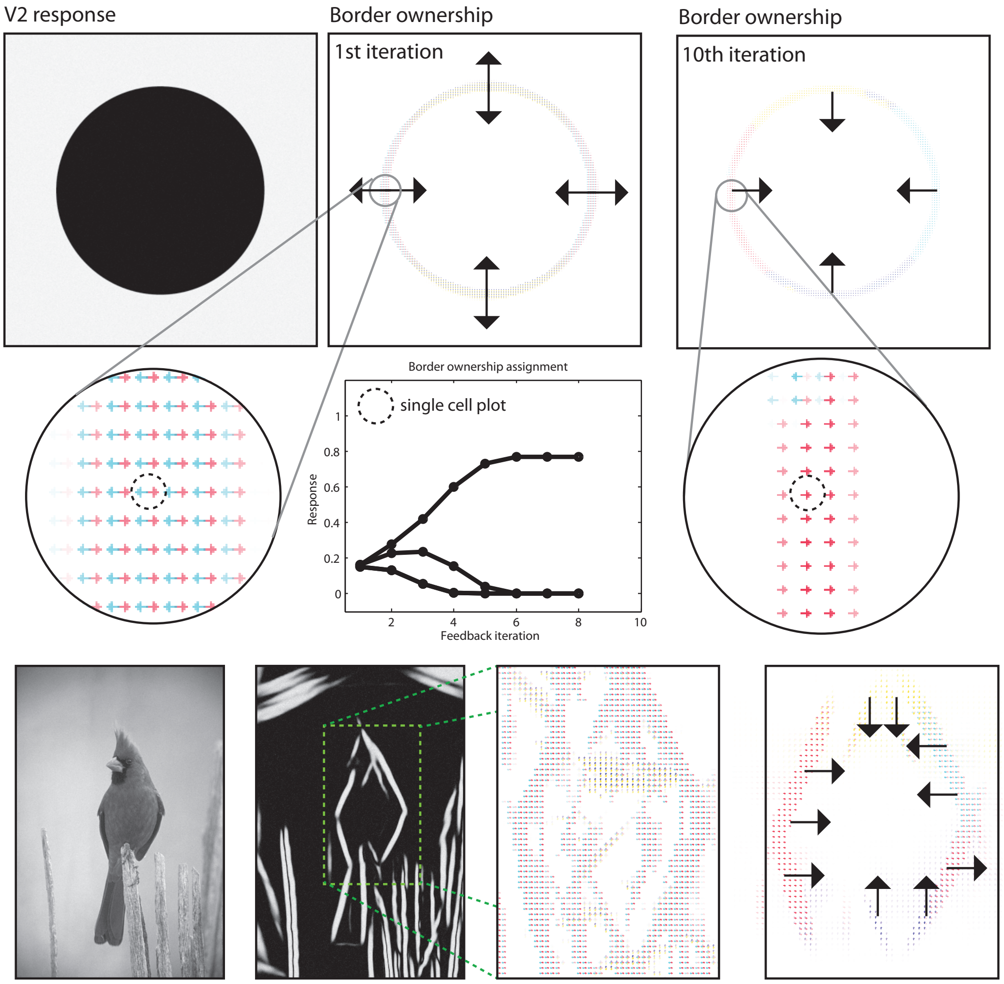


Figure 4.6: Results for border ownership assignment. First two rows: *Cells at model V2 indicate the direction of figure side at positions where boundaries exist. Initially, four hypotheses exist for possible figure direction. These are refined in model V2, where only two hypotheses remain after the orientation of the boundary is represented. Contextual integration in model V4 then provides correct estimates with modulatory feedback. Subsequent normalisation and mutual competition leaves only one hypotheses for border ownership direction. See text for details on the time phases of border ownership assignment.* Third row: *Demonstration of the boundary assignment for a natural image. The intial responses are improved after a few iteration steps.*

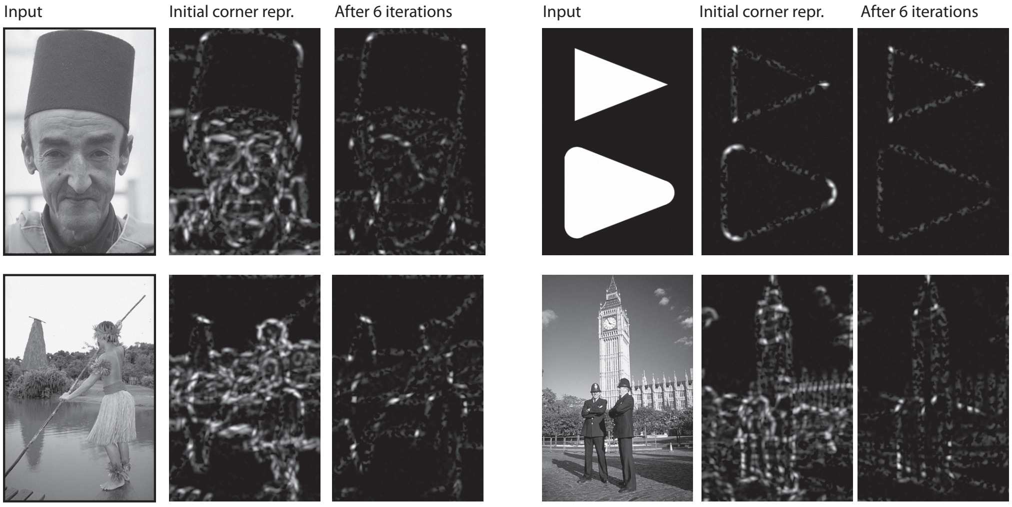


Figure 4.7: *Corner representation in model V2/V3.* For each group of three pictures: *Initially, responses of model V1 did not yet benefit from contextual feedback of model V2 neurons. Corner representation is thus distorted by noise (*second row, middle*). After a few iterations, when V1 responses have been modulated V2 feedback, the corner representation is much clearer.*

are clearly emphasized. From these V2 activations, a recurrent feedback signal is generated that modulates V1 activations. After a few iterations, the representation at V1 dramatically changed, with the outline of the figure now clearly visible.

The effect of the feedback signal is also measurable in a quantitative way, see Fig. 4.4, *right.* Along the boundary of an object we plotted the activation levels of the population of V1 neurons that represent the orientation. Initially, the neuron with preferred orientation responds best, but also those with orientation tunings close to the real contour (*first plot*). The situation changes when feedback is added (*second plot*). Now, representations of undesired orientations are attenuated and the activation of the cell representing the contextually valid orientation is highly increased.

Also in Fig. 4.4, the representation of illusory contours at V2 stage is depicted. This is illustrated using an input depicting a *Kanisza* square (*last row*). A complete square is highly salient for human observers despite the fact that only a series of circles with cut-out corners are depicted. This is reflected in the grouping responses of V2 neurons, they also show activity in the gap between the real contour fragments. Figure 4.5 shows V2 responses for the same parameter set and for a parameter set with changed receptive field sizes, to illustrate the effect even stronger (*framed part*).

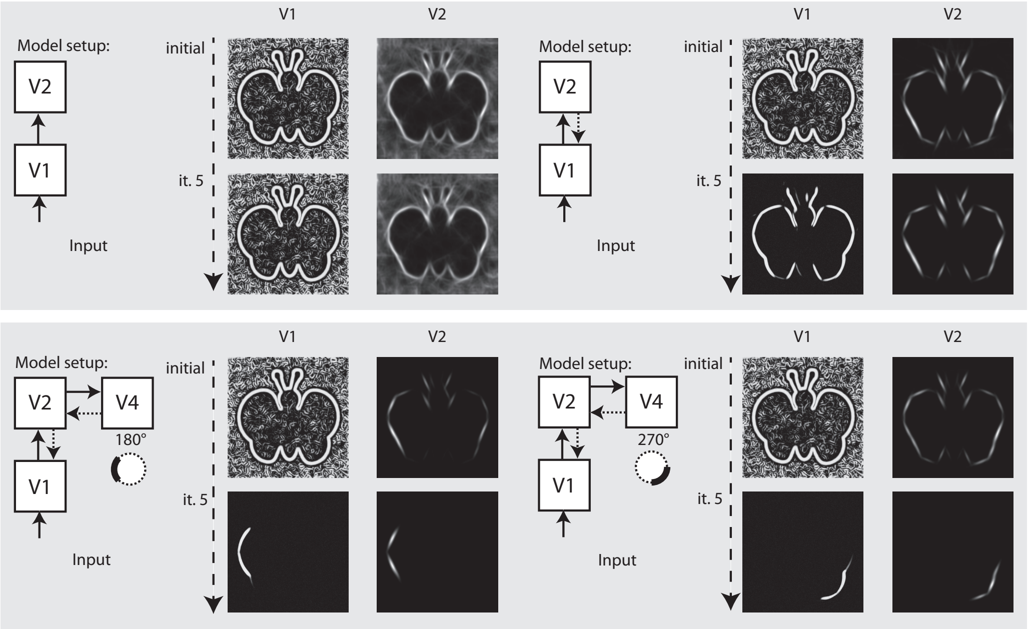


Figure 4.8: *Representation of curved boundary segments and effect of feedback.* Top left:*: Model setup without feedback connections. The intial representation at model V1 and V2 undergo no change.* Top right: *Feedback from V2 causes a refinement of elongated structure within a few iterations.* Bottom left: *Feedback from V4 allows the accentuation of boundary segments with distinct curvature strength and direction. Here, a curvature segment as found on the left part of the shape is highly emphasized by feedback.* Bottom right: *Same as* left, *but with selectivity for another segment of the shape. Note that while boundaries with the same orientation are present in the stimulus, only the one with matching curvature is emphasized.*

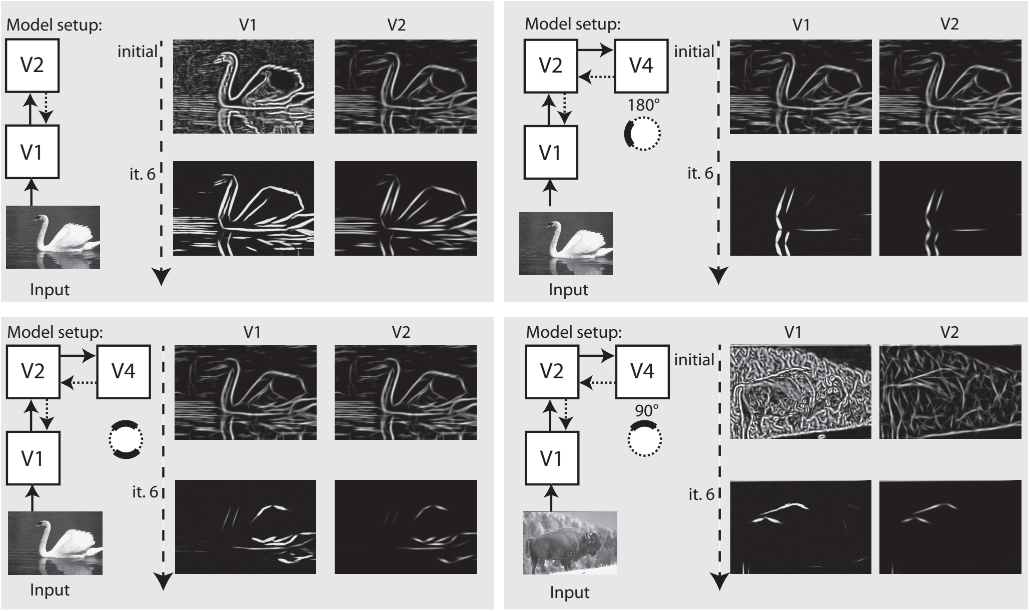


Figure 4.9: *Refinement and modulation of shape contours in a real world example.* Left: *Within 5 iterations, the outline of the animal is very well visible at V1 and at V2 stage.* Right: *With modulatory feedback from model V4, various parts of the animal like those contours with a certain curvature and orientation can be emphasized.*

4.4.2 Curvature tuning

Figure 4.3 illustrates the tuning functions we defined for model V2 curved cells. A curved cell with distinct radius tuning was selected and we presented arcs of different curvature to this cell and simulated the response. We performed this for four cells with curvature tuning to $10, 15, 20$ and 25 pixels radius. This curvature definition happens in V2, where the initial resolution of the image had been subsampled. For this reason, the value here correspond to values $40, 60, 80$ and 100 in V1 resolution. In each plot, the peak response occured when the stimulus with the matching radius was presented. In this simulation, subsampling artefacts cause the first two plots to elicit some discontinuities.

