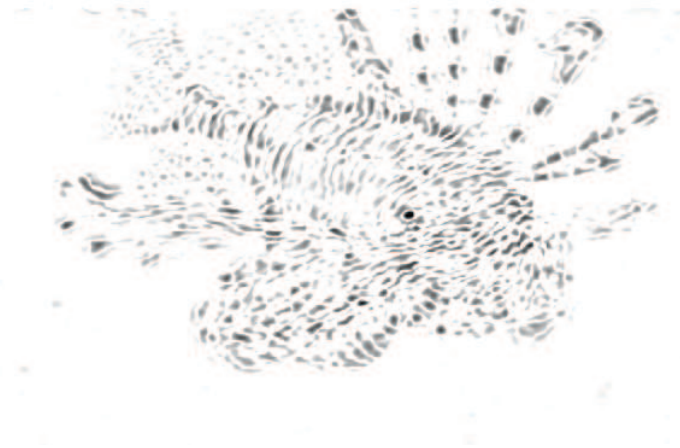
Universität Bremen
Fachbereich 03: Mathematik, Informatik

Intrinsic dimensionality in vision

Nonlinear filter design and applications

Tobias Kluth



Dissertation
zur Erlangung des Grades eines Doktors der
Ingenieurwissenschaften
– Dr.-Ing. –

vorgelegt im Fachbereich 03 der Universität Bremen am
03. März 2015
verteidigt am
15. Juli 2015

Tobias Kluth

Intrinsic dimensionality in vision

Nonlinear filter design and applications

1st Reviewer:
Dr. Christoph Zetzsche, University of Bremen
2nd Reviewer:
Prof. Dr. Udo Frese, University of Bremen

Date of submission: March 03, 2015
Date of defense: July 15, 2015

Preface

The work presented in this thesis was carried out during my work in the working group Cognitive Neuroinformatics affiliated to the University of Bremen. I started my research regarding biological vision after I had finished my studies in mathematics at the University of Bremen.

It turned out that having a background in mathematics is not always a blessing. As a mathematician I always have an intrinsic motivation to understand the structures within an abstract "mathematical" world. In the field of computational neuroscience including vision the results have commonly an empirical nature. There the dilemma starts. Roughly spoken the empirical world is "dirty" compared to the "nice" mathematical world. As fast as an empirical result is described by an abstract structural relation, even faster another explanation can be found. The number of possible explanations seems to be infinitely large. This causes big problems in what we really think to know. In the relatively young field of neuroscience the number of empirical results continuously increases. As a result, the number of conclusions how the neural structure should work also increases. But what do we really know about these structures if the conclusions are based on results within the abstract mathematical world? This discussion kept and still keeps a lot of philosophers busy. However, I believe in the existence of a relation between the empirical world and the abstract mathematical world. The important point is that without explicitly knowing anything about the relation between the empirical and the mathematical world, one has to be very careful in drawing conclusions regarding the empirical world based on results derived in the abstract mathematical world. But the advantage, in my opinion, of results developed in a mathematical way is their inherent logic. Ignoring the criticism on the axiomatic system not to be complete, mathematical results which are grounded on set theory and logic can be derived. Within this framework everything is logical and structured. This circumstance is my understanding of "beauty in mathematics". Assuming that there exists a relation between the two worlds, I hope that this beauty can somehow be transfered to the empirical world. Even if this can be done only partially, it is a success. This hope motivated my work on the visual system. I started my adventure through different disciplines like signal processing, linear system theory, nonlinear system theory, differential geometry, topology, probability theory, and information theory.

First I would like to thank my supervisor Dr. Christoph Zetzsche for his motivation, teaching, guidance and many comments which enabled me to develop my understanding of the subject. I also would like to thank Prof. Dr. Kerstin Schill for the financial support and the employment in her working group. I am heartily thankful to my co-authors Thomas Reineking, David Nakath, Konrad Gadzicki, Christoph Zetzsche, and Kerstin Schill for the beneficial work on the articles which are part of this thesis. I am indebted to all of my colleagues for their support and many discussions. I wish to thank my sister, Sarah Kluth, and my coworker,

David Nakath, for proofreading and comments. At last and most importantly I wish to thank my partner Karina Heidt who supports me in everything, cares for me and has complete understanding.

Bremen, 3rd March 2015
Tobias Kluth

Abstract

Biological vision and computer vision cannot be treated independently anymore. The digital revolution and the emergence of more and more sophisticated technical applications caused a symbiosis between the two communities. Competitive technical devices challenging the human performance rely increasingly on algorithms motivated by the human vision system. On the other hand, computational methods can be used to gain a richer understanding of neural behavior, e.g. the behavior of populations of multiple processing units. The relations between computational approaches and biological findings range from low level vision to cortical areas being responsible for higher cognitive abilities.

In early stages of the visual cortex cells have been recorded which could not be explained by the standard approach of orientation- and frequency-selective linear filters anymore. These cells did not respond to straight lines or simple gratings but they fired whenever a more complicated stimulus, like a corner or an end-stopped line, was presented within the receptive field. Using the concept of intrinsic dimensionality, these cells can be classified as intrinsic-two-dimensional systems. The intrinsic dimensionality determines the number of degrees of freedom in the domain which is required to completely determine a signal. A constant image has dimension zero, straight lines and trigonometric functions in one direction have dimension one, and the remaining signals, which require the full number of degrees of freedom, have the dimension two. In this term the reported cells respond to two dimensional signals only. Motivated by the classical approach, which can be realized by orientation- and frequency-selective Gabor-filter functions, a generalized Gabor framework is developed in the context of second-order Volterra systems. The generalized Gabor approach is then used to design intrinsic two-dimensional systems which have the same selectivity properties like the reported cells in early visual cortex.

Numerical cognition is commonly assumed to be a higher cognitive ability of humans. The estimation of the number of things from the environment requires a high degree of abstraction. Several studies showed that humans and other species have access to this abstract information. But it is still unclear how this information can be extracted by neural hardware. If one wants to deal with this issue, one has to think about the immense invariance property of number. One can apply a high number of operations to objects which do not change its number. In this work, this problem is considered from a topological perspective. Well known relations between differential geometry and topology are used to develop a computational model. Surprisingly, the resulting operators providing the features which are integrated in the system are intrinsic-two-dimensional operators. This model is used to conduct standard number estimation experiments. The results are then compared to reported human behavior.

The last topic of this work is active object recognition. The ability to move the information gathering device, like humans can move their eyes, provides the opportunity to choose the

next action. Studies of human saccade behavior suggest that this is not done in a random manner. In order to decrease the time an active object recognition system needs to reach a certain level of performance, several action selection strategies are investigated. The strategies considered within this work are based on information theoretical and probabilistic concepts. These strategies are finally compared to a strategy based on an intrinsic-two-dimensional operator.

All three topics are investigated with respect to their relation to the concept of intrinsic dimensionality from a mathematical point of view.

Zusammenfassung

Biologische und technische Sehsysteme können nicht mehr unabhängig voneinander betrachtet werden. Die digitale Revolution und die Entwicklung von immer komplexeren technischen Anwendungen haben zu einer Symbiose zwischen den beiden Felder geführt. Technische Systeme, die an die Leistungsfähigkeit des Menschen herankommen wollen, basieren zunehmend auf Mechanismen des menschlichen Sehsystems. Auf der anderen Seite ermöglicht das Nutzen technischer Ansätze den Gewinn neuer Erkenntnisse über die Funktionsweise komplexer neuronaler Systeme, wie beispielsweise das Verhalten von Netzwerken bestehend aus parallel geschalteten Verarbeitungseinheiten. Der Umfang dieser symbiotischen Verbindung reicht von frühen Verarbeitungsstufen des visuellen Systems bis hin zu höheren kognitiven Fähigkeiten.

In frühen Stufen des visuellen Kortex wurden Zellen gefunden, die nicht mehr allein durch den linearen Orientierungs- und Frequenz-selektiven Ansatz erklärt werden können. Diese Zellen reagieren nicht auf Linien oder einfache Gitterstrukturen, sondern werden durch komplexere Stimuli im rezeptiven Feld, wie zum Beispiel eine Ecke oder ein Linienende, gereizt. Unter Verwendung des Prinzips der intrinsischen Dimensionalität können diese Zellen als intrinsisch zwei-dimensional klassifiziert werden. Die intrinsische Dimensionalität bestimmt dabei die Anzahl der Freiheitsgrade im Definitionsbereich, die benötigt wird, um das Signal komplett zu bestimmen. Ein konstantes Signal hat die intrinsische Dimensionalität Null, Linien oder trigonometrische Funktionen in eine Richtung haben die Dimensionalität Eins und Signale, die beide Freiheitsgrade benötigen, haben die Dimensionalität Zwei. Die beobachteten Neurone reagieren somit nur auf intrinsisch-zwei-dimensionale Signale. Basierend auf dem klassischen Ansatz, der lineare Orientierungs- und Frequenz-selektive Gabor-Filter benutzt, wird ein generalisiertes nichtlineares Gabor-Filter im Kontext der Volterra-Systeme zweiter Ordnung entwickelt. Dieser Ansatz wird benutzt, um intrinsisch-zwei-dimensionale Systeme zu implementieren, die die Selektivitätseigenschaften der beobachteten Neurone besitzen.

Die numerische Wahrnehmung des Menschen wird oft als eine höhere kognitive Fähigkeit klassifiziert. Das Bestimmen der Anzahl aus der Umgebung erfordert ein hohes Abstraktionsvermögen. Eine Vielzahl von Studien mit Menschen und anderen Spezies hat gezeigt, dass die getesteten Individuen einen Zugang zu dieser Art von Information haben. Es ist jedoch immer noch aktueller Forschungsgegenstand, wie das Gehirn und somit ein neuronales Netzwerk diese Information aus der Umgebung extrahiert. Um mit dieser Problematik umgehen zu können, muss die starke Invarianzeigenschaft, die sich hinter der Anzahl verbirgt, berücksichtigt werden. Eine Vielzahl an Transformationen kann beispielsweise auf ein Objekt angewandt werden, ohne dass sich dessen Anzahl verändert. Diese Problemstellung wird im Rahmen dieser Arbeit aus Sicht der mathematischen Topologie untersucht. Die Verbindung zwischen Topologie und Differentialgeometrie wird ausgenutzt, um ein implementierbares Modell zu

entwickeln. Interessanterweise gibt es eine Verbindung zwischen den extrahierten Merkmalen innerhalb des Modells und dem Konzept der intrinsischen Dimensionalität. Die extrahierten Merkmale werden durch intrinsisch-zwei-dimensionale Operatoren zur Verfügung gestellt. Das entwickelte Modell wird in typischen Experimenten der numerischen Wahrnehmung getestet und mit menschlichem Verhalten verglichen.

Der letzte Themenbereich behandelt die aktive Objekterkennung. Die Fähigkeit den informationsbeschaffenden Sensor, wie der Mensch sein Auge, zu bewegen, ermöglicht es, die nächste durchzuführende Aktion auszuwählen. Untersuchungen menschlicher Augenbewegungen haben ergeben, dass diese Aktionsauswahl nicht zufällig passiert, sondern einer gewissen Systematik zu unterliegen scheint. Um die Geschwindigkeit des Erkennungsprozesses zu erhöhen, beziehungsweise die Anzahl der durchgeführten Aktionen zum Erreichen einer bestimmten Performanz zu verringern, werden unterschiedliche Auswahlverfahren zur Bestimmung der nächsten Aktion untersucht. Die Strategien basieren sowohl auf informationstheoretischen, als auch auf probabilistischen Größen. Letztendlich werden diese Ansätze mit einer Strategie basierend auf einem intrinsisch-zwei-dimensionalen Operator verglichen.

Alle Themenbereiche werden auch im Hinblick auf ihren mathematischen Zusammenhang zum Konzept der intrinsichen Dimensionalität untersucht.