4.4.3 Shape representation

In the final setup, we show how our model independently represents different elements of a shape, and how this depends on the recurrent feedback connections. Fig. 4.8 illustrates the results we achieved for an artificial image. Initially, we configured the model to only use feed-forward connections from V1 to V2. The model only achieves an representation at model V1 and a representation at V2 where the elongated boundaries are visible, but surrounded by many spurious activations. When recurrent feedback from V2 is added, the representation at V1 improves in the first few iterations before a steady representation is reached. In parallel, elongated boundaries at V2 are integrated and noise is highly reduced.

To represent prototypical objects at an intermediate level of detail, we stated that the model needs to represent different contour properties. In the second row of Fig. 4.8 we show how the model achieves to emphasize V1 responses when they contribute to a certain contour fragment with desired properties. We deliberately exaggerated the effect and chose a very narrow tuning so that all other responses become almost completely suppressed. On the *left* side, we let the model emphasize contour parts that are oriented almost vertical but in a curved context of a matching radius. As can be seen, the model highlights that parts on the left side of the stimulus that matches and leaves others suppressed, even if their local orientation would match. On the *right* side of Fig. 4.8, we perform the same for a different part of the shape outline.

In Fig. 4.9 we perform the same selection for a realistic photograph depicting an elephant. On the *left* side, we show interaction of model V1 and V2 causes an appealing representation of the animal at stages V1 and V2. On the *right* side, we configured the model using model area V4 to emphasize parts of the outline of the animal that match a certain context and configuration, here, a part of the outline.

4.4.4 Border ownership and figure-ground assignment

In the segregation of a scene into figure and ground the modeled *border ownership* cells participate by indicating the direction where the frontal surface is positioned at a boundary [Zhou et al., 2000]. Our model incorporates a mechanism using such border-ownership cells to resolve the direction of a frontal surface from local boundary cues [Zhou et al., 2000]. We performed such a assignment for our sample images, see Figure 4.6 for an illustration of the result. The output of model area V1 and of V2 long-range integration cells are acquired to generate initial hypotheses of border ownership direction at image regions where local contrasts are situated. Initially, all four border ownership directions show equal responses at a boundary location. After stimulus onset, three dynamic effects occur and their contribution to the resolution of border ownership is reflected in the time course of cell activation, see Fig. 4.6 for an illustration.

First, local feedback from V2 cells enhances two hypotheses of border ownership for the directions orthogonal to the local boundary orientations.

A local normalization causes an attenuation of the other two representations (Fig. 4.6, *second row*; Timestep 0 and 1). Second, shape-level integration at model area IT contributes positive feedback to those border ownership cells that are directed towards the inside of the figural depiction. Again, normalization leaves the net response of the cells constant (timesteps 2-4). Finally, mutual inhibition among border ownership cells with opposite direction selectivity causes the dominant direction to gain all available net energy (timestep 5 to 8). At this point, a stable point is reached and the local ambiguity for border ownership direction is resolved using feedback from higher cortical areas. The interpretation of the final representation would be that the frontal surface is to the inside of the curved boundary.

Table 4.1: *Parameter values used to generate the results of the real-world examples. References to equations are given in the second column.*

General Model Parameters used for simulations		
description	see equation	value
Number of orientations used		8
Number of feedback iterations		6
Number of BO directions		4
Model Area V1		
Network size		321 × 481
LGN σ		1.00
LGN κ		1.50
LGN Normalization α	(4.9 applied)	2.17e-03
LGN Normalization β	(4.9 applied)	2.17e-03
V1 Contrast σ_1	4.11	0.23
V1 Contrast σ_2		0.12
for **p**: Excentricity ω_1		3.00
V1 Normalization α	(4.9 applied)	2.16e-06
Non-Linearity of V1 responses	(4.6 applied)	4.00
Model Area V2/V3		
V1:V2 subsampling	4.13	1: 3
Network size		107 × 161
Filter size of V2 complex		41.00
V2 complex cell $\sigma_1, \sigma_2, \sigma_3$		0.21,0.02,0.10
in **p**: ω_{ex}^{V2}		3.70
in **p**: ω_{inh}^{V2}		2.50
V2 Inhibition Strength γ		0.10
V2 Nonlinearity k	(4.6 applied)	1.00
Strength of V2-V1 feedback		0.11
Non-Linearity of BOwn		1.00
In **c**: curvature radius ω^c		15.00
in w^{dist}: σ_1		40.00
in w_{ori}: σ_2		2.00
Strength of R^{V2Curv} feedback		0.15
Model Area V4		
V1:V4 subsampling		1: 4
Network size		81 × 121
V4 filter size		31.00
α_4	4.30	0.01
σ_4,σ_{4b}	4.32	0.43,1.35
ω^{V4}	4.32	-1.00
Model Area IT		
σ_{IT}	4.36	0.43
σ		29.00
ω_{IT}		17.00
IT Non-Linearity		3.00
α	(4.9 applied)	5.49e-05
Strength of BOwn feedback		50.00

4.5 Discussion

4.5.1 Summary of contributions

In this contribution we emphasized the role hierarchical representations have in the organization of shape features and their combinations into a coherent form. Like some previous model developments [Cadieu et al., 2007, Rodríguez-Sánchez and Tsotsos, 2012, Hatori and Sakai, 2012] our model is based on low and intermediate representations of shape features. These proposals are all based on a strictly hierarchical feed-forward processing sequence. We propose here that such shape encoding mechanisms may be based on distributed representations that are established by interacting assemblies each devoted to specific feature properties. Such interactions in the model are organized by recurrent interactions of feed-forward and feedback signals. The underlying structural principles are based on the cortical architecture of the ventral pathway with mutual interactions between such distributed representations [Markov et al., 2013]. The model architecture incorporates principles that have been predicted to minimize the computational efforts of visual systems to successfully deal with the complexity problem of perception [Tsotsos, 1988] (compare also [Tsotsos, 2005]). Among those, the hierarchical organization of representations in model areas, the specific receptive field properties of model columnar mechanisms, hierarchical pooling of spatially separated input representations, and top-down (modulatory) feedback are proposed here to account for the functional properties of cortical shape processing. We did not discuss complexity advantages in this contribution. However, given the theoretical predictions by such earlier work our proposal of a model architecture provides a evidence how distributed intermediate-level mechanisms may help to shape our understanding of modeling complex visual machinery that captures key cortical principles.

The main contributions of this chapter are twofold. First, we propose a computational network architecture that utilizes a hierarchical distributed representation of shape features. Contour features play a major role to track moving shape in which their strength parametrically change as a function of their saliency [Caplovitz and Tse, 2007]. This necessitates global configurational as well as local information to distinguish rather tiny differences in the outline of a 2-dimensional form (such as curved boundaries vs localized corners, [Pasupathy and Connor, 1999, Ito and Komatsu, 2004]). In order to generate a representation with sufficient spatial resolution combined with spatial context we suggest that multiple specialized component representations interact by feed-forward hierarchical processing that is combined with feedback from representations generated at higher stages in the hierarchy. Second, we incorporate grouping mechanisms to integrate like-oriented contour responses that are integrated if they form a smooth outline fragment of a surface boundary [Grossberg and Mingolla, 1985, Neumann and

Sepp, 1999, Ben-Shahar and Zucker, 2004]. Such grouping mechanisms operate at the stage of area V2 and are, thus, involved in the hierarchical processing of shape. Given the hierarchical processing and representation of boundary information in the ventral pathway (see the overview in [Neumann et al., 2007]) the shape processing observed in area V4 is mainly driven by the output of grouping responses. It may be supplemented by input from simple/complex cells in V1, a principle of convergent signal streams also used in the models described in [Rodríguez-Sánchez and Tsotsos, 2012, Thielscher and Neumann, 2003]. In addition, we suggest that the shape representation built at the stages of V4 and IT influences the assignment of border ownership in surface representation [Zhou et al., 2000] (see overview in [Neumann et al., 2007]). Model IT cells send modulatory feedback to those V4 cells that provide relevant input (in V4 and V2) such that the net sum of convex corners/curvatures determines the ownership direction. The proposed model thus combines separate findings about the generation of cortical shape representation with figure-ground segregation mechanisms by assigning border ownership.

4.5.2 Relation to previous models of shape representations

Visual shape recognition has already been investigated intensively by considering the 3-dimensional (3D) surface appearance for object recognition [Riesenhuber and Poggio, 1999, Serre et al., 2007, Mutch and Lowe, 2008, Yamane et al., 2008, Serre and Poggio, 2010] as well as 2-dimensional (2D) shape recognition [Schwartz et al., 1983, Mokhtarian and Mackworth, 1986, Mokhtarian, 1995, Rodríguez-Sánchez and Tsotsos, 2012]. In the context of view-based models of object recognition stable views [Logothetis et al., 1995] are associated with 2D shapes so that their analysis can be considered as an intermediate stage of object processing [Cadieu et al., 2007]. The computational model approaches of 2D shape representation can be subdivided into flat and hierarchical schemes. Examples of flat processing schemes, e.g., utilize Fourier descriptors [Schwartz et al., 1983], multi-scale representations of curvature features in the shape outline [Mokhtarian and Mackworth, 1986, Mokhtarian, 1995], or global schemes for integrating oriented line features [Wilson and Wilkinson, 1998]. Hierarchical multi-layer processing schemes are based on different stages to generate an increasingly coarse-grained representation of shape features utilizing repetitive application of local filtering operations [Riesenhuber and Poggio, 1999, Cadieu et al., 2007, Rodríguez-Sánchez and Tsotsos, 2012]. In order to resemble the feature selectivity of V4 cells in monkey cortex such cells build coarse-grained orientation-curvature representation of the shape under inspection. The hierarchical organization of a sequence of processing stages follows the idea of the Neocognitron [Fukushima, 1980, Fukushima, 1988] by developing low and intermediate representations of richer shape feature compositions [LeCun et al., 1998, Riesenhuber and Poggio, 1999, Mutch and Lowe, 2008, Tabernik et al., 2014]. The orientation-curvature representation of V4 cells reported by [Pasupathy and Connor, 1999, Connor et al., 2007] has been investigated in the models reported in [Cadieu et al., 2007, Rodríguez-Sánchez and Tsotsos, 2012, Hatori and Sakai, 2012]. We share the principles of the hierarchical organization of processing and the emergence of rich orientation-curvature sensitivity in our proposal. Initial processing utilizes orientation sensitive filters to extract local oriented contrast. Unlike the previous models we incorporate a stage of boundary grouping at the interface between low and intermediate levels of representation. Such grouping operations integrate oriented contrast responses that are arranged in the local neighborhood of a target location. The local responses are enhanced by evaluating a support function that measures feature compatibility ([Neumann and Mingolla, 2001] for an overview and taxonomy of grouping schemes). The measure of compatibility depends on the lateral integration that utilizes oriented weighting functions for contrast features arranged along a model shape outline, e.g., circular arcs with different radii [Parent and Zucker, 1989]. Such a scheme thus implicitly incorporates curvature as a local contour feature. In order to make this explicit, different contour radii and signs of curvature (for individual orientations) have been considered in [Rodríguez-Sánchez and Tsotsos, 2012]. Rather then implementing this curvature selectivity in a hard-wired scheme of local oriented filter conjunctions, we propose that this selectivity is generated via bottom-up and top-down filter mechanisms organized in a hierarchy. In this architecture the responses from model V2 contour groupings (based on different radii) are integrated by model V4 curvature sensitive cells with coarse bipartite odd-symmetric receptive fields (similar to simple cell profiles, but at much larger spatial scale). The sign of curvature is distinguished by cells of opposite polarity that mutually compete for each orientation. As a consequence responses are generated preferentially in cases where a single dominant curvature is present while responses are suppressed for straight contours which feed curvature cells symmetrically. The curvature radius is represented through a family of differently scaled integration sizes of such model V4 cells. Each of these cells have a specific peak selectivity. In the simulations we used three different sizes for each curvature sign. In order to make those cell responses selective to the feature specificity but mainly invariant to luminance contrast we suggested that each V4 cell response competes against the responses of other curvature selective cells in a local pool that interact via a mechanism of shunting inhibition. This leads to normalization of responses just like in those mechanisms proposed to account for various non-linearities at different stages in cortical processing, e.g., for context related contour responses in V1 [Carandini and Heeger, 1994, Carandini et al., 1999], attention selection [Carandini and Heeger, 2012], and higher level cognitive functions [Louie et al., 2011]. Since the curvature sensitive model V4 cells, in turn, send feedback to their input contour representations in model V2 and filter response in model V1 those corresponding input activations will be enhanced. The amplitude of responses in distributed boundary representations will be amplified as an emergent net effect such that local salient curvature features in a shape outline will be amplified to yield distributed component feature representations of figural shapes.

These local boundary and curvature representations also feed mechanisms of border ownership assignment at the level of the model V4/IT complex. Such mechanisms have been investigated before in [Zhou et al., 2000, O'Herron and von der Heydt, 2011]. Our computational framework belongs to the group of feedback models for border ownership encoding (see the overview of the current state in [Williford and von der Heydt, 2013], see discussion below). We adopted this generic scheme by integrating responses from curvature selective cells with the compatible sign of curvature. In such a way the ownership configuration favors contributions from coarsely presented convex components. If a shape with multiple convex and concave segments is present then the ownership cells with opponent direction selectivities compete in order to arrive at a disambiguated assignment of surface belongingness. This makes the testable prediction that bumpy outlines should lead to slightly longer ownership disambiguation than for smooth convex shapes since the disambiguation will take more time when initially opposite assignment hypotheses coexist.

An additional investigation was argued to be of importance in the work proposed here. Several experimental investigations have reported that cells in extra-striate cortex selectively respond to corner junctions. For example, [Ito and Komatsu, 2004] (compare also [Hegdé and Van Essen, 2000]) reported that cells in area V2 selectively respond as to generate representations of sharp corners, or angles, selective for a particular opening angle. Similarly, [Pasupathy and Connor, 1999, Yau et al., 2013a] show that area V4 cells respond to sharp shape corners with a sub-population of cells preferring sharp corners with different orientation and opening angles while another sub-population prefers smooth rounded corners. While the previous hierarchical models can account for the response selectivity for any of these generic corner types the perceptual representation of sharp localized features that allow, e.g., to distinguish between sharp and rounded corners remain unanswered. Sharp corners of any opening angle would be indistinguishable from the smooth variants of these corners given the increasing smoothing and sub-sampling of the visual representation while proceeding in the hierarchy. Our model argues in favor of a distributed representation: While shape sensitive cells at an intermediate level represent the salient shape protrusions (as in V4) the localized detail of an outline is represented at a higher spatial resolution in lower-level representations, e.g., in V1, V2, V3. In our model we suggest representations of smooth boundaries with different curvatures represented by groupings in model V2 while sharp corners are implicitly represented by convergent V1 input in local representations in model V2/V3. We assume that responses of cells in the model V2/V3 complex mutually compete such that their energy provides a measure to normalize individual responses. These provide convergent input to curvature selective contour cells in model V4 which, in turn, send feedback signals to their input sites at preceding stages. Since they are driven by either smooth or sharp contour arrangements the interaction of bottom-up sensory and top-down context-driven signals leads to selective enhancement of the particular corner configuration in the present stimulus. The specific details of the interaction between such counter-stream signal flows are discussed below.

4.5.3 Feedback as prediction mechanism to link shape components

The hierarchical model architecture proposed here is composed of multiple model areas each of which is represented by a three-stage columnar cascade model. In a nutshell, the model cascade consists of (i) an initial stage of input filtering, (ii) a stage of activity modulation of filter outputs by top-down or lateral reentrant signals, and (iii) a stage of center-surround interaction of target cells against an inhibitory pool of cells leading to activity normalization to generate the net output response of the model area. These three stages can be roughly mapped onto compartments of cortical area subdivisions (as suggested in [Self et al., 2012]). The filtering stage of the driving feedforward input signals is specific to the particular (model) area under consideration. At the output stage, the activity normalization is computed by a mechanism of shunting inhibition, like the non-linear divisive mechanisms proposed in [Carandini and Heeger, 1994, Carandini et al., 1999, Kouh and Poggio, 2008, Carandini and Heeger, 2012]. The feedback signal is generated at higher-level cortical stages or parallel processing pathways and is thought to provide context information that is re-entered at the stage earlier in the processing hierarchy [Grossberg, 1980, Edelman, 1993].

The functional role feedback signals play still remains controversial. Different proposals how feedback signals interact and combine with the driving feed-forward stream have been discussed in the literature which have received different support from the experimental literature [Markov et al., 2013]. One such framework proposes that the goal of computation is to reduce the residual error between the different signal streams in order to approach the sensory prediction generated by higher stages of processing [Ullman, 1995, Bastos et al., 2012]. This idea is rooted in the Bayesian theory of predictor-corrector mechanisms which yields to the Kalman optimal filter realization under some restricting assumptions [Rao and Ballard, 1999]. We follow an alternative route in which the feedback mechanism is modulatory in nature. Unlike predictive coding which tried to drive the difference between driving signals and the prediction to zero bottom-up input signals are amplified by matching feedback signals. This leads to a gain enhancement for those cell responses where a matching top-down predictive signal template has been generated. This feedback signal amplifies the sensory signal such that the subsequent competition between neurons yields a competitive advantage for the enhanced response patterns (biased competition; [Girard and Bullier, 1989, Desimone, 1998, Roelfsema et al., 2002, Reynolds and Heeger, 2009]). The modulation mechanism is reminiscent of the linking mechanism suggested by [Eckhorn et al., 1990, Eckhorn, 1999] to account for activity synchronization in networks of spiking neurons. We have recently demonstrated [Brosch and Neumann, 2014a] that such mechanism of convergent bottom-up feedforward and top-down feedback signal correlation accounts for the signal amplification as measured at the level of cortical pyramidal cells [Larkum, 2013].

In the shape processing architecture described here the modulatory feedback serves the role of a predictor [Spratling, 2008]. For example, bottom-up input in oriented contrast is integrated by mechanisms of contour grouping and integration to generate continuous boundary representations. This is similar in spirit as the recent investigation of [Piëch et al., 2013] who emphasized how context information at higher cortical stages influence more local feature representation at lower levels. Here, the same principle is replicated over different stages of model cortical processing. Contour representations after grouping in model V2 and junction configurations in model V3 send their output activations to curvature sensitive cells in model V4 where the activities are integrated. These cells, in turn, send their feedback to the input populations of neurons that have generated their input. The computational logic is that the curvature responses provide a template of context-related information about the local presence of oriented shape features. The modulatory feedback amplifies those inputs that are consistent with the curvature feature representation. The mutual competition of responses in a pool of cells at the lower level leads to a suppression of inputs that do not contribute to the present curvature feature. In all, a distributed representation of shape information is created that contains coarse-grained configurational information about stimulus shape and, at the same time, the spatially localized detail needed to distinguish between sharp and smooth corners. Similarly, the action of feedback sent from ownership sensitive cells (in the V4/IT complex of the model) to curvature sensitive and grouping cells in model V2 and V4 also provides context information for the assignment of configurational information. Here, the ownership assignment is based on the consolidation of evidence which convex shape elements make to establish a closed shape region in the visual field. This context is delivered via feedback to their input that represents fragments of shape components (irrespective of the sign of curvature) and also to the grouping representations. Those shapes that finally receive assigned direction of border ownership, and thus figure-ground direction, will enhance the associated inputs at the intermediate level orientation-curvature representations.

In all, the hierarchical processing scheme proposed here relies on extensive bidirectional flow of information in which the feedback signals that represent context-sensitive templates are gated by feed-forward driving input signals. Such a modulating feedback driven gain control mechanism relates to mechanisms proposed by Roelfsema and colleagues [Lamme and Roelfsema, 2000, Roelfsema et al., 2002, Roelfsema, 2006] in which spatial detail is generated by feature-driven low-level processes and representations and subsequently associated with coarse-grained context information provided by intermediate and higher-levels of cortical computation. The mechanisms implemented in the proposed model are consistent with theoretical predictions from computational constraints visual perception imposes on the underlying architecture [Tsotsos, 1988]. The advantages in computational complexity have been calculated for principles such as hierarchical organization, localized receptive field computations, and dedicated (distributed) maps of feature representation and their combination. Feedback has been suggested to steer an attentional beam by selecting a spatial region and their computational resources [Tsotsos, 2005]. In the proposed architecture feedback also selectively enhances representations of features by increasing their gain which are coherent with the predictions generated at higher-level stages with more condensed coding of shape and figural properties. Also we emphasize that this provides a key to enhance (and make accessible) localized shape features, such as sharp edges, as part of a shape configuration that is represented on a coarser scale.

4.5.4 Model limitations and further extensions

The proposed model architecture emphasized the computational role of feed-forward and feedback mechanisms in order to generate a hierarchical distributed representation of shape information. For that reason, we focused on the representational aspects as steady-state solutions of an otherwise dynamic interaction between neuronal populations and representations distributed over several model areas. We did not, so far, investigate the temporal response phases observed for shape sensitive cells in V4 [Yau et al., 2013a]. The work of Roelfsema and colleagues has shown that different response phases exist that can be reliably assigned to different mechanisms in processing, namely for feature detection, figure-ground segregation, and attention [Roelfsema et al., 2007]. We have demonstrated that such separate but temporally overlapping phases can be accounted for by a recurrent network of mutually interacting neuronal sites. The network model has been composed of the same components like the present model architecture [Raudies and Neumann, 2010]. It would thus be interesting to reveal whether similar temporal phases can be identified for model V4 cells that may give rise to identify different signatures indicative of contributions from delayed neuronal mechanisms that are involved in the computation of figural shape information.

Different signal streams (particularly in the feed-forward sweep of feature processing) operate on different temporal scales. Several lines of evidence suggest that the dorsal and the ventral streams of processing do not operate entirely in isolation but mutually interact at different levels [Felleman and Van Essen, 1991, Markov et al., 2013]. Also different response characteristics of cells may define different temporal routes of fast and slow processing [Born, 2001] that may help fusing information from different pathways. Here, we did not take into account such interactions based on different temporal effectivenesses. However, other model investigations capitalized on combining information from different channels to improve the selectivity of representation. For example, edge detection and grouping (in the ventral pathway) could be enhanced through mutually inhibitory gain control (which is similar as the normalization stage described here) generated by representations in the dorsal pathway. Since the dorsal representation is created by magno-cellular responses, such inhibition arrives already early to shape the selectivity of shape representations in the ventral path that is mainly driven by parvo-cellular responses [Shi et al., 2013]. Similarly, interactions between the motion and form pathway have been suggested to help disambiguating localized features that give rise to occlusion cues which, in turn, support the disambiguation of object representation in the motion representation [Bayerl and Neumann, 2007, Beck and Neumann, 2010]. Such detailed mechanisms would further enhance the proposed model architecture in refining the selectivities at different levels of low and intermediate representation.