Contents

1 Introduction **1**
1.1 The visual pathway 2
1.2 Linear models in the visual system 4
1.3 The concept of intrinsic dimensionality 7

2 A generalized Gabor approach for $i2D$-feature extraction **12**
2.1 Related Work 12
2.2 Mathematical preliminaries 14
2.3 Generalized Gabor to obtain second-order selectivity 15
2.3.1 Classical Gabor approach 15
2.3.2 Second-order Volterra system 20
2.3.3 Analysis of selected $i0D$, $i1D$, and $i2D$ signals 22
2.3.4 Generalized Gabor approach 29
2.3.5 Results 32
2.4 Article: Statistical invariants of spatial form: From local AND to numerosity 39

3 Role of curvature i2D-features in numerical cognition **50**
3.1 Related work 50
3.2 Mathematical preliminaries 52
3.3 Article: Spatial numerosity: A computational model based on a topological invariant 63
3.4 Article: Numerosity as a topological invariant 80

4 Action selection for object recognition and the influence of isotropic $i2D$-features **125**
4.1 Related Work 125
4.2 Mathematical preliminaries 126
4.3 Article: Active sensorimotor object recognition in three-dimensional space . . 131
4.4 Article: Affordance-based object recognition using interactions obtained from a utility maximization principle 145
4.5 Article: Adaptive information selection in images: Efficient naive bayes nearest neighbor classification 153

5 Summary and outlook **166**

Preliminary remark

In consultation with the reviewers the articles which had not been published at the date of submission of this work were replaced by the latest version which was available at the date of publication of this work.

1 Introduction

The visual system of humans and animals is one of the most important modalities for interacting with the world. It extracts relevant information nearly on-the-fly. Knowing what runs in one's direction, is it an enemy or not, can decide on survival. Similarly, the decision where the higher amount of food is located plays an important role as well. Although the study of vision has a long history, the definite knowledge about the functional principals underlying the information extraction process is still limited. Nowadays, the study of biological vision cannot be seen independent of computational applications anymore. The digital revolution within the last decades, the emergence of fields like computer vision as a sub-discipline of computer science, and the development of more and more sophisticated technical devices relying on visual information caused a symbiosis between these fields. On the one hand, technical applications trying to be as efficient as humans incorporate methods motivated by the human visual system. Assuming that the visual system is optimally adapted to the natural world and its structures, gives the justification to hope that applying biological principals yield better technical systems. On the other hand, the study of more and more complex artificial systems in computer science helps to obtain new insights in reported brain functionalities. This holds true for low level vision, i.e. information extracted in the early visual cortex, as well as higher cognitive abilities of humans. In low level vision the modeling of neuronal behavior can yield better strategies to encode natural images or to provide relevant features for high level classification tasks. The understanding of higher cortical functionalities can yield better algorithms to extract relevant information with computer vision systems. For example, the number of objects or the identity of an object.

The subtitle of this work "nonlinear filter design and applications" recalls the content of this work. Three different main topics are addressed here. On the one hand, functional models for neurons in early visual cortex are developed and investigated. This belongs to the field of low level vision. And on the other hand, higher cognitive abilities are investigated. In particular, these applications are numerical cognition and active object recognition. But what motivates the combination of these fields? One argument would be that all investigated functionalities are realized in the human brain differing in their location only. This would be a relatively weak argument for this compilation. The stronger argument is that each topic is influenced significantly by a specific concept, the concept of intrinsic dimensionality. The influence is considered within each chapter addressing another research question. This directly leads to the three major questions affecting this work.

- How can neurons of the visual cortex be modeled so that they show a significantly nonlinear behavior in line with the concept of intrinsic dimensionality? (Section 2)
- How can numerical cognition be modeled from operations determined by the concept of intrinsic dimensionality so that human behavior can be explained? (Section 3)

- How can the action selection for active object recognition be influenced by information theoretical quantities and operations determined by the concept of intrinsic dimensionality? (Section 4)

Possible answers to these questions are presented in the referenced sections. Each section contains a related work section and a mathematical preliminaries section. In particular, the relation to the concept of intrinsic dimensionality for the second and the third research question is stated in Section 3.2 and 4.2. A short overview about the biological vision system is given in Section 1.1 and linear modeling approaches of functionalities in the early stages of the visual pathway are presented in Section 1.2. The concept of intrinsic dimensionality is introduced in a mathematical way in Section 1.3.

1.1 The visual pathway

This section gives a coarse overview about the anatomy and the physiology of the first parts of the visual pathway. For considerations in more detail, the reader is referred to textbooks like [34].

The visual information is gathered by the eye. An image is projected onto the retina which is an area equipped with photoreceptor cells. This area is located at the inner surface of the back part of the eye. One important observation regarding visual perception is that there exist regional differences in the information processing within the visual field. This goes back to first behavioral findings regarding letter perception in the periphery by Aubert and Foerster in 1857 [3]. In 1935 Osterberg published his results about the receptor density distribution on the retina in dependence on the eccentricity [60]. Two regions in the visual field are distinguished, the periphery and the fovea. The information which is gathered by a huge number of receptor cells on the retina is then processed by bipolar cells. The bipolar cells send the information to retinal ganglion cells which encode different aspects of the visual stimulus. Information like stimulus size, color, and movement, for example, is carried to the thalamus and then to the visual cortex. The important parts which are considered in more detail are the retina, the lateral geniculate nucleus, and the primary visual cortex V1.

The *retina* converts the projected image into neural responses. It is the innermost layer of the eye consisting of neurons and supporting cells and covering the choroid. The retina is derived from the neural tube such that it is part of the central nervous system. The neural retina contains five types of neurons: visual receptor cells, horizontal cells, bipolar cells, amacrine cells, and retinal ganglion cells. These cells are organized in multiple layers from the outside to the inside within the retina. This means that the light passing through the lens must pass through the other layers before it reaches the light-sensitive photo receptors. The area which corresponds to the central visual field, i.e. the fovea, is organized differently. The retina there consists of fewer layers such that a clearer image without obstacles is projected to the retina. The remaining cells are located in the surrounding of the fovea so that this

region is thicker. The union of both regions is referred to as macula. This part of the retina corresponds to the region with the highest resolution in the visual field. The retinal ganglion cells exit the retina in one specific region. There they build the optic nerve. In this region no photo receptors are located which causes the "blind spot" in the visual field. Humans have two kinds of photoreceptors, rods and cones. These types differ not only in their structure but also in their functionality. Rods are responsible for scotopic vision. They are more sensitive to light and can deal with low levels of illumination. The visual periphery is dominated by this kind of photoreceptors. Cones are responsible for photopic vision. There exist three different cell types which are sensitive to different bandwidths of the spectrum of light. Thus they are color sensitive. The fovea consists of cones only. This region is characterized by high visual acuity and color vision. The next layer within the retina consists of bipolar and horizontal cells. These cells have synapses to the photoreceptors. The bipolar cells also have synapses with the amacrine cells and the ganglion cells. Thus the horizontal cells have an implicit connection to the ganglion cells only. There exist two types of bipolar cells, ON-cells and OFF-cells. The ON-cells detect light regions in a dark background and OFF-cells detect dark regions in a light background [20]. Bipolar cells do not generate action potentials. The receptive field of bipolar cells, i.e. all photoreceptors having a synapse with the bipolar cell, differs by the type of photoreceptor. The receptive field of cells with cones are very small, i.e. up to one cone only. In contrast, the receptive fields with rods vary from a few up to fifty or more receptors. The ganglion cells, whose axons exit the retina to the lateral geniculate nucleus of the thalamus, are the connection to the brain. The receptive fields of the bipolar cells which have synapse with the respective ganglion cell determine the receptive field of the ganglion cell. A ganglion cell which has a synapse with an ON bipolar cell thus has an ON-center/OFF-surround receptive field. Analogously, a ganglion cell which has a synapse with an OFF bipolar cell has an OFF-center/ON-surround receptive field. In summary the visual information is processed by ∼125 million photoreceptors which converge to ∼10 million bipolar cells which again converge to ∼1 million ganglion cells. Further detail information about the retina can be found in [19].

The axons of the ganglion cells terminate in the *lateral geniculate nucleus (LGN)* which is responsible for visual perception. This is one destination beside three other nuclei (*superior colliculus* - control of eye movements, *pretectum* - control of pupillary reflex, *suprachiasmatic nucleus* - control of hormonal changes) [31, 62]. The LGN consists of three types of cells and is structured in six layers. Two magnocellular layers consist of larger mLGN cells which have a relatively large center-surround receptive field. These cells are insensitive to color and they are most sensitive to the movement of visual stimuli. Four parvocellular layer consist of smaller pLGN cells which have relatively small center-surround receptive fields. These cells are sensitive to color and can detect contrasts which build the basis shape discrimination. The third class of the smallest koniocellular neurons builds thin layers which are located between

the six principal layers. These neurons have a stronger color sensitivity such that they are well suited to support shape discrimination. The axons of these three cell types terminate in different layers of the primary visual cortex.

The *primary visual cortex (V1)* is located in the occipital lobe of the brain. The brain region is also referred to as striate cortex[1]. The striate cortex does the initial cortical processing of all visual information and sends the information to higher cortical areas. V1 is structured in six layers. The axons from the LGN terminate in layer 4 with lateral connections to layer 6. Layer 6 has a connection to the thalamus again. Other cortical areas are connected with layer 2 and layer 3. And layer 5 has outputs to other subcortical regions. The area V1 differs from other cortical regions in their number of neurons. Especially in layer 4 it has a higher density of neurons [63]. V1 has other remarkable properties. First, it is retinotopically organized which means that it contains a complete map of the visual field. Positions which are nearby in the visual field are also nearby in V1 [1]. This visual field position to cortex position mapping is not an isometric map. The small central part of the visual field corresponds to approximately 50% of the neurons located in V1 [80]. Second, this mapping transforms concentric circles and radial lines in the visual field to orthogonal lines in V1 [70]. For further detail information about the primary visual cortex, the reader is referred to [16].

1.2 Linear models in the visual system

The information processing in the visual pathway is commonly assumed to be parallel, i.e. different kinds of information are extracted simultaneously. This goes back to Campbell and Robson [11] who found out that the desensitization to high-contrast gratings depends on the orientation and the spatial frequency of the grating. They concluded that the visual pathway contains various orientation- and frequency-selective features which are processed in parallel channels. It turned out that linear system theory plays an important role in modeling these functionalities of the visual system. Before we start with examples from the literature in which the observed phenomena were modeled successfully with linear systems, we recall the following definition of a linear system. Any further information regarding properties of linear systems can be found in standard textbooks, e.g. [78], and is not part of this work.

Definition 1.1 (Linear system)**.** Let $T : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ be an operator defining a system which maps the input signal u to a response signal r, i.e. $r = T(u)$. The operator T is a *linear* operator if and only if

$$T(k_1u_1 + k_2u_2) = k_1T(u_1) + k_2T(u_2) \tag{1.1}$$

for arbitrary signals $u_1, u_2 \in L^2(\mathbb{R}^2)$ and $k_1, k_2 \in \mathbb{R}$. The system defined by the operator T is

[1] The name "striate cortex" is motivated by a visible stripe of axons which have their origin in the LGN and which terminate in this region.

(a) (b) (c)

Figure 1.1: The filter kernel of the Mexican hat function ($\sigma = 0.025$) is illustrated in (a). The filter response (c) of the image (b) is computed by the convolution of the image and the filter kernel.

then a linear system.