As already pointed out above, the focus here is on the processing of 2D shape representations. In [Cadieu et al., 2007] the authors have highlighted that their specific model investigation on shape representation in V4 is part of a larger hierarchically organized architecture for object recognition [Riesenhuber and Poggio, 1999, Serre and Poggio, 2010]. Since their model principles relied on purely feedforward processing the insights provided in the work presented here might also shed some light on the mutual interactions between different processes on an even larger scale of object recognition processes. In addition, it would be interesting to find out how the representation of 3D surface patches [Yamane et al., 2008] seamlessly fit into a model computational architecture of recurrent shape computation.

In the presented coverage our model does not respond to contours elicited by contrasts of spatial luminance statistics caused by differently textured regions. However, the core mechanisms, including initial filtering, modulatory feedback and competitive interaction for normalization, are like those proposed in the current contribution. A model that focuses on the processing of such boundaries has been developed in [Thielscher and Neumann, 2003]. It is thus very likely that the recent model architecture proposed here can be extended with processing stages capable to process texture define boundaries as well without changing the basic architecture and computational principles. Also not considered in the current version is a multi-scale approach. We acknowledge the theoretical justification of hierarchical multi-stage processing to build up a pyramid-like structure [Tsotsos, 2005]. Incorporating this representational diversity would allow the processing of a wider range of curvature configurations in shape outlines. In addition, this would support a more robust segregation of border ownership on the basis of convexities in the figural outline. We have focused our efforts on the specification of a hierarchically organized network architecture that utilized bottom-up and top-down convergent processing flows. In order to keep the computational efforts and the simulation times within reasonable bounds we restricted our description to single scale components at the different model stages within the hierarchy. A more extended realization of components is certainly desired but left for future investigations.

Intermediate level representations involve cells with receptive fields that recruit multiple sub-field components [Yau et al., 2013a, Mineault et al., 2012]. The model of [Cadieu et al., 2007] accounts for this by sequentially fitting the subunits of intermediate level receptive field models to match the response profiles of V4 responses measured experimentally. This yields a sampling structure of statistically significant inputs in a feature space that contributes a significant amount of feature input to generate the final response of a shape selective cell. So far, in our modeling we sampled the spatial and the feature domains regularly. This of course demands high representational as well as computational resources. Consequently, it would be of interest to see how an irregularly sampled 4D space-feature domain (with orientation and curvature features) can be embedded into the scheme of shape representation proposed here.

Publications

The contributions contained in this chapter are the accumulated outcome of the preceding research period. Prior investigations and earlier results have been presented in individual publications. The applicability of neurally inspired mechanisms to actual problems of computer vision and machine learning was presented in [Layher et al., 2011] and [Schels et al., 2013].

The contributions of optical flow for the perception of figure and ground has been investigated over a period of time and lead to various publications [Tschechne and Neumann, 2011a, Tschechne and Neumann, 2011b, Tschechne and Neumann, 2012]. In this thesis, the contribution of motion information to the resolving of spatial arrangements was neglected, however the component of contextual modulation using area IT for the resolving of border ownership assignments from shape outline was included. The model presented here has been published in [Tschechne and Neumann, 2014a]. Some depictions in this thesis have been adapted from this publication.

5 Conclusion of the thesis

One of the wonders of evolution is the brain, a fascinating matter that hosts a conglomerate of neurons and synapses. The complexity, the density of neurons and the level of interconnectedness provides us and other species with highly developed cognitive and intellectual capabilities, that still outplay artificial intelligence systems in terms of performance, flexibility, robustness and energy consumption. How does this grayish matter work, where are the gear-wheels and transmission lines of perception, knowledge, emotion and instincts? Researchers deal with these questions for many decades now, but many challenges remain on the way to a complete understanding.

5.1 Contextual modulation as generic principle

One established insight into the principles of visual processing is that it takes place along the visual pathway and follows mainly two tracks, one 'what' pathway that is dedicated to form processing and one 'where' pathway that is dedicated to motion processing. Along these tracks, neural processing areas receive input from increasingly large receptive fields, which are the regions of the visual field they receive input from. At first sight, early areas with small receptive fields seem to only respond to their local input. Interestingly, this is not the case. The surround of the receptive fields affects the response of cells, and this happens in different parts of visual cortex [Nurminen and Angelucci, 2014]. The authors of [Krause and Pack, 2014] review the contribution of contextual modulation with many examples taken from different parts of the visual system. In a nutshell, contextual modulation influences local representations with information from outside their receptive fields. This hypotheses is backed up with results from anatomical and physiological investigations that revealed that there is a considerable amount of downward-directed connections in the cortex. However, these downward directed connections and thus, contextual modulation, is often ignored in models for the sake of simplicity, but it is believed that they contribute significantly to a list of nonlinear effects that can be observed in visual processing.

The hierarchical model architecture proposed here is composed of multiple model areas each of which is represented by a three-stage columnar cascade model. In a nutshell, the model cascade consists of (i) an initial stage of input filtering, (ii) a stage of activity modulation of filter outputs by top-down or lateral reentrant signals, and (iii) a stage of center-surround interaction of target cells against an inhibitory pool of cells leading to activity normalization to generate the net output response of the model area. These three stages can be roughly mapped onto compartments of cortical area subdivisions [Self et al., 2012], see Sec. 2.3. The filtering stage performs linear or nonlinear integration of afferent input connections. The modulatory feedback amplifies those inputs that are consistent with the feature representation. Those enhanced features have a competitive advantage against others that did not receive top-down modulation. Such a modulating feedback driven gain control mechanism relates to mechanisms proposed by Roelfsema and colleagues. Feedback is discussed controversially in literature [Markov et al., 2013] and can also be regarded as predictor-corrector mechanisms in a Bayesian framework [Rao and Ballard, 1999, Ullman, 1995, Bastos et al., 2012]. In this thesis, we follow a path where feedback is modulatory. The following mutual competition of responses in a pool of cells at the lower level leads to a suppression of inputs that do not contribute to the present feature. Chap. 2 introduces the reader to the computational foundations of these mechanisms. These computational building blocks for models of visual processing have previously been used in architectures for form and motion processing. [Brosch and Neumann, 2014a] recently investigated that the pool normalization accounts for signal amplification at the level of cortical pyramidal cells.

This concept is applied to two very different models of visual processing, but each of them makes use of the same generic principles. Our models rely on extensive bidirectional flow of information in which the feedback signals that represent context-sensitive templates are gated by feed-forward driving input signals. By this we show of contextual information can be made available even across different modalities using the same mechanisms, and how the quality of the representations benefits from this influence.

5.2 Event-based motion estimation

We present a model of motion estimation based on asynchronous input from a neuromorphic vision sensor in Chap. 3. We suggest that the receptive field structure of spatio-temporally motion selective V1 neurons resemble the likelihood distribution in the $x-y-t$-domain. Our model is directly motivated by recent physiological findings described in [De Valois et al., 2000]. These findings suggest that simple cells in V1 can be subdivided into two groups, namely those that are separable in space-time (which prefer stationary stimuli), and those that are inseparable in space-time (which are sensitive to moving stimuli). A detailed singular-value decomposition analysis of the spatio-temporal response properties of such cells revealed that direction-selective cells with spatio-temporally inseparable receptive fields are composed by two singular values of one temporally biphasic and one temporally monophasic cell [De Valois and Cottaris, 1998]. We propose a simple model to generate the receptive profiles and their linear superposition. We suggest a simple parameterized model that is capable to generate a bank of direction sensitive filters in multiple scales. Our investigation is similar to the investigation of [Escobar and Kornprobst, 2012] who seek to specify the spatio-temporal selectivity of direction-selective filters to be employed. Our mechanism is directly derived from physiological findings and we were able to derive simple mechanisms for parameterization. Perhaps the most similar scheme in comparison to our model is the one proposed by [Adelson and Bergen, 1985] which also suggests to derive spatio-temporally selective kernels by superposing different receptive fields. We rely on the superposition of space-time separable filters with out-of-phase temporal modulation filter-responses. Our test applications of the model implementation successfully demonstrates the functionality of such initial filtering for motion detection from the spatio-temporal event cloud. Our model also contains a processing stage that creates a representation of motion streaks. Those streaks or speed lines occur from the temporal integration of visual features and result in an activation of cells tuned to form features. We show how the event-based format directly leads to an activation of form cells using existing mechanisms of temporal and spatial integration. We show how the representation of these speed lines may influence a representation of motion signals in another example of form-motion-interaction.

5.3 Hierarchical representation of form features

The generic cortical mechanisms are also applied throughout a model of hierarchical form processing that is presented in Chap. 4. Here we propose a model of mechanisms in ventral pathway that represents visual scenes as image regions and allows their combination of surfaces and prototypical objects. Multiple mutually connected areas in the ventral cortical pathway receive visual input and extract local form features that are subsequently grouped into increasingly complex, more meaningful image elements. We propose a mechanism how such a distributed network of processing is capable of representing highly articulated changes in shape boundary as well as subtle curvature changes. We propose a recurrent computational network architecture that utilizes hierarchical distributed representations of shape features to encode surface and object boundaries over different scales of resolution. Our model makes use of neural mechanisms that model the processing capabilities of early and intermediate stages in visual cortex, namely areas V1-V4 and IT. We suggest that multiple specialized component representations interact by feedforward hierarchical processing that is combined with feedback signals driven by representations generated at higher stages. Based on this, global configurational as well as local information is made available to distinguish changes in the object's contour. Once the outline of a shape has been established, contextual contour configurations are used to assign border ownership directions and thus achieve segregation of figure and ground. The model, thus, proposes how separate mechanisms contribute to distributed hierarchical cortical shape representation and combine with processes of figure-ground segregation. Our model is probed with a selection of stimuli to illustrate processing results at different processing stages. We especially highlight how modulatory feedback connections contribute to the processing of visual input at various stages in the processing hierarchy.

The two models showed how contextual information is contributed to early areas using a hierarchical model of generic building blocks. Contextual information in the shape of modulatory feedback is elegantly contained in a physiologically plausible three-stage processing cascade that has versatile use for the processing of visual features. Such models of visual processing contribute to a precise understanding of neural principles, the building block they are comprised of and possible processing strategies in vision and perception that lead to our detailed and accurate visual capabilities. In addition, the investigation of such neural principles and their applicability to computer systems has a huge potential for upcoming generations of signal-processing hard- and software. Current and future requirements to computer systems are at the moment tough to realize with traditional methods of computer science. The investigation of biologically inspired mechanisms provided alternative strategies to achieve robust solutions to complex problems with the application of mechanisms of neural information processing.

Acknowledgments

I want to thank everybody wo contributed in one way or another to the making of this thesis. Thanks to my supervisor Heiko Neumann, who is patient and relaxed and supports his students' individual skills with time for discussion and very elaborate corrections and contributions of paper manuscripts. Thanks to my colleagues and all staff members of the institute who supported me many times with answers to smaller or bigger problems and had an open ear for a broad range of related and unrelated discussions about the meaning of doing a doctor's degree, life, and everything. Some of the colleagues became friends over time and I hope that there will be many chances to get together long after the day I write these lines. Thanks to Y.T. for motivating me and for bearing with me through times of stress, high and low moods. Last but not least I send my gratitude to T.F. for the support and encouragement to start this endeavor. Many thanks to my family for their support being there and offering me retreat and diversion whenever I was feeling for it.