Remark 1.2. Let T be additionally shift-invariant, i.e. $r(x-\delta x) = T(u(x-\delta x))$ for arbitrary δx and a given response function $r = T(u)$. Knowing the impulse response $h \in L^1(\mathbb{R}^2)$[2] of the system T, the linear shift-invariant system T is determined by

$$r(x) = T(u)(x) = \int_{\mathbb{R}^2} h(x-y)u(y)\ dy = (h * u)(x). \tag{1.2}$$

The receptive field of a neuron is defined by all cells which have a synapse with this neuron. In the linear system the synaptic weighting of this receptive field corresponds to the impulse response h. Positive function values of h define the excitatory part of the receptive field. The negative function values define the inhibitory part. Thus the response of the neuron is completely determined by this linear filter operation. The stimulus u can be split into various filter outputs defined by different impulse responses. Consequently, the extracted information can be processed in parallel.

The receptive field of retinal ganglion cells is circular symmetric with an excitatory center and an inhibitory surrounding [47]. The interplay between excitatory and inhibitory regions is also known as lateral inhibition and increases the contrast at sharp edges in the stimulus. In order to model this functionality Marr and Hildreth [50], for example, proposed a filter kernel defined by the Laplace operator applied to a two-dimensional Gaussian function. The kernel is defined by

$$h(x) = -\frac{1}{\pi\sigma^4}\left(1 - \frac{\|x\|^2}{2\sigma^2}\right) e^{-\frac{\|x\|^2}{2\sigma^2}},\ x \in \mathbb{R}^2, \tag{1.3}$$

where the size of the excitatory center can be controlled by the parameter σ. The filter kernel,

[2] $h \in L^1(\mathbb{R}^2)$ results from Young's inequality for convolution.

which is also referred to as *Mexican hat function*[3], and example filter outputs are illustrated in Figure 1.1. The linear system defined by this kernel allowed to explain physiological data from retinal ganglion cells [51].

The receptive fields of cells in the visual cortex are more complicated. Hubel and Wiesel [35] identified three different types of cells: simple cells, complex cells, and hypercomplex cells[4]. The distinction criterion can be found in more detail in [76]. The authors also propose a formal method to identify the cell type.

Simple cells have receptive fields similar to the receptive fields of ganglion cells. But the fundamental difference is that this cell type additionally is selective to the orientation of the stimulus. This results in an increase of selectivity in general. The simple cell behavior, i.e. spatially localized receptive fields which consist of distinct elongated excitatory and inhibitory regions, can be modeled by even- and odd-symmetric Gabor filter kernels [15, 52]

$$h_{even}(V(\phi)^T x) = h_{even}(y) = e^{-\frac{1}{2}y^T\Sigma^{-1}y}\cos(fy_1+\theta),\ x\in\mathbb{R}^2, \tag{1.4}$$

$$\text{and } h_{odd}(V(\phi)^T x) = h_{odd}(y) = e^{-\frac{1}{2}y^T\Sigma^{-1}y}\sin(fy_1+\theta),\ x\in\mathbb{R}^2, \tag{1.5}$$

where

$$V(\phi) = \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix}$$

is the rotation matrix into the rotated y coordinate system. Σ is the covariance matrix which determines the speed of decrease in amplitude, f is the spatial frequency of the trigonometric function, and θ determines the phase shift. The filter kernels in different instantiations and corresponding example filter outputs are illustrated in Figure 1.2. This decomposition in different orientations and spatial frequencies is the standard model of V1. Similar to simple cells, *complex cells* are also orientation selective but they respond independent of the exact position of the presented stimulus. These cells introduce a position invariance in a certain neighborhood. The third cell type, the *hypercomplex cell*, also has an orientation-selective property but it does not respond to elongated stimuli like lines or gratings. These cells respond to end-stopped lines or corners, for example. Complex cells and hypercomplex cells cannot be modeled linear-only anymore. The theoretical justification for not modeling the behavior of these cells by a linear system is given in the following section.

[3] The Mexican hat function can be approximated by the difference of two Gaussian functions [50].

[4] The term *hypercomplex* was replaced by the term *end-stopped* as there is evidence that simple cells exist which have the selectivity property [36].

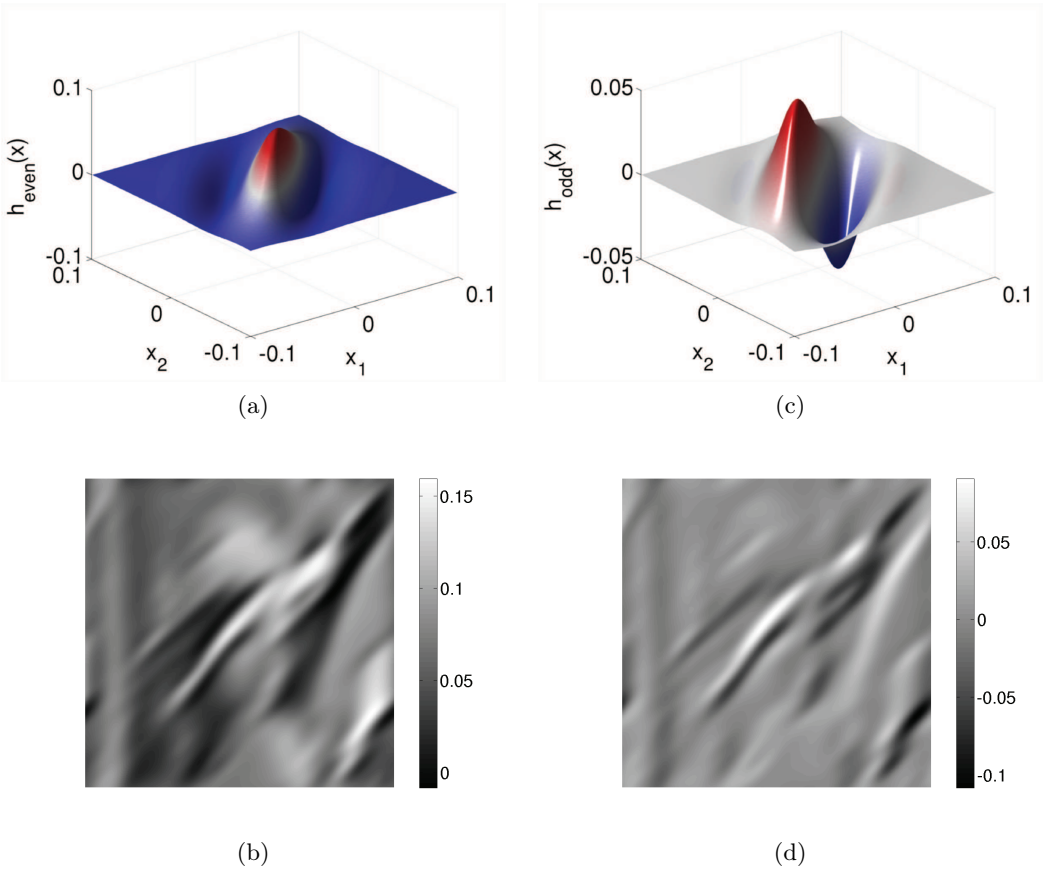


Figure 1.2: The even-symmetric (a) and the odd-symmetric (c) Gabor filter kernels ($\phi = 135°$, $f = 20\pi$, $\Sigma_{11} = 0.025$, $\Sigma_{22} = 0.05$, and $\Sigma_{21} = \Sigma_{12} = 0$) are illustrated. The filter response of the respective filters applied to the image in Figure 1.1(b) are illustrated in (b) for the even-symmetric and in (d) for the odd-symmetric filter kernel.

1.3 The concept of intrinsic dimensionality

The concept of intrinsic dimensionality was developed by Zetzsche and Barth in the early nineties [86, 87]. It connects the dimensions of the input of a signal with the shape of the signal. The intrinsic dimensionality of a signal is defined by the degrees of freedom in the input space which are necessary to determine a constant path of function values within the signal uniquely. In the following, we consider signals with a two-dimensional domain, i.e. images, and their intrinsic dimensionality as defined in the following.

Definition 1.3 (Intrinsic dimensionality)**.** Let u be a signal which is defined by a function $u : \mathbb{R}^2 \to \mathbb{R}$, $u \in L^2(\mathbb{R}^2)$, with a compact support $\Omega \subset \mathbb{R}^2$. The *intrinsic dimensionality id* of

u is then defined by

$$id(u) := \begin{cases} 0 & \text{, if } \forall x \in \Omega: \ u(x) = const., \\ 1 & \text{, if } \exists v \in S^1 \subset \mathbb{R}^2 : \forall z \in \Omega : \forall t \in \mathbb{R} \text{ s.t. } z + tv \in \Omega: \ u(z+tv) = const., \\ 2 & \text{, else,} \end{cases} \tag{1.6}$$

where $S^1 \subset \mathbb{R}^2$ is the two-dimensional unit sphere. A signal belonging to the class of signals which have the same intrinsic dimensionality n is referred to as an inD-signal.

Remark 1.4. The degrees of freedom which are necessary to determine the constant path of function values (amplitude) in the domain of a signal increases with the intrinsic dimensionality.

An $i0D$-signal is a constant function. It can only vary in its overall amplitude. An $i1D$-signal's domain has one characteristic direction which defines the lines in the domain on which the signal amplitude does not vary, i.e. the signal can be written as a function of one variable in an appropriately rotated coordinate system. Typical $i1D$-signals are dirac-lines, oriented sign functions, or oriented two-dimensional sinus functions, for example. The class of $i2D$-signals is the biggest class as it comprises all other possible signals which are not $i0D$ or $i1D$. An $i2D$-signal has no direction of constant amplitude. The amplitude varies in all, i.e. both, dimensions of the domain. Easy examples are corners, crossing dirac-lines, or bounded lines. An overview of examples for $i0D$-, $i1D$-, and $i2D$-signals can be found in Figure 1.3.

In order to analyze local regions in natural images, we need a modified definition of intrinsic dimensionality to be able to characterize these regions within an image.

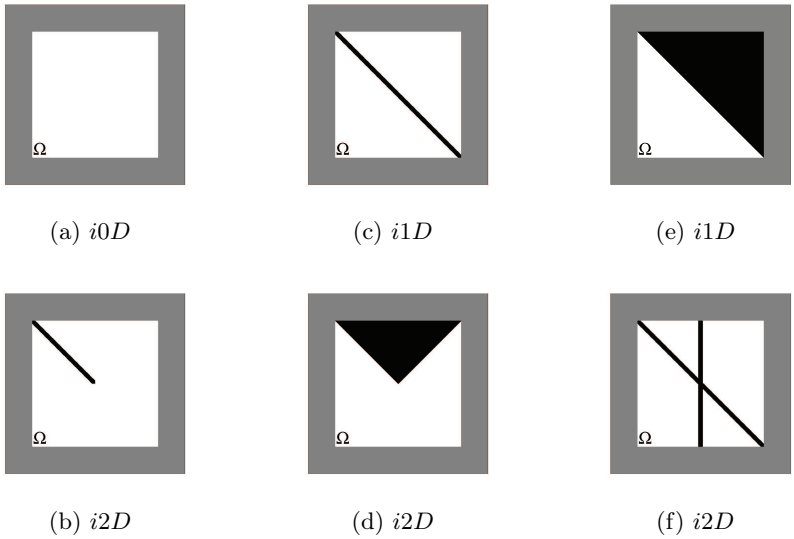
Definition 1.5 (Local intrinsic dimensionality)**.** Let $u \in L^2(\mathbb{R}^2)$ be a signal. Let x_0 be a single point in the domain of u and $\Omega_{x_0} \subset \mathbb{R}^2$ is a compact neighborhood around x_0. The local intrinsic dimensionality id_{loc} of the signal u in the point x_0 is defined by

$$id_{loc}(x_0, u) := id(u_{|\Omega_{x_0}}). \tag{1.7}$$

The point x_0 with $id_{loc}(x_0, u) = n$ is then referred to as an inD-point. The set of all inD-points with respect to the signal u is defined by

$$I_n(u) := \{x \in \mathbb{R}^2 | id_{loc}(x, u) = n\}. \tag{1.8}$$

Many neurons in the early visual cortex, i.e. V1 and V2, exhibit a selectivity for $i2D$-signals. That means they suppress or give reduced responses to $i1D$-signals while responding strongly to $i2D$-signals. Such neurons have been called "hypercomplex" [35], "end-stopped" [57], "dot-responsive" [64], or having "surround suppression" [14]. These reported neurons



(a) $i0D$ (c) $i1D$ (e) $i1D$

(b) $i2D$ (d) $i2D$ (f) $i2D$

Figure 1.3: This figure shows typical signals of the $i0D$-, $i1D$-, and $i2D$-type within the region Ω. A constant $i0D$-signal is illustrated in (a). A single line (c) and an edge (e) are examples for $i1D$-signals. The bottom row shows typical $i2D$-signals like an end-stopped line (b), a corner (d), or crossing lines (f).

share one essential property. None of them reacts to longer straight lines, extended sinusoidal gratings, or any other elongated pattern, i.e. they do not react to $i0D$- and $i1D$-signals. Instead, they respond to $i2D$-signals like spots, corners, line ends, and similar patterns. In order to be able to deal with such $i2D$-selective systems, its formal meaning is clarified in the following definition.

Definition 1.6 ($i2D$-system)**.** Let $T : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ be an operator defining a shift-invariant system which maps the input signal u to a response signal r, i.e. $r = T(u)$. The operator T is an $i2D$-operator, if and only if

$$r(x) = T(u)(x) = 0 \quad , \forall x \in I_0(u) \cup I_1(u), \tag{1.9}$$

for arbitrary signals $u \in L^2(\mathbb{R}^2)$. The system defined by the operator T is then an $i2D$-system. The response r is referred to as an $i2D$-feature.

Remark 1.7. Systems being shift-invariant are often referred to as time-invariant in the signal processing literature. Shift-invariant means invariant with respect to the input argument, i.e. $r(x - \delta x) = T(u(x - \delta x))$ for arbitrary δx and a given $r = T(u)$.

The definition of an $i2D$-system on its own does not give much information about the properties of the system. The following lemma cancels out a specific set of systems which is not able to be an $i2D$-system and concludes this section.

Lemma 1.8. *No shift-invariant linear system $T : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ with an impulse response function unequal to zero can be an $i2D$-system.*

Proof. The statement is proved by contradicting the counter statement: *There exists a shift-invariant linear system T with impulse response function unequal to zero which is an $i2D$-system.*
Let T be a shift-invariant linear operator and u be an arbitrary signal. We then can rewrite every shift-invariant linear system by

$$T(u)(x) = \int_{\mathbb{R}^2} h(x-y)u(y)\ dy = 2\pi\mathcal{F}^{-1}(\mathcal{F}(h)\mathcal{F}(u))(x), \tag{1.10}$$

where $h \in L^1(\mathbb{R}^2)$ is the respective impulse response of the linear system [4] and $\mathcal{F}$ is the Fourier transform operator, see Section 2.2 for a definition. The system has to satisfy the following equation for $i0D$-signals, i.e. $u(x) = k = const.,\ \forall x \in \mathbb{R}^2$.

$$\begin{aligned} 0 = &\int_{\mathbb{R}^2} h(x-y)u(y)\ dy = k\int_{\mathbb{R}^2} h(x-y)\ dy \quad ,\ \forall x \in \mathbb{R}^2 \\ \Leftrightarrow &\int_{\mathbb{R}^2} h(y)\ dy = 0 \\ \Leftrightarrow &\mathcal{F}(h)(0) = 0. \end{aligned} \tag{1.11}$$

The system also must not respond to $i1D$-signals, i.e. signals u for which the following holds: $\exists v \in S^1 : \forall x_0 \in \mathbb{R}^2 : \forall t \in \mathbb{R}$ s.t. $x_0 + tv \in \Omega :\ u(x_0 + tv) = const.$. Let u be an $i1D$-signal with respect to the arbitrary direction $v \in \mathbb{R}^2,\ \|v\|_2 = 1$, and let $n \in \mathbb{R}^2,\ \|n\|_2 = 1,\ n \perp v$. In particular

$$u(x) = e^{i(x\cdot sn)},\ s > 0, \tag{1.12}$$

fulfill the $i1D$-property. The zero response requirement then becomes

$$\begin{aligned} 0 =& \int_{\mathbb{R}^2} h(y)u(x-y)\ dy \\ =& \int_{\mathbb{R}^2} h(y)e^{i((x-y)\cdot sn)}\ dy \\ =& e^{i(x\cdot sn)}\int_{\mathbb{R}^2} h(y)e^{-i(y\cdot sn)}\ dy \\ =& e^{i(x\cdot sn)}2\pi\mathcal{F}(h)(sn),\ \forall n \in S^1,\ \forall s > 0. \end{aligned} \tag{1.13}$$

Equation (1.11) and (1.13) imply $\mathcal{F}(h) = 0$. This contradicts the counter statement.

□

Remark 1.9. The shift-invariance is not a necessary precondition for the previous lemma.

The statement can be proved in an analog way for more general linear systems like

$$T(u)(x) = \int_{\mathbb{R}^2} h(x,y)u(y)\ dy. \tag{1.14}$$

This is non-relevant because shift-invariant systems are considered only within the context of this work.

Given this lemma we can conclude that the behavior of the previously mentioned "hypercomplex", "end-stopped", and "dot-responsive" cells or cells having a "surround suppression" cannot be modeled by a linear system. For this reason more sophisticated models are required which directly leads to the following section.

2 A generalized Gabor approach for $i2D$-feature extraction

What kind of system is able to model neurons which are completely quiet whenever a "boring" stimulus is presented and which are totally excited if a more complex stimulus, like a curved line, an end-stopped line, a corner, etc., is presented? The rising evidence in the neurophysiological literature for nonlinear neurons, which are highly selective to intrinsic two-dimensional features, raises the question for new model approaches being able to describe both reported phenomena. The formalism has to be powerful in such a way that it can describe the reported linear phenomena as well as the highly nonlinear behavior of neurons in early visual cortex. As has already been stated in the early nineties [86] and proved in Section 1.3 (Lemma 1.8), we cannot draw on linear systems anymore. Linear systems alone are not powerful enough to model an $i2D$-selective neuron, respectively an $i2D$-system, cf. Definition 1.6. The linear approach to describe neural behavior in early stages of the visual system is well accepted and it is able to explain a wide range of reported phenomena, cf. Section 1.2. The standard model of linear, frequency-selective mechanisms is a systematic approach by using the formal framework of linear systems theory. Furthermore it provides a low-parametric description by the Gabor filters with center frequency, bandwidth, and orientation. This model gives a clear account of neural selectivity. It has already been attempted to adapt and to modify the linear approach or extend it by some "small" nonlinear operations. All effort was expended to be able to explain phenomena which could not be explained by the linear model. A standard model for these problems has not been established yet so that the primary question is: Does a similar approach exist for nonlinear vision?

In this chapter we consider the simplest nonlinear extension of linear systems, the second-order Volterra-series expansion of a nonlinear system, cf. Section 2.2. The concept of orientation- and spatial frequency-selectivity is applied and adapted to the second-order Volterra-system in Section 2.3 in order to provide a generalized Gabor framework to formulate simple $i2D$-systems corresponding to functionalities reported in early stages of the visual cortex. Furthermore in Section 2.4 multiple approaches including $i2D$-selective operators are considered and extended by a subsequent spatial pooling to extract relevant object features.

2.1 Related Work

The concept of intrinsic two-dimensional features proposed by Zetzsche and Barth [87] can be found in various applications reported in the literature. For example, $i2D$-features are relevant for object recognition as shown in classic experiments of Attneave [2] and by the "Recognition by Components"-theory [7]. $i2D$-features and their respective neurons appear to have a role in the bottom-up control of saccadic eye movements [45, 68]. In natural scenes there is a strong relation between statistical redundancies and $i2D$-features [87, 6, 88]. The

probability of occurrence from $i0D$ to $i2D$ in natural images has a decreasing order [90]. As a result $i2D$-features are highly predictive so that it is possible to reconstruct an image from the mere knowledge of the $i2D$-regions only [6]. The $i2D$-features also can achieve a nonlinear whitening of the higher-order-statistics as expressed by the bi-spectrum of natural images [91]. A generalization to continuous intrinsic dimensionality was considered in [46].

It was also shown that a multiplicative AND-like combination is required to obtain a system which is optimally adapted to the statistics of natural scenes [92]. This AND-like combination which can be interpreted as a multiplicative combination is an essential property of $i2D$-systems as can be seen in the subsequent sections. The relation between higher-order statistics and Volterra systems was considered explicitly in [93]. These findings were also used to develop an optimized coding scheme for natural images [94] and to learn selectivity properties of cortical cells of V2 and V4 [56].

That cortical neurons cannot be modeled by linear systems only, is supported by a variety of neural findings. The cortical gain control [13] as a normalization of the output of a linear unit is one example for the adaptation of the linear system approach to be able to explain the behavioral findings of neurons. Another example is the complex cell found in early visual cortex. For this cell type it is argued that in comparison to the linear simple cell the complex cell has a phase invariance [12], i.e. it responds independently of the exact location of the stimulus [41]. This phase invariance also cannot be modeled by a linear model. Therefore, it has to be extended by a nonlinear mechanism, too. These are two examples of found phenomena which can be explained by slight adaptations of the linear approach. But cells which cannot be explained by the "semi-linear" approach have also been reported in the literature. In visual cortex "hypercomplex" [35], "end-stopped" [57], and "dot-responsive" [64] cells have been found. Even on the frog's retina these highly nonlinear cells have been found in the form of a "bug-detector" [49]. More recent findings also give evidence for $i2D$-selectivity implemented by the neural hardware.

In [14] the authors investigated the influence between the center and the surround of the classical receptive fields and it turned out that the recorded cells in V1 show a divisive surround suppressed behavior similar to the mechanisms of cortical gain control. In [74] it is reported that the suppression depends on the stimulus orientation presented to the receptive field and it has its maximum for same orientations. In terms of intrinsic dimensionality this means that $i1D$-signals are maximally suppressed. In V1 cells were found which respond to stimuli with differently oriented gratings in the center and the surround [75]. But these cells do not respond anymore if the center orientation and surround orientation is equal. Furthermore, cells were found which prefer a specific angle between the two orientations. This is also supported by similar findings in the cat's striate cortex [73, 30]. The selectivity to specific opening angles of oriented corners was also reported in [37]. In [72] the oriented gratings in the center and in the surround were varied with respect to different properties like luminance, contrast, color,

or orientation. The discontinuity between the center region and the surround region could cause an response of an appropriately selective neuron. The authors reported no responses in all cases where the center and the surround share the same properties, i.e. no response to $i1D$-signals. But the cells fired whenever a property of the center and the surround differed. A similar behavior was also reported in a study regarding discontinuities in presented stimuli [69].

2.2 Mathematical preliminaries

In this chapter we review the most important mathematical definitions and equations which are relevant for the further analysis and synthesis of nonlinear systems. The first important definition is the Volterra series of a nonlinear system. The definition for signals with one-dimensional arguments and further results regarding Volterra systems can be found in the book by Schetzen [65]. As the main focus in this work is vision, his definition is extended to signals with two-dimensional arguments.