Bibliography

[Abdul-Kreem and Neumann, 2014] Abdul-Kreem, L. I. and Neumann, H. (2014). Bio–Inspired Model For Motion Estimation Using An Address-Event Representation (accepted). In *10th Int. Conf. on Computer Vision Theory and Application (VISAPP) 2015*.

[Adams, 2010] Adams, A. (2010). Canon 5d: How much dynamic range does it have, really? `http://provideocoalition.com/aadams/story/canon_5d_how_much_dynamic_range_does_it_have_really/`. Accessed February 12, 2015.

[Adelson and Bergen, 1985] Adelson, E. H. and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am.*, 2(2):284–299.

[Albright, 1984] Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area MT. *J Neurophys*, 52:1106–1130.

[Anzai et al., 2007] Anzai, A., Peng, X., and Van Essen, D. (2007). Neurons in monkey visual area V2 encode combinations of orientations. *Nature Neuroscience*, 10:1313–1321.

[Apthorp et al., 2010] Apthorp, D., Cass, J., and Alais, D. (2010). Orientation tuning of contrast masking caused by motion streaks. *Journal of Vision*, 10(10):1–13.

[Apthorp et al., 2013] Apthorp, D., Schwarzkopf, D. S., Kaul, C., Bahrami, B., Alais, D., and Rees, G. (2013). Direct Evidence for encoding of motion streaks in human visual cortex. *Proc R Soc B*, 400(20122339).

[Apthorp et al., 2009] Apthorp, D., Wenderoth, P., and Alais, D. (2009). Motion streaks in fast motion rivalry cause orientation-selective suppression. *Journal of Vision*, 9(5):1–14.

[Barlow, 1958] Barlow, H. (1958). Temporal and spatial summation in human vision at different background intensities. *Journal of Physiology*, 141:337–350.

[Barlow and Levick, 1965] Barlow, H. B. and Levick, W. R. (1965). The Mechanism of Directionally Selective Units in Rabbits Retina. *The Journal of Physiology*, 178(3):477–504.

[Barranco et al., 2014] Barranco, F., Fermüller, C., and Aloimonos, Y. (2014). Contour Motion Estimation for Asynchronous Event–Driven Cameras. *Proceedings of the IEEE*, 102(10):1537–1556.

[Barron et al., 1994] Barron, J., Beauchemin, S., and DJ, F. (1994). Performance of optical flow techniques. *Int. J. of Computer Vision*, 12(1):43–77.

[Bastos et al., 2012] Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron*, 76(4):695–711.

[Bayerl and Neumann, 2004] Bayerl, P. and Neumann, H. (2004). Disambiguating visual motion through contextual feedback modulation. *Neural Computation*, 16(10):2041–66.

[Bayerl and Neumann, 2007] Bayerl, P. and Neumann, H. (2007). Disambiguating visual motion by form-motion interaction – a computational model. *Int. Journal of Computer Vision*, 72(1):27–45.

[Baylis and Driver, 2001] Baylis, G. C. and Driver, J. (2001). Shape coding in IT cells generalizes over contrast and mirror reversal but not figure–ground reversal. *Nature Neuroscience*, 4:937–942.

[Beck and Neumann, 2010] Beck, C. and Neumann, H. (2010). Interactions of motion and form in visual cortex -– a neural model. *Journal of Physiology Paris*, 104:61–70.

[Beck and Neumann, 2011] Beck, C. and Neumann, H. (2011). Combining feature selection and integration. a neural model for mt motion selectivity. *PLoS ONE*, 6(7).

[Bedell et al., 2010] Bedell, H. E., Tong, J., and Aydin, M. (2010). The pereception of motion smear during eye and head movements. *Vision Research*, 50:2692–2701.

[Ben-Shahar and Zucker, 2004] Ben-Shahar, O. and Zucker, S. (2004). Geometrical Computations Explain Projection Patterns of Long–Range Horizontal Connections in Visual Cortex. *Neural Computation*, 16(3):445–76.

[Benosman et al., 2014] Benosman, R., Clercq, C., Lagorce, X., Ieng, S. H., and Bartolozzi, C. (2014). Event-Based visual flow. *IEEE Trans. on Neural Networks and Learning Systems*, 25(2):407–417.

[Benosman et al., 2011] Benosman, R., Ien, S.-H., Rogister, P., and Posch, C. (2011). Asynchronous Event–Based Hebbian Epipolar Geometry. *IEEE Transactions on Neural Networks*, 22(11):1723–1734.

[Benosman et al., 2012] Benosman, R., Ieng, S.-H., Clercq, C., Bartolozzi, C., and Srinivasan, M. (2012). Asynchronous frameless event–based optical flow. *Neural Networks*, 27:32–37.

[Born and Bradley, 2005] Born, R. and Bradley, D. (2005). Structure and function of visual area MT. *Ann. Rev. Neurosci.*, 28:157–189.

[Born, 2001] Born, R. T. (2001). Visual processing: Parallel–er and Parallel–er. *Current Biology*, 11:R566—-R568.

[Born et al., 2010] Born, R. T., Tsui, J. M. G., and Pack, C. C. (2010). Temporal dynamics of motion integration. In Ilg, U. J. and Masson, G. S., editors, *Dynamics of Visual Motion Processing: Neuronal, Behavioral, and Computational Approaches*, chapter 2. Springer Berlin.

[Borst and Euler, 2011] Borst, A. and Euler, T. (2011). Seeing Things in Motion: Models, Circuits, and Mechanisms. *Neuron*, 71(6):974–994.

[Bosking et al., 1997] 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(6):2112–27.

[Brandli et al., 2014] Brandli, C., Berner, R., Yang, M., Liu, S. C., and Delbruck, T. (2014). A 240×180 130dB 3μs Latency Global Shutter Spatiotemporal Vision Sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341.

[Brosch and Neumann, 2014a] Brosch, T. and Neumann, H. (2014a). Computing with a canonical neural circuits model with pool normalization and modulating feedback. *Neural Computation.*

[Brosch and Neumann, 2014b] Brosch, T. and Neumann, H. (2014b). Interaction of feedforward and feedback streams in visual cortex in a firing-rate model of columnar computations. *Neural Networks*, 54:11–6.

[Brosch et al., 2014] Brosch, T., Tschechne, S., and Neumann, H. (2014). On event–based motion estimation. *Frontiers in Neuroscience.*

[Burr, 2000] Burr, D. C. (2000). Motion vision: are 'speed lines' used in human visual motion? *Curr Biol.*, 10:R440–443.

[Burr and Morgan, 1997] Burr, D. C. and Morgan, M. J. (1997). Motion deblurring in human vision. *Proceedings of the Royal Society of London*, 264:431–436.

[Burr and Ross, 2002] Burr, D. C. and Ross, J. (2002). Direct evidence that 'Speedlines' influence motion mechanisms. *Journal of Neuroscience*, 22(19):8661–8664.

[Cadieu et al., 2007] Cadieu, C., Kouh, M., Pasupathy, A., Connor, C., Riesenhuber, M., and Poggio, T. (2007). A model of V4 shape selectivity and invariance. *J. Neurophysiology*, 98(1733–1750).

[Camunas-Mesa et al., 2014] Camunas-Mesa, L. A., Serrano-Gotarredona, T., Ieng, S. H., Benosman, R., and Linares-Barronco, B. (2014). On the use of orientation filters for 3D reconstruction in event-driven stereo vision. *Frontiers in Neuroscience*, 8(48).

[Caplovitz and Tse, 2007] Caplovitz, G. and Tse, P. (2007). V3A processes contour curvature as a trackable feature for the perception of rotational motion. *Cerebral Cortex*, 17:1179–1189.

[Carandini and Heeger, 1994] Carandini, M. and Heeger, D. J. (1994). Summation and Division by Neurons in Primate Visual Cortex. *Science*, 264(5163):1333–6.

[Carandini and Heeger, 2012] Carandini, M. and Heeger, D. J. (2012). Normalization as a Canonical Neural Computation. *Nature Reviews Neuroscience*, 13:51–62.

[Carandini et al., 1997] Carandini, M., Heeger, D. J., and Movshon, J. A. (1997). Linearity and Normalization in Simple Cells of the Macaque Primary Visual Cortex. *Journal of Neuroscience*, 17(21):8621–8644.

[Carandini et al., 1999] Carandini, M., Heeger, D. J., and Movshon, J. A. (1999). Linearity and Gain Control in V1 Simple Cells. In Ulinski, P. S., Jones, E. G., and Peters, A., editors, *Cerebral Cortex Volume 13 Models of Cortical Circuits*, pages 401–443. Kluwer Academic / Plenum Publishers.

[Cerda and Girau, 2008] Cerda, M. and Girau, B. (2008). A Neural Model with Feedback for Robust Disambiguation of Motion. *European Symposium on Artificial Neural Networks*, 567:505–510.

[Churchland et al., 1990] Churchland, P., Koch, C., and Sejnowsky, T. (1990). *What is Computational Neuroscience?*, chapter 5, pages 46–55. MIT Press.

[Clady et al., 2014] Clady, X., Clercq, C., Ieng, S.-H., Houseini, F., Randazzo, M., Natale, L., Bartolozzi, C., and Benosman, R. (2014). Asynchronous visual event–based time–to–contact. *Frontiers in Neuroscience*, 8(9):1–10.

[Connor et al., 2007] Connor, C. E., Brincat, S. L., and Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. *Current Opinion in Neurobiology*, 17:140–147.

[Conradt et al., 2009] Conradt, J., Berner, R., Cook, M., and Delbruck, T. (2009). An Embedded AER Dynamic Vision Sensor for Low–Latency Pole Balancing. *5th IEEE Workshop on Embedded Computer Vision (in conjunction with ICCV 2009)*, pages 1–6.

[De Pasquale and Murray Sherman, 2013] De Pasquale, R. and Murray Sherman, S. (2013). A modulatory effect of the feedback from higher visual areas to V1 in the mouse. *J Neurophysiology*, 109:2618–2631.

[De Valois and Cottaris, 1998] De Valois, R. and Cottaris, N. P. (1998). Inputs to Directionally Selective Simple Cells in Macaque Striate Cortex. *PNAS*, 95:14488–93.

[De Valois et al., 2000] De Valois, R., Cottaris, N. P., Mahon, L. E., Elfar, S. D., and Wilson, J. A. (2000). Spatial and temporal receptive fields of geniculate and cortical cells and directional selectivity. *Vision Research*, 40:3685–3702.

[DeAngelis et al., 1995] DeAngelis, G. C., Ohzawa, I., and Freeman, R. D. (1995). Receptive-field dynamics in the central visual pathways. *TINS*, 18(10):451–458.

[Delbrück, 2012] Delbrück, T. (2012). Fun with asynchronous vision sensors and processing. In *Fusellio et al. (Eds.) ECCV 2012 Ws/Demos, Part 1, LNCS*, volume 1, pages 506–515.

[Delbrück and Lang, 2013] Delbrück, T. and Lang, M. (2013). Robotic goalie with 3ms reaction time at 4% CPU load using event-based dynamic vision sensor. *Frontiers in Neuroscience*, 7(223):1–7.

[Desimone, 1998] Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philos Trans R Soc Lond B*, 353(1373):1245–55.

[DeValois et al., 1982a] DeValois, R. L., Albrecht, D. G., and Thorell, L. G. (1982a). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, 22:545–559.

[DeValois et al., 1982b] DeValois, R. L., Yund, E. W., and Hepler, N. (1982b). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, 22:531–544.