Definition 2.1 (Volterra system)**.** Let T be a shift-invariant and continuous operator which maps the input signal $u : \mathbb{R}^2 \to \mathbb{R}$ to an output signal $r : \mathbb{R}^2 \to \mathbb{R}$. Under certain conditions (cf. following remark), it can be shown that the relation between the output and the input can be expressed by

$$\begin{aligned} T(u)(x) = \quad & \int_{\mathbb{R}^2} h_1(y_1)u(x-y_1)\ dy_1 + \int_{\mathbb{R}^4} h_2(y_1,y_2)u(x-y_1)u(x-y_2)\ d(y_1,y_2) \\ & + \int_{\mathbb{R}^{2n}} h_n(y_1,\ldots,y_n)\prod_{i=1}^{n} u(x-y_i)\ d(y_1,\ldots,y_n) + \ldots \end{aligned} \tag{2.1}$$

where the functions h_n are elements of the corresponding $L^1(\mathbb{R}^{2n})$ such that the integrals exist[5]. This functional series is referred to as *Volterra series* and the functions h_n are called *Volterra kernels* of the system. An equivalent expression of the Volterra series is the operator series

$$T(u)(x) = H_1(u)(x) + H_2(u)(x) + \ldots + H_n(u)(x) + \ldots \tag{2.2}$$

with

$$H_n(u)(x) := \int_{\mathbb{R}^{2n}} h_n(y_1,\ldots,y_n)\Pi_{i=1}^{n}u(x-y_i)\ d(y_1,\ldots,y_n). \tag{2.3}$$

The operator H_n is called *nth-order Volterra operator*. A system which can be represented by a finite number of Volterra operators with a maximum order of n is an *nth-order Volterra system*.

[5]This follows from Young's inequality and $\|u(\bullet)u(\bullet)\|^2_{L^2(\mathbb{R}^4)} = \int_{\mathbb{R}^4} |u(x)u(y)|^2(x,y) \leq \|u\|^2_{L^2(\mathbb{R}^2)}\|u\|^2_{L^2(\mathbb{R}^2)} \leq \infty$ for $u \in L^2(\mathbb{R}^2)$.

Remark 2.2. It was shown by Brilliant [10] that any continuous nonlinear system can be approximated sufficiently well by the Taylor expansion if the input signal lies in $L^2(\Omega)$ where Ω is a compact subset of $\mathbb{R}^2$.

Within this work the second-order Volterra kernel and its spectral representation is of major interest. In order to be able to design a nonlinear system in Fourier space, the 4-dimensional Fourier transformation is required. In the following the n-dimensional Fourier transformation is defined. Further information can be found in standard signal processing literature [77].

Definition 2.3 (n-dimensional Fourier transformation)**.** Let $f \in L^1(\mathbb{R}^n)$ be an integrable function. The Fourier transformation of f is defined by

$$\mathcal{F}(f)(x) := \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\mathbb{R}^n} f(z) e^{-ix\cdot z}\, dz. \tag{2.4}$$

The following theorem is a generalization to n-dimensional integration domains of the substitution rule for integrals. It is an important tool to obtain the parametrization by orientation of the nonlinear system kernel. A proof can be found in standard calculus textbooks like [33].

Theorem 2.4 (Transformation theorem)**.** *Let $\Omega \subset \mathbb{R}^n$ be an open subset and let $\Phi : \Omega \to \Phi(\Omega)$ be a diffeomorphism. Then f is integrable on $\Phi(\Omega)$ if and only if the function $x \mapsto f(\Phi(x))|\det(D\Phi(x))|$ is integrable on Ω. It also holds*

$$\int_{\Phi(\Omega)} f(y)\, dy = \int_{\Omega} f(\Phi(x))|\det(D\Phi(x)|\, dx. \tag{2.5}$$

2.3 Generalized Gabor to obtain second-order selectivity

This section presents a generalized Gabor approach within the framework of nonlinear systems and in line with the concept of intrinsic dimensionality. First, the classical Gabor approach for linear systems is reviewed briefly. Second, the theoretical framework of second-order Volterra systems is prepared such that thirdly common signals can be analyzed with respect to their nonlinear representation. Finally, the generalized Gabor approach is developed and tested for various parameter settings.

2.3.1 Classical Gabor approach

The main motivation for the filter design based on second-order Volterra systems is the classical Gabor filter approach. It turned out that this approach works well to describe the behavior of simple cells in the early visual cortex, e.g. the cat's striate cortex [39]. The classical approach and its spectral representation is considered in the following. Subsequently this approach is transfered to second-order Volterra systems to provide a framework which can be parametrized low-dimensionally. The main goal is a low-parametrized approach which

is able to explain the i2D-selectivity of reported neurons in early visual cortex. The following definition includes the filter kernel presented in Section 1.2.

Definition 2.5 (Classical Gabor filter kernel)**.** Given the rotation matrix $V(\phi)$ into the coordinate system y rotated by the angle $\phi \in [0, 2\pi)$, i.e. $y = V(\phi)^T x$ with

$$V(\phi) = \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix}, \tag{2.6}$$

and the covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}, \tag{2.7}$$

the classical *Gabor filter* kernels, i.e. even-symmetric and odd-symmetric, are defined by the product of a Gaussian function and a trigonometric function, i.e.

$$h_{even}(x) = e^{-\frac{1}{2} x^T V(\phi) \Sigma^{-1} V(\phi)^T x} \cos(f_r(x_1 \cos(\phi) + x_2 \sin(\phi))), \tag{2.8}$$

$$\text{and } h_{odd}(x) = e^{-\frac{1}{2} x^T V(\phi) \Sigma^{-1} V(\phi)^T x} \sin(f_r(x_1 \cos(\phi) + x_2 \sin(\phi))), \tag{2.9}$$

where $f_r \geq 0$ defines the spatial frequency of the trigonometric function.

Remark 2.6. The filter kernels are completely determined by the parameters ϕ, f_r, σ_1, and σ_2. ϕ determines the direction of the two-dimensional sinus- or cosine-wave and it determines the orientation of the elliptic shaped Gaussian function. f_r determines the frequency of the trigonometric function in the direction of the first dimension of the rotated coordinate system. σ_1 determines the semi-axis of the elliptically-shaped Gaussian function in the first direction of the rotated coordinate system, i.e. in the direction of the sinus- or cosine-wave. σ_2 determines the semi-axis in the second direction of the rotated coordinate system, i.e. in the direction in which the trigonometric function is constant. In case the semi-axes of the Gaussian do not correspond to the direction of the trigonometric wave function, the minor diagonal of Σ is then unequal to zero.

The filter kernels are defined in state space which has two disadvantages. The first one is that it is not obvious whether a filter decomposition of an input signal, i.e. an image, in different filter channels is a complete representation of the signal. The second disadvantage is that the filter operation implemented by a convolution operation is computationally expensive compared to the solution by the Fourier transform. To overcome these disadvantages the Fourier transform of the Gabor functions is provided in the following lemma. With the aid of these kernels the system can be described in spectral space.

Lemma 2.7. *Given the Gabor filter kernels defined in Definition 2.5, the corresponding Fourier transforms are*

$$\begin{aligned} &H_{even}(f) \\ =&\frac{1}{4\pi|\det(\Sigma^{-1/2})|}\left(e^{-\frac{1}{2}(f-f_r v)^T V(\phi)\Sigma V(\phi)^T(f-f_r v)}+e^{-\frac{1}{2}(f+f_r v)^T V(\phi)\Sigma V(\phi)^T(f+f_r v)}\right) \end{aligned} \tag{2.10}$$

and

$$\begin{aligned} &H_{odd}(f) \\ =&\frac{1}{4\pi|\det(\Sigma^{-1/2})|}i\left(e^{-\frac{1}{2}(f-f_r v)^T V(\phi)\Sigma V(\phi)^T(f-f_r v)}-e^{-\frac{1}{2}(f+f_r v)^T V(\phi)\Sigma V(\phi)^T(f+f_r v)}\right) \end{aligned} \tag{2.11}$$

where $v := (\cos(\phi), \sin(\phi))^T$.

Proof. The Fourier transform of the product of two functions f and g can be rewritten by

$$\mathcal{F}(fg) = \frac{1}{2\pi}\mathcal{F}(f) * \mathcal{F}(g). \tag{2.12}$$

First, we start with the Fourier transform of the Gaussian function

$$g(x) := e^{-\frac{1}{2}x^T V(\phi)\Sigma^{-1}V(\phi)^T x}. \tag{2.13}$$

In the following we write just V instead of $V(\phi)$ for convenience.

$$\begin{aligned} \mathcal{F}(g)(x) &= \frac{1}{2\pi}\int_{\mathbb{R}^2} e^{-\frac{1}{2}z^T V\Sigma^{-1}V^T z}e^{-i(x\cdot z)}\,dz \\ &= \frac{1}{2\pi}\int_{\mathbb{R}^2} e^{-\frac{1}{2}z^T V\Sigma^{-1}V^T z - i(x\cdot z)}\,dz. \end{aligned} \tag{2.14}$$

Substituting by $\Phi(z) := \Sigma^{-1/2}V^T z$ with $D\Phi(z) = V\Sigma^{-1/2}$ and $\det(D\Phi(z)) = \det(\Sigma^{-1/2})$ yields

$$\begin{aligned} \mathcal{F}(g)(x) &= \frac{1}{2\pi}\frac{1}{|\det(D\Phi)|}\int_{\mathbb{R}^2} e^{-\frac{1}{2}\Phi(z)^T\Phi(z) - i(x\cdot V\Sigma^{1/2}\Phi(z))}|\det(D\Phi)|\,dz \\ &= \frac{1}{2\pi}\frac{1}{|\det(\Sigma^{-1/2})|}\int_{\mathbb{R}^2} e^{-\frac{1}{2}y^T y - i(x\cdot V\Sigma^{1/2}y)}\,dy. \end{aligned} \tag{2.15}$$

Note that $det(V) = 1$ by definition. By doing a quadratic expansion, the exponent can be

rewritten by ($\langle \bullet, \bullet \rangle$ denotes the standard scalar product)

$$\begin{aligned} & -\frac{1}{2}\left(\langle y, y\rangle + i2\langle x, V\Sigma^{1/2}y\rangle\right) \\ =& -\frac{1}{2}\left(\langle y, y\rangle + 2\langle i\Sigma^{1/2}V^T x, y\rangle\right) \\ =& -\frac{1}{2}\left(\langle y + i\Sigma^{1/2}V^T x, y + i\Sigma^{1/2}V^T x\rangle + \langle \Sigma^{1/2}V^T x, \Sigma^{1/2}V^T x\rangle\right). \end{aligned} \tag{2.16}$$

We thus get

$$\mathcal{F}(g)(x) = \frac{1}{2\pi}\frac{1}{|\det(\Sigma^{-1/2})|}e^{-\frac{1}{2}x^T V\Sigma V^T x}\int_{\mathbb{R}^2} e^{-\frac{1}{2}(y+i\Sigma^{1/2}V^T x)^T(y+i\Sigma^{1/2}V^T x)}\, dy. \tag{2.17}$$

Substituting by $\Theta(y) := y + i\Sigma^{1/2}V^T x$ with $D\Theta(y) = Id$ and $\det(D\Theta(y)) = 1$ yields

$$\begin{aligned} \mathcal{F}(g)(x) &= \frac{1}{2\pi}\frac{1}{|\det(\Sigma^{-1/2})|}e^{-\frac{1}{2}x^T V\Sigma V^T x}\underbrace{\int_{\mathbb{R}^2} e^{-\frac{1}{2}w^T w}\, dw}_{=2\pi} \\ &= \frac{1}{|\det(\Sigma^{-1/2})|}e^{-\frac{1}{2}x^T V\Sigma V^T x}. \end{aligned} \tag{2.18}$$

The Fourier transform of the function $c(x) := \cos(f_r(v \cdot x))$ is derived by using the cosine represented by complex exponential functions. This yields

$$\begin{aligned} \mathcal{F}(c)(x) &= \frac{1}{2\pi}\int_{\mathbb{R}^2} \cos(f_r(z \cdot v))e^{-i(x\cdot z)}\, dz \\ &= \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{1}{2}(e^{if_r(v\cdot z)} + e^{-if_r(v\cdot z)})e^{-i(x\cdot z)}\, dz \\ &= \frac{1}{4\pi}\int_{\mathbb{R}^2} e^{-i(x-f_r v)\cdot z} + e^{-i(x+f_r v)\cdot z}\, dz \\ &= \frac{1}{2}(\delta(x - f_r v) + \delta(x + f_r v)), \end{aligned} \tag{2.19}$$

where δ is the δ-distribution. Analogously the Fourier transform of the sinus function $s(x)$ becomes

$$\mathcal{F}(s)(x) = \frac{1}{2}i(\delta(x - f_r v) - \delta(x + f_r v)). \tag{2.20}$$

The convolution of the respective functions yields the assumption.

□

Remark 2.8. Note that the covariance matrix of the Gaussian function of the Fourier transformed filter kernels is not Σ itself. The missing inversion in the formula has to be taken into

account such that the covariance matrix in Fourier representation is

$$\Sigma_f = \begin{pmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_\phi^2 \end{pmatrix} = \Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix}. \tag{2.21}$$

σ_r determines the semi-axis of the Gaussian function in radial direction and is related to σ_1 by $\sigma_r = \frac{1}{\sigma_1}$. $\sigma_\phi = \frac{1}{\sigma_2}$ determines the semi-axis which is orthogonal to the radial direction. This direction can be approximately interpreted as an angular direction. But it is not the angular direction which causes problems if one tries to find a perfect partition of unity by Gabor filter functions. Other approaches using "Gabor-like" filters have been developed to solve this problem [82, 94]. They thus can be applied to image encoding without causing any significant distortions.

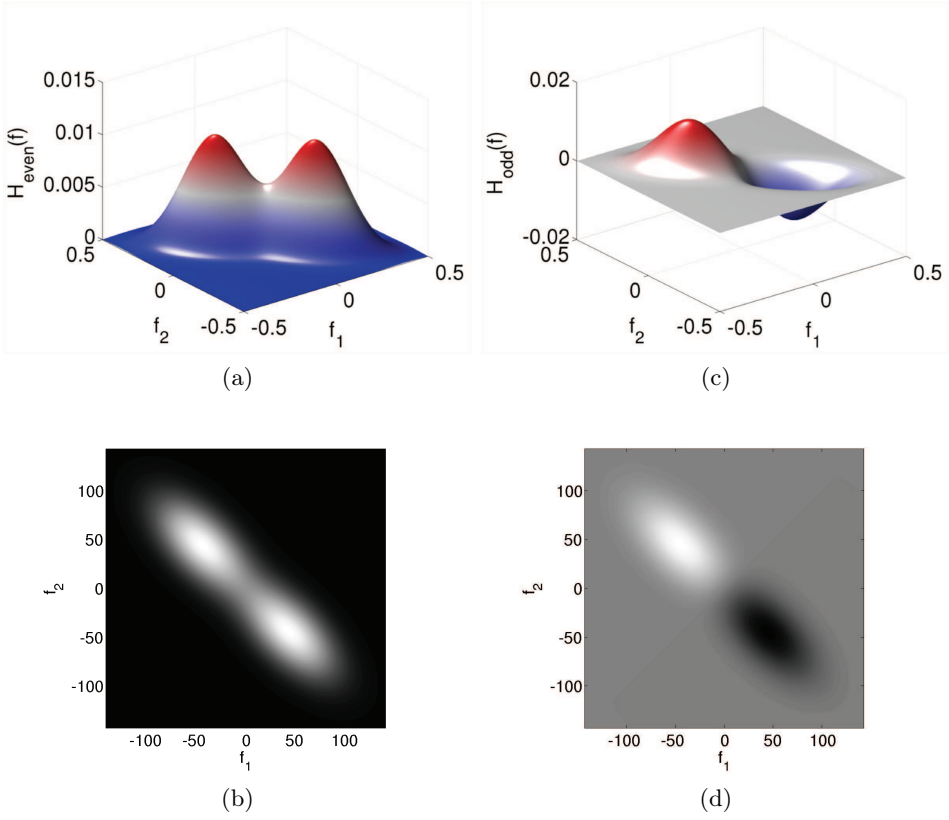


Figure 2.1: The Gabor filter kernels of Figure 1.2 are illustrated in Fourier space. The even-symmetric filter kernel in (a) and (b) has positive amplitude only. The odd-symmetric filter function in (c) and (d) is point-symmetric with respect to the origin. Both filter functions are Gaussian functions shifted to the positions $(f\cos(\phi), \sin(\phi))^T$ and $-(f\cos(\phi), \sin(\phi))^T$. The standard deviations of the Gaussian functions which characterize the elliptic shape are determined by Equation (2.21).

Remark 2.9. One common approach in the literature for the Gabor filter is edge detection [53]. Note that in the proposed parametrization of filter kernels an oriented edge with orientation θ is not detected by a Gabor filter with the same orientation. The Gabor filter which detects the oriented edge with orientation θ must have the orientation $\phi = \theta - \frac{\pi}{2}$. Odd-symmetric kernels are suitable to detect edges with a real step in its function value, like the sign function. The even symmetric kernels are appropriate to detect lines.

The filter kernels are bandpass filters as can be seen in their Fourier transform in Equations (2.10) and (2.11) and as illustrated in Figure 2.1. In summary the passband is defined by the parameters

- orientation of the angular center frequency ϕ,
- radial center frequency f_r,
- bandwidth in radial direction determined by σ_r, where the full width at half maximum is given by $2\sqrt{2\ln(2)}\sigma_r$), and
- bandwidth in direction orthogonal to the radial direction determined by σ_ϕ, where the full width at half maximum is given by $2\sqrt{2\ln(2)}\sigma_r$.

In Section 1.3 it is already stated that the abilities of linear systems defined by an impulse response or its Fourier transform are strongly limited. An $i2D$-system cannot be realized by this approach. We thus make use of second-order Volterra systems in the following.

2.3.2 Second-order Volterra system

A second-order Volterra system by Definition 2.1 is the sum of a linear first-order Volterra-operator and the second-order Volterra operator. As linear systems have been studied intensely in the past, we restrict the system of interest to the nonlinear system defined by

$$\begin{aligned} T(u)(x) =& \int_{\mathbb{R}^4} h(\tilde{x}_1, \tilde{x}_2) \underbrace{u(x-\tilde{x}_1)u(x-\tilde{x}_2)}_{:=g(x-\tilde{x}_1, x-\tilde{x}_2)} d(\tilde{x}_1, \tilde{x}_2) = (h * g)((x,x)^T) \\ =& (2\pi)^2 \mathcal{F}^{-1}(\underbrace{\mathcal{F}(h)}_{=:H}\underbrace{\mathcal{F}(g)}_{=:G})((x,x)^T) \end{aligned} \tag{2.22}$$

with the input signal $u \in L^2(\mathbb{R}^2)$ and the second-order kernel $h \in L^1(\mathbb{R}^4)$, which guarantees the existence of the integral. As can be seen in the previous equation, by defining the function $g \in L^2(\mathbb{R}^4)$ the system can be interpreted as a linear system of signals with a four-dimensional domain. This means that results from multi-dimensional linear systems theory are applicable. Before a generalized Gabor approach is applied to this kind of system, it has to be clarified how specific signals are represented in this nonlinear fashion. Especially $i0D$ and $i1D$ signals must

be analyzed to identify possible stop-bands in the spectral representation of h. In the following a specific parametrization is applied to the Fourier transformation of g and subsequently its Fourier transform G is investigated for various input signals u.

We introduce the parametrization $\phi(t) := x_0 + t_1 v_1 + t_2 v_2,\ x_0, v_1, v_2, t \in \mathbb{R}^2, \|v_1\| = \|v_2\| = 1,\ |\langle v_1, v_2 \rangle| \neq 1$, and $t = (t_1, t_2)^T$. Later investigations and the differentiation between $i0D$, $i1D$, and $i2D$ signals are easier with this parametrization as the specific directions v_1 and v_2 are explicit. For convenience in writing, the functions $\tilde{u}$ and $\tilde{g}$ are defined within the following equation

$$\begin{aligned} g(x,y) =& u(x)u(y) \\ =& u(\phi(t^x))u(\phi(t^y)) =: \tilde{u}(t^x)\tilde{u}(t^y) \\ =:& \tilde{g}(t^x, t^y) \end{aligned} \tag{2.23}$$

with $x,\ y,\ t^x,\ t^y \in \mathbb{R}^2$. Thus, $\tilde{u}$ is assigned with two directions v_1 and v_2 and an origin x_0. This representation has the advantage that the definition of edges, corners, and crossing lines with specific direction can be realized easily. Applying the definition of the n-dimensional Fourier transform (cf. Definition 2.3) yields

$$\mathcal{F}(g)(z_1, z_2) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^4} g(x,y) e^{-i(z_1 \cdot x + z_2 \cdot y)}\ d(x,y) \tag{2.24}$$

where $z = (z_1^T, z_2^T)^T$ with $z_1,\ z_2 \in \mathbb{R}^2$. We apply the coordinate transformation $\Phi(t) := (\phi(t^x), \phi(t^y))^T$, $t = (t^x, t^y)^T$, with

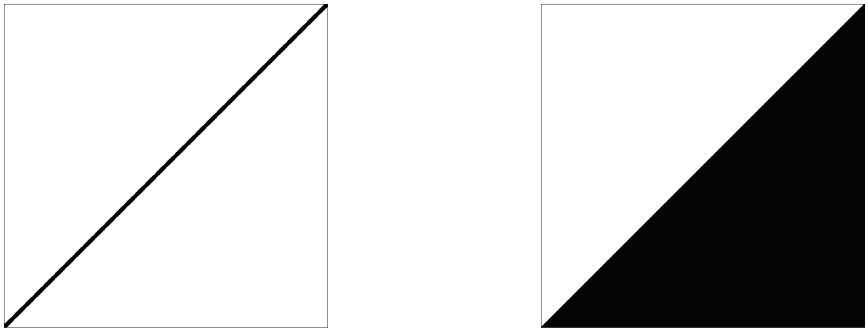
$$D\Phi(t)^T = \begin{bmatrix} v_1 & v_2 & 0 & 0 \\ 0 & 0 & v_1 & v_2 \end{bmatrix} \quad ,\ |\det(D\Phi)| =: k \overset{!}{\neq} 0\ . \tag{2.25}$$

With Theorem 2.4 and Equation (2.23) follows

$$\begin{aligned} \mathcal{F}(g)(z) =& \frac{k}{(2\pi)^2} \int_{\mathbb{R}^4} \tilde{g}(t^x, t^y) e^{-i(z_1 \cdot \phi(t^x) + z_2 \cdot \phi(t^y))}\ d(t^x, t^y) \\ =& \frac{k}{(2\pi)^2} \underbrace{\int_{\mathbb{R}^2} \tilde{u}(t^x) e^{-i z_1 \cdot \phi(t^x)}\ dt^x}_{=: S(\tilde{u})(z_1)} \int_{\mathbb{R}^2} \tilde{u}(t^y) e^{-i z_2 \cdot \phi(t^y)}\ dt^y. \end{aligned} \tag{2.26}$$

For the further analysis of signals the integral defined by S has to be determined for each specific input signal. By using the definition of ϕ the operator S becomes

$$S(\tilde{u})(z) = \int_{\mathbb{R}^2} \tilde{u}(t) e^{-iz \cdot x_0} e^{-iz \cdot (t_1 v_1 + t_2 v_2)}\ dt$$



(a) (b)

Figure 2.2: A discretized δ-line (a) and sign function (b) are illustrated as typical $i1D$-signal examples.

$$= e^{-iz\cdot x_0} \int_{\mathbb{R}^2} \tilde{u}(t) e^{-iz\cdot(Vt)}\, dt \quad , \; V := [v_1|v_2]. \tag{2.27}$$

In order to determine possible pass-bands and stop-bands in the four-dimensional frequency space in which the filters are defined, common signal types are analyzed with respect to their Fourier transform G.