[Drazen et al., 2011] Drazen, D., Lichtsteiner, P., Häflinger, P., Delbrück, T., and Jensen, A. (2011). Toward real-time particle tracking using an event-based dynamic vision sensor. *Exp. Fluids*, 51:1465–1469.

[Eckhorn, 1999] Eckhorn, R. (1999). Neural mechanisms of visual feature binding investigated with microelectrodes and models. *Visual Cognition*, 6(3):231–65.

[Eckhorn et al., 1990] 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.

[Edelman, 1993] Edelman, G. (1993). Neural Darwinism: Selection and reentrant signaling in higher brain function. *Neuron*, 10:115–125.

[Escobar and Kornprobst, 2012] Escobar, M. J. and Kornprobst, P. (2012). Action recognition via bio-inspired features: The richness of center-surround interaction. *Computer Vision and Image Understanding*, 116(5):593–605.

[Felleman and Van Essen, 1987] Felleman, D. J. and Van Essen, D. C. (1987). Receptive field properties of neurons in area V3 of macaque monkey extrastriate cortex. *J. of Neurophysiol.*, 57(4):889–920.

[Felleman and Van Essen, 1991] Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47.

[Fermüller and Aloimonos, 1995] Fermüller, C. and Aloimonos, Y. (1995). Qualitative egomotion. *Int. J. of Computer Vision*, 15(1–2):7–29.

[Field et al., 1993] Field, D. J., Hayes, A., and Hass, R. F. (1993). Contour integration by the human visual system: evidence for a local association field. *Vision Research*, 33(2):173–193.

[Fitzhugh, 1969] Fitzhugh, R. (1969). *Mathematical models of excitation and propagation in nerve*, pages 1—85. McGraw-Hill Book Co.

[Fowlkes et al., 2007] Fowlkes, C. C., Martin, D. R., and Malik, J. (2007). Local figure-ground cues are valid for natural images. *Journal of Vision*, 7(8):1–9.

[Fu et al., 2008] Fu, Z., Delbrück, T., Lichtsteiner, P., and Culurciello, E. (2008). An address-event fall detector for assisted living applications. *IEEE Trans. on Biomedical Circuits and Systems*, 2(2):88–96.

[Fukushima, 1980] 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.

[Fukushima, 1988] Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1:119–130.

[Geisler, 1999] Geisler, W. S. (1999). Motion streaks provide a spatial code for motion direction. *Nature*, 400:65–69.

[Gibson et al., 2014] Gibson, T. A., Heath, S., Quinn, R. P., Lee, A. H., Arnold, J. T., Sonti, T. S., Whalley, A., Shannon, G. P., Song, B. T., Henderson, J. A., and Wiles, J. (2014). Event–Based Visual Data Sets for Prediction Tasks in Spiking Neural Networks. In *ICANN 2014 LNCS 8681*, pages 635–642. Springer Int. Pub. Switzerland.

[Gilbert and Li, 2013] Gilbert, C. D. and Li, W. (2013). Top–Down Influences on Visual Processing. *Nature Reviews Neuroscience*, 14:350–63.

[Girard and Bullier, 1989] Girard, P. and Bullier, J. (1989). Visual activity in area V2 during reversible inactivation of area 17 in the macaque monkey. *Journal of Neurophysiology*, 62(6):1287–302.

[Grossberg, 1980] Grossberg, S. (1980). How does the brain build a cognitive code? *Psychological Review*, 87(1):1–51.

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

[Grossberg and Mingolla, 1985] Grossberg, S. and Mingolla, E. (1985). Neural dynamics of perceptual grouping: textual, boundaries, and emergent segmentations. *Perception and Psychophysics*, 38(2):141–171.

[Hansen and Neumann, 2004] Hansen, T. and Neumann, H. (2004). Neural Mechanisms for the Robust Representation of Junctions. *Neural Computation*, 16(5):1013–37.

[Hassenstein and Reichard, 1956] Hassenstein, B. and Reichard, W. (1956). Functional structure of a mechanism of perception of optical movement. *Proc. 1st Intl. Congress on Cybernetics*, pages 797–801.

[Hatori and Sakai, 2012] Hatori, Y. and Sakai, K. (2012). Surface–based construction of curvature selectivity from the integration of local orientations. In *19th International Conference on Neural Information Processing (ICONIP)*, pages 425–434.

[Heeger, 1992] Heeger, D. J. (1992). Normalization of Cell Responses in Cat Striate Cortex. *Visual Neuroscience*, 9(2):185–203.

[Hegdé and Van Essen, 2000] Hegdé, J. and Van Essen, D. (2000). Selectivity for complex shapes in primate visual area V2. *J. Neuroscience*, 20(RC61):1–6.

[Heitger et al., 1998] Heitger, F., von der Heydt, R., Peterhans, E., Rosenthaler, L., and Kübler, O. (1998). Simulation of neural contour mechanisms: Representing anomalous contours. *Image and Vision Computing*, 16:407–421.

[Herz et al., 2006] Herz, A., Gollisch, T., Machens, C., and Jaeger, D. (2006). Modeling single-neuron dynamics and computations: a balance of detail and abstraction. *Science*, 314(5796):80–85.

[Hodgkin and Huxley, 1990] Hodgkin, A. L. and Huxley, A. F. (1990). A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve. *Journal of Physiology*, 117:500—544.

[Horn and Schunck, 1981] Horn, B. K. P. and Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, 17:185–203.

[Hubel and Wiesel, 1959] Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591.

[Hubel and Wiesel, 1962] Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):106–54.

[Hupé et al., 1998] 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*, 394:784–7.

[Inilabs, 2014] Inilabs (2014). Inilabs pushbot. `http://www.inilabs.com/products/robots/pushbot`. Accessed January 15, 2015.

[Ito and Komatsu, 2004] 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, 2007] Izhikevich, E. M. (2007). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. MIT Press.

[Jehee et al., 2006] Jehee, J. F. M., Lamme, V. A. F., and Roelfsema, P. R. (2006). Boundary Assignment in a Recurrent Network Architecture. *Vision Research*, 47:1153–1165.

[Kanizsa, 1955] Kanizsa, G. (1955). Margini quasi-percettivi in campi con stimolozione omogenea. *Rivista di Psicologia*, 49(1):7–30.

[Koch, 1998] Koch, C. (1998). *Biophysics of Computation: Information Processing in Single Neurons*. Oxford Univ. Press.

[Koo, 2010] Koo, R. (2010). What is the dynamic range of a canon 5d mark ii? `http://nofilmschool.com/2010/09/what-is-the-dynamic-range-of-a-canon-5d-mark-ii`. Accessed February 12, 2015.

[Kouh and Poggio, 2008] Kouh, M. and Poggio, T. (2008). A Canonical Neural Circuit for Cortical Nonlinear Operations. *Neural Computation*, 20(6):1427–51.

[Krause and Pack, 2014] Krause, M. R. and Pack, C. C. (2014). Contextual modulation and stimulus selectivity in extrastriate cortex. *Vision Research*, 104:36–46.

[Kurzweil, 2012] Kurzweil, R. (2012). *How to Create A Mind: The Secret of Human Thought Revealed*. Viking Penguin.

[Lamme and Roelfsema, 2000] Lamme, V. and Roelfsema, P. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579.

[Larkum, 2013] Larkum, M. (2013). A cellular mechanism for cortical associations: An organizing principle for the cerebral cortex. *Trends in Neurosciences*, 36(3):141–51.

[Layher et al., 2014] Layher, G., Giese, M., and Neumann, H. (2014). Learning representations of animated motion sequences - a neural model. *Topics in Cognitive Science*, 6(1):170–82.

[Layher et al., 2011] Layher, G., Tschechne, S., Scherer, S., Brosch, T., Curio, C., and Neumann, H. (2011). Social signal processing in companion systems-challenges ahead. In *GI-Edition LNI*, page 239. Gesellschaft für Informatik.

[Layton et al., 2012] Layton, O. W., Mingolla, E., and Yazdanbakhsh, A. (2012). Dynamic Coding of Border–Ownership in Visual Cortex. *Journal of Vision*, 12(13):1–21.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient–based learning applied to document recognition. *Proc. of IEEE*, 86(11):2278–2324.

[Lee et al., 2012] Lee, J., Delbruck, T., Park, P., Pfeiffer, M., Shin, C., Ryu, H., and Kang, B. (2012). Gesture Based remote control using stereo pair of dynamic vision sensors. In *IEEE Intl. Symp. on Circuits and Systems (ISCAS 2012).*, pages 741–745, Seoul.

[Lichtsteiner et al., 2008] Lichtsteiner, P., Posch, C., and Delbrück, T. (2008). A 128×128 120dB 15μs Latency Asychronous Temporal Contrast Vision Sensor. *IEEE Journal of Solid–State Circuits*, 43(2):566–576.

[Loeffler and Orbach, 1999] Loeffler, G. and Orbach, H. S. (1999). Computing feature motion without feature detectors: A model for terminator motion without end-stopped cells. *Vision Research*, 39:859–871.

[Logothetis et al., 1995] Logothetis, N., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563.

[Louie et al., 2011] Louie, K., Grattan, L. E., and Glimcher, P. W. (2011). Reward Value–Based Gain Control: Divisive Normalization in Parietal Cortex. *Journal of Neuroscience*, 31(29):10627–39.

[Lucas and Kanade, 1981] Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679.

[Mahowald, 1994] Mahowald, M. (1994). An Analog VLSI System for Stereoscopic Vision. In *The Springer International Series in Engineering and Computer Science*, volume 265. Springer.

[Markov et al., 2013] Markov, N., Ercsey-Ravasz, M., Van Essen, D., Knoblauch, K., Toroczkai, Z., and Kennedy, H. (2013). Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *Science*, 342(1238406).

[Martin and Marshall, 1993] Martin, K. and Marshall, J. (1993). *Unsmearing Visual Motion: Development of Long–Range Horizontal Intrinsic Connections*, pages 417–424. Morgan Kaufmann Publishers, New York, NY, USA.

[McClurkin et al., 1994] McClurkin, J., Optican, L. M., , and Richmond, B. J. (1994). Cortical Feedback increases Visual Information transmitted by Monkey parvocellular lateral geniculate Nucleus Neurons. *Visual Neuroscience*, 640(11):601–617.

[Mineault et al., 2012] Mineault, P., Khawaja, F., Butts, D., and Pack, C. (2012). Hierarchical processing of complex motion along the primate dorsal visual pathway. *Proc. Nat'l. Acad. of Sciences USA, publ. online*, 109(E972–E980).

[Mishkin et al., 1983] Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: Two central pathways. *Trends in Neuroscience*, 6:414–417.

[Mokhtarian, 1995] Mokhtarian, F. (1995). Silhouette–based isolated object recognition through curvature scale space. *IEEE Trans. on PAMI*, 17(5):539–544.

[Mokhtarian and Mackworth, 1986] Mokhtarian, F. and Mackworth, A. (1986). Scale–based description and recognition of planar curves and two–dimensional shapes. *IEEE Trans. on PAMI*, 8(1):34–43.

[Mutch and Lowe, 2008] Mutch, J. and Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *Int. J. of Computer Vision*, 80(1):45–57.

[Nagumo et al., 1962] Nagumo, J., Arimoto, S., and Yoshizawa, S. (1962). An active pulse transmission line simulating nerve axon. In *Proc IRE.*, volume 50, pages 2061–70.

[Nakayama and Silverman, 1988] Nakayama, K. and Silverman, G. H. (1988). The Aperture Problem I. Perception of Nonrigidity and Motion Direction in Translating Sinusoidal Lines. *Vision Research*, 26(6):739–746.