2.3.3 Analysis of selected $i0D$, $i1D$, and $i2D$ signals

The first signal type which is considered is the **$i0D$-type**, i.e. the constant function $\tilde{u}(t) = 1, \; \forall t \in \mathbb{R}^2$. Without loss of generality let V be the identity and $x_0 = 0$. It thus follows

$$S(\tilde{u})(z) = 2\pi \mathcal{F}(1)(z) = \delta(z). \tag{2.28}$$

Note that this $\mathcal{F}$ is the two-dimensional Fourier transform. Inserting in Equation (2.26) yields

$$\mathcal{F}(g)(z) = \delta(z_1)\delta(z_2) = \begin{cases} \infty & , \; z_1 = z_2 = 0 \\ 0 & , \; \text{else.} \end{cases} \tag{2.29}$$

From this equation it can be concluded that the support of all $i0D$-functions in the frequency domain is $M_0 := \{0\} \subset \mathbb{R}^4$.

The second signal type is the **$i1D$-type**. Here two cases are distinguished. The first case is a simple line, i.e. $\tilde{u}(t) = \delta(t_2), \; \forall t \in \mathbb{R}^2$. This describes the one-dimensional δ-line in v_1-direction. A discretized example is illustrated in Figure 2.2. With this definition of $\tilde{u}$ it

can be obtained that

$$\begin{aligned} S(\tilde{u})(z) =& e^{-iz\cdot x_0} \int_{\mathbb{R}} \int_{\mathbb{R}} \delta(t_2) e^{-iz\cdot(Vt)} \, dt_2 \, dt_1 \\ =& e^{-iz\cdot x_0} \int_{\mathbb{R}} e^{-i(z\cdot v_1)t_1} dt_1 \\ =& e^{-iz\cdot x_0} (2\pi)^{\frac{1}{2}} \mathcal{F}(1)(z\cdot v_1) \\ =& e^{-iz\cdot x_0} \delta(z\cdot v_1). \end{aligned} \tag{2.30}$$

Inserting in Equation (2.26) yields

$$\mathcal{F}(g)(z) = k e^{-iz_1\cdot x_0} e^{-iz_2\cdot x_0} \delta(z_1\cdot v_1)\delta(z_2\cdot v_1) = \begin{cases} \neq 0 & , \ \langle z_1, v_1\rangle = 0 \wedge \langle z_2, v_1\rangle = 0 \\ 0 & , \text{ else.} \end{cases} \tag{2.31}$$

This holds for arbitrary directions $v_1 = v$.

The second case of $i1D$-type signals is an edge, which means that the signal is constant except for a jump in its function value. It is defined by $\tilde{u}(t) = \frac{1}{2}(1 + sign(t_2))$. An illustration of the signal can be found in Figure 2.2. It thus follows

$$\begin{aligned} S(\tilde{u})(z) =& e^{-iz\cdot x_0} \int_{\mathbb{R}^2} \frac{1}{2}(1 + sign(t_2)) e^{-iz\cdot(Vt)} \, dt \\ =& e^{-iz\cdot x_0} \frac{1}{2} \left(\int_{\mathbb{R}^2} e^{-iz\cdot(Vt)} \, dt + \int_{\mathbb{R}} e^{-i(z\cdot v_1)t_1} \int_{\mathbb{R}} sign(t_2) e^{-i(z\cdot v_2)t_2} \, dt_2 \, dt_1 \right) \\ =& e^{-iz\cdot x_0} \frac{1}{2} \left(\frac{2\pi}{|\det(V)|} \delta(z) + \int_{\mathbb{R}} e^{-i(z\cdot v_1)t_1} \frac{(2\pi)^{1/2}}{i\pi(z\cdot v_2)} \, dt_1 \right) \\ =& e^{-iz\cdot x_0} \pi \left(\frac{1}{|\det(V)|} \delta(z) + \delta(z\cdot v_1) \frac{1}{i\pi(z\cdot v_2)} \right). \end{aligned} \tag{2.32}$$

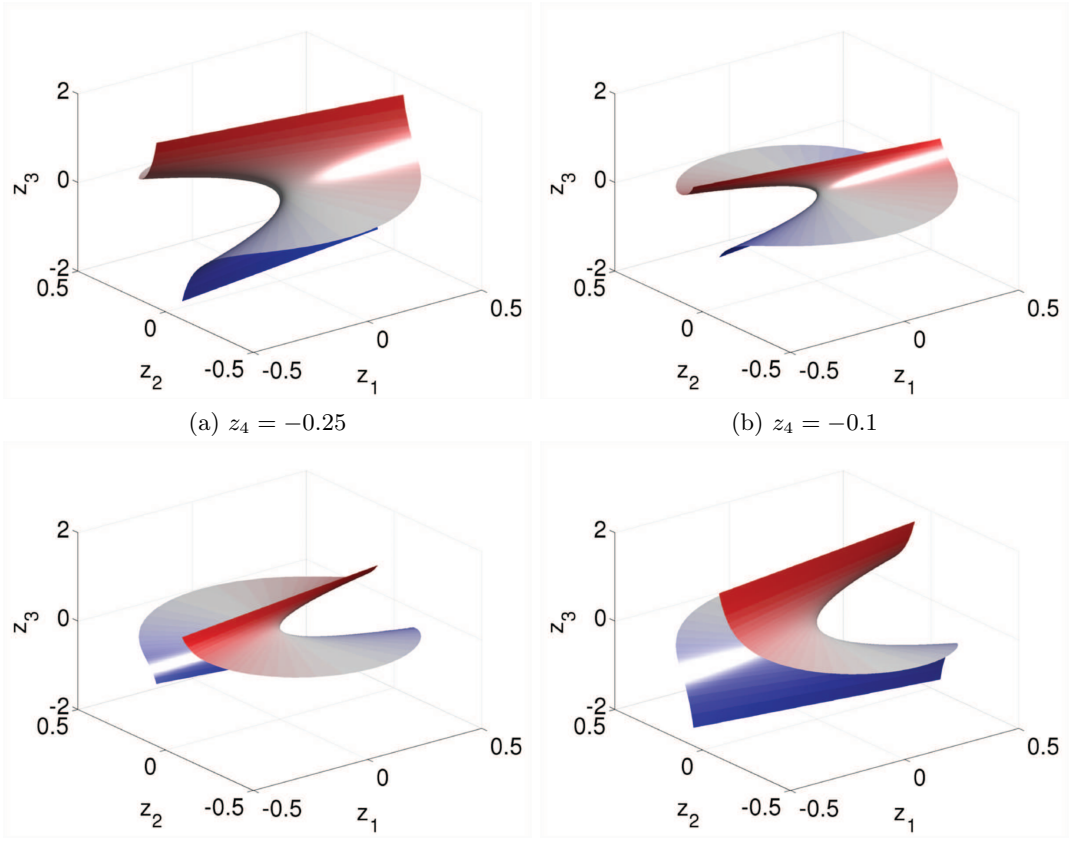
From this equation the same qualitative Fourier transform as in the first case can be derived, i.e.

$$\mathcal{F}(g)(z) = \begin{cases} \neq 0 & , \ \langle z_1, v_1\rangle = 0 \wedge \langle z_2, v_1\rangle = 0 \\ 0 & , \text{ else.} \end{cases} \tag{2.33}$$

In both cases the Fourier transform of the second-order function g has entries unequal to zero in the following two-dimensional planes of the $\mathbb{R}^4$ caused by the $i1D$ signals. Let $n \in \mathbb{R}^2$ be the unit normal vector to $v \in \mathbb{R}^2$. We define the plane $T_1 \subset \mathbb{R}^4$ for each $n = (\cos(\psi), \sin(\psi))^T$, $\psi \in [0, \pi]$ by

$$T_1(n) : z = s_1 \begin{pmatrix} n \\ 0 \end{pmatrix} + s_2 \begin{pmatrix} 0 \\ n \end{pmatrix} \quad , \ \forall s_1, s_2 \in \mathbb{R}. \tag{2.34}$$

The subset $M_1 := \{z \in \mathbb{R}^4 | \exists \psi \in [0, \pi] : n = (\cos(\psi), \sin(\psi))^T \wedge z \in T_1(n)\}$ of $\mathbb{R}^4$ describes the three-dimensional subset which comprises the support of all possible $i1D$-signals in the frequency domain. With $M_0 \subset M_1$ it also contains the $i0D$-functions. The Fourier transform of a second-order Volterra kernel defining an $i1D$-selective system, which is not affected by a $i0D$-signal, must be supported on $M_1 \setminus M_0$ in the frequency domain. More importantly, the support of an $i2D$-selective system must exclude M_1 completely. The three-dimensional set M_1 is illustrated in Figure 2.3. The following theorem is motivated by the previous considerations and states in which case a second-order Volterra system is an $i2D$-system.



(a) $z_4 = -0.25$ (b) $z_4 = -0.1$

(c) $z_4 = 0.1$ (d) $z_4 = 0.25$

Figure 2.3: In this figure the forbidden region M defined in Equation (2.36) is illustrated in Cartesian coordinates for various fixed z_4 (a)-(d). For $z_4 = 0$ the whole $z_1 - z_2$ plane belongs to the forbidden region.

Theorem 2.10 (Second-order Volterra $i2D$-system)**.** *Let T be a second-order Volterra system of the form*

$$T(u)(x) = \int_{\mathbb{R}^4} h(\tilde{x}_1, \tilde{x}_2) \underbrace{u(x - \tilde{x}_1) u(x - \tilde{x}_2)}_{:= g(x - \tilde{x}_1, x - \tilde{x}_2)} \, d(\tilde{x}_1, \tilde{x}_2)$$

$$=(2\pi)^2\mathcal{F}^{-1}(\underbrace{\mathcal{F}(h)}_{=:H}\underbrace{\mathcal{F}(g)}_{=:G})((x,x)^T) \tag{2.35}$$

and let the set M be given by

$$M=\left\{z=s_1\begin{pmatrix}n(\phi)\\0\end{pmatrix}+s_2\begin{pmatrix}0\\n(\phi)\end{pmatrix}\in\mathbb{R}^4\,\middle|\,n(\phi)=\begin{pmatrix}\cos(\phi)\\\sin(\phi)\end{pmatrix},\ s_1,s_2\in\mathbb{R},\ \text{and } \phi\in[0,\pi]\right\}. \tag{2.36}$$

Then T is an i2D-system if and only if $H(z)=0$ for all $z\in M$.