[Neumann and Mingolla, 2001] Neumann, H. and Mingolla, E. (2001). Computational neural models of spatial integration in perceptual grouping. In Shipley, T. and Kellman, P., editors, *From Fragments to Objects: Grouping and Segmentation in Vision*, chapter 12, pages 353–400. Elsevier, Amsterdam.

[Neumann and Ottenberg, 1992] Neumann, H. and Ottenberg, K. (1992). *EUSIPCO-92: Theories and Applications*, volume I, chapter Estimating ramp-edge attributes from scale-space, pages 603–607. Elsevier.

[Neumann and Sepp, 1999] Neumann, H. and Sepp, W. (1999). Recurrent v1-v2 interaction in early visual boundary processing. *Biological Cybernetics*, 81:425–44.

[Neumann et al., 2007] Neumann, H., Yazdanbakhsh, A., and Mingolla, E. (2007). Seeing surfaces: The brain's vision of the world. *Physics of Life Rev.*, 4:189–222.

[Noback et al., 2005] Noback, Strominger, Demarest, Ruggiero, C., Norman, Robert, and David (2005). *The Human Nervous System: Structure and Function (Sixth. Ed.).* Humana Press.

[Nurminen and Angelucci, 2014] Nurminen, L. and Angelucci, A. (2014). Multiple components of surround modulation in primary visual cortex: Multiple neural circuits with multiple functions? *Vision Research*, 104:47–56.

[O'Herron and von der Heydt, 2011] O'Herron, P. and von der Heydt, R. (2011). Representation of object continuity in the visual cortex. *Journal of Vision*, 11(2):1–9.

[O'Herron and von der Heydt, 2013] O'Herron, P. and von der Heydt, R. (2013). Remapping of Border Ownership in the Visual Cortex. *J. Neuroscience*, 33(5):1964–1974.

[Pack and Born, 2001] 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. *Letters to Nature*, 409(6823):1040–1042.

[Parent and Zucker, 1989] Parent, P. and Zucker, S. (1989). Trace Inference, Curvature Consistency, and Curve Detection. *IEEE Trans. PAMI*, 11(8):823–839.

[Pasupathy and Connor, 1999] Pasupathy, A. and Connor, C. (1999). Responses to contour features in macaque area V4. *J. Neurophysiology*, 82:2490–2502.

[Pavan et al., 2013] Pavan, A., Marotti, R. B., and Mather, G. (2013). Motion–form interactions beyond the motion integration level: Evidence for interactions between orientation and optic flow signals. *Journal of Vision*, 13(6):1–13.

[Peterhans and von der Heydt, 1991] Peterhans, E. and von der Heydt, R. (1991). Subjective contours – bridging the gap between psychophysics and physiology. *Trends in Neuroscience*, 14(3):112–119.

[Piëch et al., 2013] Piëch, V., Li, W., Reeke, G. N., and Gilbert, C. D. (2013). Network Model of Top–Down Influences on Local Gain and Contextual Interactions in Visual Cortex. *Proc. Nat'l. Acad. of Sciences USA*, 110(43).

[Qiu and von der Heydt, 2005] Qiu, F. T. and von der Heydt, R. (2005). Figure and Ground in the Visual Cortex: V2 combines Stereoscopic Cues with Gestalt Rules. *Neuron*, 47:155–166.

[Rao and Ballard, 1999] Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

[Raudies et al., 2011] Raudies, F., Mingolla, E., and Neumann, H. (2011). A model of motion transparency processing with local center-surround interactions and feedback. *Neural Computation*, 23(11):2868–914.

[Raudies and Neumann, 2010] Raudies, F. and Neumann, H. (2010). A Model of Neural Mechanisms in Monocular Transparent Motion Perception. *Journal of Physiology Paris*, 104:71–83.

[Reynolds and Heeger, 2009] Reynolds, J. H. and Heeger, D. J. (2009). The Normalization Model of Attention. *Neuron*, 61:168–85.

[Riesenhuber and Poggio, 1999] Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025.

[Riesenhuber and Poggio, 2000] Riesenhuber, M. and Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, 3:1199–1204.

[Rodríguez-Sánchez and Tsotsos, 2012] Rodríguez-Sánchez, A. J. and Tsotsos, J. K. (2012). The Roles of Endstopped and Curvature Tuned Computations in a Hierarchical Representation of 2D Shape. *PLoS ONE*, 7(8):1–13.

[Roelfsema et al., 2002] Roelfsema, P., Lamme, V., Spekreijse, H., and Bosch, H. (2002). Figure–ground segregation in a recurrent network architecture. *J. Cognitive Neuroscience*, 14(4):525–537.

[Roelfsema, 2006] Roelfsema, P. R. (2006). Cortical Algorithms for Perceptual Grouping. *Annual Review of Neuroscience*, 29:203–227.

[Roelfsema et al., 2007] Roelfsema, P. R., Tolboom, M., and Khayat, P. S. (2007). Different processing phases for features, figures, and selective attention in the primary visual cortex. *Neuron*, 56(5):785–92.

[Rust et al., 2006] Rust, N. C., Mante, V., Simoncelli, E. P., and Movshon, J. A. (2006). How mt cells analyze the motion of visual patterns. *Nature Neurosci.*, 9:1421–1431.

[Salin and Bullier, 1995] Salin, P.-A. and Bullier, J. (1995). Corticocortical connections in the visual system: Structure and function. *Physiological Reviews*, 75(1):107–54.

[Sceniak et al., 1999] Sceniak, M. P., Ringach, D. L., Hawken, M. J., and Shapley, R. (1999). Contrasts Effect on Spatial Summation by Macaque V1 Neurons. *Nature Neuroscience*, 2(8):733—-9.

[Schels et al., 2013] Schels, M., Glodek, M., Meudt, S., Scherer, S., Schmidt, M., Layher, G., Tschechne, S., Brosch, T., Hrabal, D., Walter, S., Traue, H. C., Palm, G., Neumann, G., and Schwenker, F. (2013). Multi–Modal Classifier–Fusion for the Recognition of Emotions. In Rojc, M. and Campbell, N., editors, *Coverbal Synchrony in Human–Machine Interaction*, chapter 4, pages 73–98. Science Publishers.

[Schwartz et al., 1983] Schwartz, E., Dee, R., Albright, T., and Gross, C. (1983). Shape Recognition and inferior temporal neurons. *Proc. Nat'l. Acad. of Science USA*, 80:5776–5778.

[Self et al., 2012] Self, M. W., Kooijmans, R. N., Supér, H., Lamme, V. A., and Roelfsema, P. R. (2012). Different Glutamate Receptors Convey Feedforward and Recurrent Processing in Macaque V1. *Proc. Nat'l. Acad. of Sciences USA*, 109(27):11031–6.

[Serrano-Gotorredona et al., 2009] Serrano-Gotorredona, R., Oster, M., Lichtsteiner, P., Linares-Barranco, A., Paz-Vicente, R., Gómez-Rodríguez, F., Camuñas Mesa, L., Berner, R., Rivas-Pérez, M., Delbr"uck, T., Liu, S.-C., Douglas, R., H"afliger, P., Jiménez-M, Ballcels, A. C., Serrano-Gotarredona, T., J, A.-J. A., and Linares-Barranco, B. (2009). CAVIAR: A 45k Neuron, 5M Synapse, 12G Connects/s AER Hardware Sensory–Processing–Learning–Actuating System for High–Speed Visual Object Recognition and Tracking. *IEEE Transactions on Neural Networks*, 20(9):1417–1438.

[Serre and Poggio, 2010] Serre, T. and Poggio, T. (2010). A neuromorphic approach to computer vision. *Communications of the ACM*, 53(10):54.

[Serre et al., 2007] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *TPAMI*, 29(3):411–26.

[Shi et al., 2013] Shi, X., Wang, B., and Tsotsos, J. K. (2013). Early Recurrence Improves Edge Detection. In *Proc. 24th British Machine Vision Conference (BMVC)*, Bristol, UK.

[Simoncelli and Heeger, 1998] Simoncelli, E. P. and Heeger, D. J. (1998). A model of neural responses in visual area mt. *Vision Res.*, 38:743–761.

[Spratling, 2008] Spratling, M. W. (2008). Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience*, 2(4):1–8.

[Tabernik et al., 2014] Tabernik, D., Kristan, M., Boben, M., and Leonardis, A. (2014). Using discriminative analysis for improving hierarchical compositional models. *Proc. 19th Computer Vision Winter Workshop in Z. Kukelova, Krtiny, Czech Republic, Feb. 3–5.*

[Thielscher and Neumann, 2003] Thielscher, A. and Neumann, H. (2003). Neural Mechanisms of Cortico–Cortical Interaction in Texture Boundary Detection: A Modeling Approach. *Neuroscience*, 122:921–939.

[Trappenberg, 2002] Trappenberg, T. (2002). *Fundamentals of Computational Neuroscience.* Oxford Univ. Press.

[Tschechne et al., 2014a] Tschechne, S., Brosch, T., Sailer, R., von Egloffstein, N., Abdul-Kreem, L. I., and Neumann, H. (2014a). On event-based motion detection and integration. In *8th International Conference on Bio-inspired Information and Communications Technologies (BICT)*.

[Tschechne and Neumann, 2011a] Tschechne, S. and Neumann, H. (2011a). Kinetic occlusions and ordinal depth assignments – a neural model. In *European Conference on Vision and Perception (ECVP)*, volume 40, page 103.

[Tschechne and Neumann, 2011b] Tschechne, S. and Neumann, H. (2011b). Ordinal depth from occlusion using optical flow: A neural model. *Journal of Vision (abstracts)*, 11(11):Article 716.

[Tschechne and Neumann, 2012] Tschechne, S. and Neumann, H. (2012). The structure of optical flow for figure–ground segregation. *Journal of Vision (abstracts)*, 12(9):Article 242.

[Tschechne and Neumann, 2014a] Tschechne, S. and Neumann, H. (2014a). Hierarchical representation of shapes in visual cortex -– from localized features to figural shape segregation. *Front. Comput. Neurosci.*, 8(93).

[Tschechne and Neumann, 2014b] Tschechne, S. and Neumann, H. (2014b). Unified Representation of Motion and Motion Streak Patterns in a Model of Cortical Form–Motion–Interaction. *Journal of Vision (abstracts)*, 14(10):Article 18.

[Tschechne et al., 2014b] Tschechne, S., Sailer, R., and Neumann, H. (2014b). Bio–inspired optic flow from event–based neuromorphic sensor input. In *ANNPR 2014*, LNAI 8774, pages 171–182.

[Tsotsos, 1988] Tsotsos, J. K. (1988). *Computational Processes in Human Vision: An Interdisciplinary Perspective*, chapter How Does Human Vision beat the Computational Complexity of Visual Perception?, pages 286–338. Ablex Press, Norwood, NJ.

[Tsotsos, 2005] Tsotsos, J. K. (2005). *Neurobiology of Attention*, chapter Computational foundations for attentive processes, pages 3–7. Elsevier, Amsterdam, NL.

[Tsui et al., 2010] Tsui, J. M. G., Hunter, J. N., Born, R. T., and Pack, C. C. (2010). The role of v1 surround suppression in mt motion integration. *J Neurophysiol.*, 103:3123–3138.

[Ullman, 1995] Ullman, S. (1995). Sequence-seeking and counter streams: A computational model for bi-directional information flow in the visual cortex. *Cerebral Cortex*, 5(1):1–11.

[Ungerleider and Haxby, 1994] Ungerleider, L. G. and Haxby, J. V. (1994). 'what' and 'where' in the human brain. *Current Opinion in Neurobiology*, 4(2):157–65.

[Van Essen and Gallant, 1994] Van Essen, D. C. and Gallant, J. L. (1994). Neural Mechanisms of form and motion processing in the primate visual system. *Neuron*, 13(1):1–10.

[vanSanten and Sperling, 1985] vanSanten, J. P. H. and Sperling, G. (1985). Elaborated Reichardt detectors. *Journal of the Optical Society of America*, 2(2):300–321.

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

[Wallach, 1935] Wallach, H. (1935). Über visuell wahrgenommene Bewegungsrichtung. Dissertationsschrift, Philosophische Fakultät der Universität Berlin.

[Wallis and Arnold, 2009] Wallis, T. and Arnold, D. (2009). Motion-induced blindness and motion streak suppression. *Curr Biol.*, 19(4):325–329.

[Weidenbacher and Neumann, 2009] Weidenbacher, U. and Neumann, H. (2009). Extraction of Surface–Related Features in a Recurrent Model of V1–V2 Interactions. *PloS ONE*, 4(6):e5909.

[Williford and von der Heydt, 2013] Williford, J. and von der Heydt, R. (2013). Border–Ownership Coding. *Scholarpedia*, 8(19,30040).

[Wilson and Wilkinson, 1998] Wilson, H. and Wilkinson, F. (1998). Detection of global structure in Glass patterns: implications for form vision. *Vision Research*, 38:2933–2947.

[Wuerger et al., 1996] Wuerger, S., Shapley, R., and Rubin, N. (1996). On the visually perceived direction of motion – by Hans Wallach: 60 years later. *Perception*, 25:1317–1367.

[Yamane et al., 2008] Yamane, Y., Carlson, E., Bowman, K., Wang, Z., and Connor, C. (2008). A neural code for three–dimensional object shape in macque inferotemporal cortex. *Nature Neuroscience*, 11(11):1352–1360.

[Yau et al., 2013a] Yau, J., Pasupathy, A., Brincat, S., and Connor, C. (2013a). Curvature processing dynamics in macaque area V4. *Cerebral cortex*, 23:198–209.

[Yau et al., 2013b] Yau, J. M., Pasupathy, A., Brincat, S. L., and Connor, C. E. (2013b). Curvature Processing Dynamics in Macaque V4. *Cerebral Cortex*, 23:198–209.

[Zhaoping, 1998] Zhaoping, L. (1998). A neural model of contour integration in the primary visual cortex. *Neural Computation*, 10:903–940.

[Zhou et al., 2000] 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.

Summary

This thesis presents models of visual form and motion processing and how neural representations in either modality are influenced by the context their are embedded within. The work presents the physiological and theoretical foundations that the investigations are based upon, as well as two models of visual processing. The work's content is organized into three main chapters.

Chapter 2 presents the physiological background of vision including a short summary of the optical tract from the retina to the occipital cortex. More highlight is put on the cortical mechanisms of visual processing and the neural representation of visual stimuli. While keeping the mechanisms of single-compartment-models in mind, a generic cascade model of visual processing in presented that presents canonical mechanisms of model processing areas. It has been used in previous models of visual processing and contains three major stages of processing for a model visual area, namely a stage of initial filtering, a stage of contextual feedback modulation and a stage of non-linear filter operations by means of normalization mechanisms. The subsequent contributions in this thesis are based on these generic components.

Chapter 3 proposes a model of motion processing along the dorsal pathway using the event-based Dynamic Vision Sensor (DVS). The sensor provides a constant stream of visual events that densely samples the plenoptic function. Using the neuromorphic output of the sensor, the proposed model applies physiological findings for the generation of a model of optic flow estimation. Our model makes use of neural mechanisms that model the processing capabilities of early and intermediate stages in dorsal pathway of visual cortex. Spatio-temporally inseparable initial motion filters are combined from separable components to yield direction-selective filters for the initial stage of motion estimation (V1). The thesis elaborates on subsequent processing mechanisms that pick up physiological and psychophysical findings along the dorsal pathway. Motion integration and its effects on increasing the robustness of initial motion representation to noise and outliers is presented. Physiological investigations suggest that the solution of the

aperture problem is achieved at an MT level, but early mechanisms at V1 might also contribute to the solution. We present mechanisms of early surround inhibition that show to contribute to the solution of the aperture problem. The dense processing allows an representation of motion streaks in the ventral (form) pathway, that comes with the unique representation of visual events. These motion streaks or speed lines provide a spatial code for motion processing. In this context the interaction of the dorsal and ventral stream for motion processing is shown. The functionality is demonstrated with various stimuli. The chapter closes with considerations of how a motion algorithm based on an event-based representation of visual input is designed in order to keep the unique processing advantages like instantaneous estimation of local estimates and sparse representation. The dense representation of visual events requires altered processing mechanisms in that a continuous updating process is required. To conclude the chapter, the computational advantages of such and algorithm is described using theoretical and empirical data as well as knowledge about the hardware.

Chapter 4 proposes a model of mechanisms in ventral pathway that segments visual scenes into image regions and allows their combination to surfaces and prototypical objects. Multiple mutually connected areas in the ventral cortical pathway receive visual input and extract local form features that are subsequently grouped into increasingly complex, more meaningful image elements. We propose a mechanism how such a distributed network of processing is capable of representing highly articulated changes in shape boundary as well as subtle curvature changes. We propose a recurrent computational network architecture that utilizes hierarchical distributed representations of shape features to encode surface and object boundaries over different scales of resolution. Our model makes use of neural mechanisms that model the processing capabilities of early and intermediate stages in visual cortex, namely areas V1-V4 and IT. We suggest that multiple specialized component representations interact by feedforward hierarchical processing that is combined with feedback signals driven by representations generated at higher stages. Based on this, global configurational as well as local information is made available to distinguish changes in the object's contour. Once the outline of a shape has been established, contextual contour configurations are used to assign border ownership directions and thus achieve segregation of figure and ground. The model, thus, proposes how separate mechanisms contribute to distributed hierarchical cortical shape representation and combine with processes of figure-ground segregation. Our model is probed with a selection of stimuli to illustrate processing results at different processing stages. We especially highlight how modulatory feedback connections contribute to the processing of visual input at various stages in the processing hierarchy.

Zusammenfassung

Diese Arbeit präsentiert Modelle visueller Form- und Bewegungsverarbeitung und zeigt wie der Einfluss kontextueller (umgebender) Aktivitäten die neuronale Repräsentation beeinflusst. Die Arbeit führt die physiologischen und theoretischen Grundlagen ein, auf denen die Untersuchungen beruhen, danach werden zwei Modelle visueller Informationsverarbeitung untersucht. Die Arbeit gliedert sich in drei Kapitel.

Kapitel 2 zeigt den physiologischen Hintergrund der Untersuchungen und bietet eine kurze Zusammenfassung der Komponenten des optischen Pfades. Dabei wird auf die zu Grunde liegenden kortikalen Mechanismen visueller Informationsverarbeitung und die neuronale Repräsentation visueller Signale fokussiert. Hierbei zeigt die Arbeit die Mechanismen einfacher Neuronenmodelle (*single-compartment models*) und präsentiert die modulare Verarbeitungskaskade, mit denen optische Areale modelliert werden. Diese wurde bereits ihn vorausgehenden Arbeiten eingesetzt. Sie gliedert sich in drei hauptsächliche Verarbeitungsschritte: Eine initiale Filterung, eine modulierende Stufe, bei der Kontextinformation eingebracht wird und eine Stufe nichtlinearer Operationen in Form eines Normalisierungsmechanismus. Die weiteren Ausführungen dieser Arbeit beruhen auf diesem Kaskadenmodell.

Kapitel 3 zeigt ein Modell der Bewegungsverarbeitung entlang des dorsalen Pfades, das die ereignisbasierten Ausgabedaten des *Dynamic Vision Sensors (DVS)* verarbeitet. Dieser Sensortyp tastet optischen Input in hoher zeitlicher Auflösung ab und liefert einen asynchronen Strom visueller Ereignisse. Das Modell verknüpft diese neuromorphe Repräsentation mit physiologischen Erkenntnissen über Modelle der Bewegungsschätzung und benutzt die neuronalen Mechanismen der frühen und mittleren Verarbeitungsschritte im visuellen Kortex entlang des dorsalen Pfades. In einer initialen Stufe (V1) werden raum-zeitlich separierbare Filterkomponenten zu raum-zeitlich nicht separierbaren Filtern zusammengesetzt, die der Bewegungserkennung dienen. Die Arbeit führt hierbei die weiterführenden Verarbeitungsschritte entlang des dorsalen Pfades auf, die auf physiologischen

Erkenntnissen beruhen. Der Einfluss der Bewegungsintegration und ihr positiver Effekt auf das enthaltene Rauschen in der initialen Schätzung wird aufgezeigt. Die Arbeit präsentiert den Einfluss von frühen hemmenden Signalen auf das Blendenproblem mit Hilfe kontextueller Modulation. Die dichte zeitliche Abtastung des DVS ermöglicht weiterhin eine Repräsentation von Bewegungsslinien (*motion streaks*) in einem ursprünglich der Formverarbeitung zugeordneten Areal (V2). In diesem Kontext wird eine mögliche Interaktion zwischen dorsalem und ventralem Pfad aufgezeigt. Die Funktionalität unseres Modells wird mit mehreren Eingabesignalen getestet. Das Kapitel schließt mit Details über eine algorithmische Implementierung des Modells, das die einzigartigen Eigenschaften der ereignisbasierten Daten erhält. Zu diesen Eigenschaften zählt eine spärliche Datenhaltung und kontinuierliche, auf einzelnen Ereignissen beruhende Abarbeitung, die angepasste Verarbeitungsschritte erfordern. Am Ende des Kapitels wird auf die Fragestellung nach möglichen Verarbeitungsvorteilen mit einer Komplexitätsanalyse eingegangen.

Kapitel 4 präsentiert ein Modell der Formverarbeitung entlang des ventralen Pfades, das eine verteilte Repräsentation von Szenen in Form von lokalen Komponenten ermöglicht und diese zu einer Repräsentation von Oberflächen und prototypischen Objekten zusammenfasst. Mehrere miteinander verbundene Modellareale entlang des ventralen Pfades erhalten visuellen Input und extrahieren lokale Formeigenschaften. Diese werden nachfolgend zu komplexeren, bedeutungsvolleren Bildelementen zusammengefasst. Wir präsentieren einen Mechanismus, der in einem verteilten Verarbeitungsnetzwerk die Repräsentation deutlicher Schwankungen im Konturverlauf ebenso ermöglicht wie kleinere Veränderungen. Unser rekurrentes Verarbeitungsnetzwerk nutzt hierarchisch verteile Formrepräsentationen, um Oberflächen und Objektgrenzen über mehrere Skalen zu repräsentieren. Wir benutzen hierzu die Verarbeitungseigenschaften früher und mittlerer Areale im visuellen Kortex, die Areale V1 bis V4 und IT. In unserem Modell interagieren mehrere spezialisierte Komponenten in einem hierarchischen, vorwärts gerichteten Verarbeitungsschritt mit Repräsentationen, die durch rückwärts gerichtete Verbindungen aus späteren Verarbeitungsstufen stammen. Dadurch werden globale, kontextuelle und lokale Informationen verfügbar, um Änderungen im Konturverlauf festzustellen. Sobald der Konturverlauf fest steht wird kontextuelle Information benutzt, um die Zugehörigkeit einer Objektkante zum Vorder- oder Hintergrund festzulegen. Unser Modell präsentiert dadurch die einzelne Mechanismen in einem hierarchischen Modell zur Objektrepräsentation beitragen. Besonderes Augenmerk wird hierbei auf den Einfluss modulierender kontextueller Verbindungen gelegt und wie sich diese auf die Verarbeitung in verschiedenen Stufen entlang der Verarbeitungskette auswirken.