Proof. Let $u\in L^2(\mathbb{R}^2)$ be a signal with the set of $i0D$- and $i1D$-points $I_0(u)\cup I_1(u)$. Let $x_0\in I_0(u)\cup I_1(u)$ with respect to the neighborhood Ω_{x_0} and the direction $v\in S^1$ with $v=(\cos(\phi),\sin(\phi))^T$ for a given angle ϕ. As $x_0\in I_0(u)\cup I_1(u)$ we can rewrite u by

$$u(x)=u(x_0+tv+sn)=\tilde{u}(t,s)=f(s) \tag{2.37}$$

with an appropriate function f in the direction of n for all $s,t\in\mathbb{R}$ such that $x_0+tv+sn\in\Omega_{x_0}$. Without loss of generality let $x_0=0$ and $t\in[-a,a]$. The Fourier transform of u thus becomes

$$\begin{aligned}\mathcal{F}(u)(z)&=\frac{1}{2\pi}\int_{\mathbb{R}^2}u(x)e^{-i(x\cdot z)}\,dx\\&=\frac{1}{2\pi}\int_{\mathbb{R}^2}\tilde{u}(t,s)e^{-i(v\cdot z)t-i(n\cdot z)s}\,dt\,ds\\&=\frac{1}{2\pi}\int_{\mathbb{R}}\int_{[-a,a]}e^{-i(v\cdot z)t}\,dt\;f(s)e^{-i(n\cdot z)s}\,ds\\&=\frac{1}{(2\pi)^{1/2}}\frac{2\sin((v\cdot z)a)}{(v\cdot z)}\mathcal{F}(f)(n\cdot z).\end{aligned} \tag{2.38}$$

For $a\to\infty$ the Fourier transform becomes

$$\mathcal{F}(u)(z)=\delta(v\cdot z)\mathcal{F}(f)(n\cdot z). \tag{2.39}$$

With $x_0=0$ the operator T becomes

$$T(u)(x_0)=(2\pi)^2\int_{\mathbb{R}^4}H(z_1,z_2)\underbrace{\delta(v\cdot z_1)\mathcal{F}(f)(n\cdot z_1)\delta(v\cdot z_2)\mathcal{F}(f)(n\cdot z_2)}_{=(*)}\,dz_1\,dz_2. \tag{2.40}$$

The support of $(*)$ is a subset of M. Thus, the integral becomes zero if and only if $H(z_1,z_2)=0$ for $(z_1,z_2)^T\in M$. This holds for arbitrary direction angles ϕ which concludes the proof. □

Remark 2.11. Note that the limit $a\to\infty$ taken in the proof increases the neighborhood Ω_{x_0} to infinite length in the direction of v. Restricting the neighborhood to a bounded interval

in the direction of v is equivalent to a windowed Fourier transform which causes some side effects by the sinus function emerging in the corresponding equation. Without taking the limit an $i2D$-signal is created implicitly as it is assumed that the signal is zero outside the neighborhood. For the moment this result is sufficient.

As a consequence $i2D$**-type** signals can be supported on the whole $\mathbb{R}^4$ in the frequency domain. An $i2D$-selective filter which is not affected by signals with lower intrinsic dimensionality must be supported on $\mathbb{R}^4 \backslash M$. In order to derive a generalized version of Gabor-filters, different cases of $i2D$-signals are investigated. First, two "crossing lines" are considered. It is assumed that the signal consists of two δ-lines in different directions $v_1 \neq v_2$ intersecting in x_0. This means $\tilde{u}(t) = \delta(t_1) + \delta(t_2),\ \forall t \in \mathbb{R}^2$. It thus follows

$$\begin{aligned} S(\tilde{u})(z) =& e^{-iz\cdot x_0} \int_{\mathbb{R}^2} (\delta(t_1) + \delta(t_2)) e^{-iz\cdot(Vt)}\, dt \\ =& e^{-iz\cdot x_0} (2\pi)^{1/2} (\delta(z \cdot v_1) + \delta(z \cdot v_2)). \end{aligned} \tag{2.41}$$

The Fourier transform of g thus becomes qualitatively

$$\begin{aligned} \mathcal{F}(g)(z) =& e^{-iz_1\cdot x_0} e^{-iz_2\cdot x_0} \frac{k}{2\pi} (\delta(z_1 \cdot v_1)\delta(z_2 \cdot v_1) + \delta(z_1 \cdot v_1)\delta(z_2 \cdot v_2) \\ &+ \delta(z_1 \cdot v_2)\delta(z_2 \cdot v_1) + \delta(z_1 \cdot v_2)\delta(z_2 \cdot v_2)) \\ =& \begin{cases} \neq 0 \quad , \ [\langle z_1, v_1\rangle = 0 \wedge \langle z_2, v_1\rangle = 0] \quad (\subset M) \\ \qquad\qquad \vee [\langle z_1, v_1\rangle = 0 \wedge \langle z_2, v_2\rangle = 0] \\ \qquad\qquad \vee [\langle z_1, v_2\rangle = 0 \wedge \langle z_2, v_1\rangle = 0] \\ \qquad \vee [\langle z_1, v_2\rangle = 0 \wedge \langle z_2, v_2\rangle = 0] \quad (\subset M), \\ 0 \qquad\qquad\qquad\qquad , \text{ else.} \end{cases} \end{aligned} \tag{2.42}$$

Note that the Fourier transform of this signal type has function values unequal to zero on a subset of M. This subset cannot be used to design an $i2D$-selective filter. But the set defined by the constraints $\langle z_1, v_1\rangle = 0 \wedge \langle z_2, v_2\rangle = 0$ and $\langle z_1, v_2\rangle = 0 \wedge \langle z_2, v_1\rangle = 0$ can be used to design a bandpass filter which is selective to signals with characteristic directions v_1 and v_2. This means that the center frequency f_{center} of a suitable bandpass should be positioned at some point defined by

$$f_{center} = s_1 \begin{pmatrix} n_1 \\ 0 \end{pmatrix} + s_2 \begin{pmatrix} 0 \\ n_2 \end{pmatrix} \quad \text{or} \quad f_{center} = s_2 \begin{pmatrix} n_2 \\ 0 \end{pmatrix} + s_1 \begin{pmatrix} 0 \\ n_1 \end{pmatrix} \tag{2.43}$$

where $s_1, s_2 \in \mathbb{R}$ with $s_1 \neq 0 \wedge s_2 \neq 0$ and $n_1, n_2 \in \mathbb{R}^2$ with $n_1 \perp v_1$ and $n_2 \perp v_2$.

The second case is the so called "end-stopped" line. The signal is assumed to be one end

point with a line leaving in direction v_1. With $\tilde{u}(t) = \delta(t_2)\frac{1}{2}(1 + sign(t_1))$ the following holds

$$\begin{aligned}
S(\tilde{u})(z) =& e^{-iz\cdot x_0} \int_{\mathbb{R}} \frac{1}{2}(1 + sign(t_1)) \int_{\mathbb{R}} \delta(t_2) e^{-iz\cdot(Vt)}\ dt_2\ dt_1 \\
=& e^{-iz\cdot x_0} \int_{\mathbb{R}} \frac{1}{2}(1 + sign(t_1)) e^{-i(z\cdot v_1)t_1} dt_1 \\
=& e^{-iz\cdot x_0} (2\pi)^{\frac{1}{2}} \mathcal{F}(\frac{1}{2}(1 + sign(t_1)))(z \cdot v_1) \\
=& e^{-iz\cdot x_0} \frac{1}{2}(2\pi)^{\frac{1}{2}} \left(\delta(z \cdot v_1) + \frac{1}{i\pi(z \cdot v_1)} \right) \\
=& e^{-iz\cdot x_0} \frac{1}{2}(2\pi)^{\frac{1}{2}} \begin{cases} \delta(z \cdot v_1) & ,\ \langle z, v_1 \rangle = 0, \\ \frac{1}{i\pi(z\cdot v_1)} & ,\ \text{else.} \end{cases}
\end{aligned} \tag{2.44}$$

Using this result, it follows

$$\begin{aligned}
&\mathcal{F}(g)(z) \\
=& e^{-iz_1\cdot x_0} e^{-iz_2\cdot x_0} \frac{k}{4(2\pi)} \begin{cases} \delta(z_1 \cdot v_1)\delta(z_2 \cdot v_1) & ,\ \langle z_1, v_1 \rangle = 0 \wedge \langle z_2, v_1 \rangle = 0 \quad (\subset M), \\ \frac{\delta(z_1\cdot v_1)}{i\pi(z_2\cdot v_1)} & ,\ \langle z_1, v_1 \rangle = 0 \wedge \langle z_2, v_1 \rangle \neq 0, \\ \frac{\delta(z_2\cdot v_1)}{i\pi(z_1\cdot v_1)} & ,\ \langle z_1, v_1 \rangle \neq 0 \wedge \langle z_2, v_1 \rangle = 0, \\ -\frac{1}{\pi^2(z_1\cdot v_1)(z_2\cdot v_1)} & ,\ \langle z_1, v_1 \rangle \neq 0 \wedge \langle z_2, v_1 \rangle \neq 0 \quad (\cap M \neq \emptyset). \end{cases}
\end{aligned} \tag{2.45}$$

This signal type has significantly high complex function values in non-forbidden regions defined by $\langle z_1, v_1 \rangle = 0 \wedge \langle z_2, v_1 \rangle \neq 0$ and $\langle z_1, v_1 \rangle \neq 0 \wedge \langle z_2, v_1 \rangle = 0$. The first line in the case distinction lies in the forbidden region for $i2D$-operators and the last line intersects the forbidden region as well as it decreases quadratically with the length of z_1 and z_2. Therefore, the center frequency of a suitable bandpass filter f_{center} could be

$$f_{center} = s_1 \begin{pmatrix} n_1 \\ 0 \end{pmatrix} + t_1 \begin{pmatrix} 0 \\ v_1 \end{pmatrix} \quad \text{or} \quad f_{center} = t_1 \begin{pmatrix} v_1 \\ 0 \end{pmatrix} + s_1 \begin{pmatrix} 0 \\ n_1 \end{pmatrix} \tag{2.46}$$

where $s_1, t_1 \in \mathbb{R}$ with $s_1 \neq 0 \wedge t_1 \neq 0$ and $n_1 \in \mathbb{R}^2$ with $n_1 \perp v_1$.

The last signal of $i2D$-type, we consider, is an "oriented corner" constructed by two end-stopped lines. The signal consists of two δ-lines which have the same origin but different leaving directions v_1 and v_2. Thus, $\tilde{u}(t) = \frac{1}{2}\delta(t_1)(1+sign(t_2))+\delta(t_2)\frac{1}{2}(1+sign(t_1))$, $v_1 \neq \alpha v_2$, $\alpha \in \mathbb{R}$, such that the following holds

$$\begin{aligned}
S(\tilde{u})(z) =& e^{-iz\cdot x_0} [\int_{\mathbb{R}} \frac{1}{2}(1 + sign(t_1)) \int_{\mathbb{R}} \delta(t_2) e^{-iz\cdot(Vt)}\ dt_2\ dt_1 \\
&+ \int_{\mathbb{R}} \frac{1}{2}(1 + sign(t_2)) \int_{\mathbb{R}} \delta(t_1) e^{-iz\cdot(Vt)}\ dt_1\ dt_2]
\end{aligned}$$

$$
\begin{aligned}
=&e^{-iz\cdot x_0}(2\pi)^{\frac{1}{2}}\left[\mathcal{F}(\frac{1}{2}(1+sign(t_1)))(z\cdot v_1)+\mathcal{F}(\frac{1}{2}(1+sign(t_2)))(z\cdot v_2)\right]\\
=&e^{-iz\cdot x_0}(2\pi)^{\frac{1}{2}}\frac{1}{2}\left[\delta(z\cdot v_1)+\frac{1}{i\pi(z\cdot v_1)}+\delta(z\cdot v_2)+\frac{1}{i\pi(z\cdot v_2)}\right]\\
=&e^{-iz\cdot x_0}(2\pi)^{\frac{1}{2}}\frac{1}{2}\begin{cases}\delta(z\cdot v_1)+\delta(z\cdot v_2) & ,\ \langle z,v_1\rangle=0\wedge\langle z,v_2\rangle=0,\\ \frac{1}{i\pi(z\cdot v_1)}+\delta(z\cdot v_2) & ,\ \langle z,v_1\rangle\neq 0\wedge\langle z,v_2\rangle=0,\\ \delta(z\cdot v_1)+\frac{1}{i\pi(z\cdot v_2)} & ,\ \langle z,v_1\rangle=0\wedge\langle z,v_2\rangle\neq 0,\\ \frac{1}{i\pi(z\cdot v_1)}+\frac{1}{i\pi(z\cdot v_2)} & ,\ \langle z,v_1\rangle\neq 0\wedge\langle z,v_2\rangle\neq 0,\end{cases}\\
=&e^{-iz\cdot x_0}(2\pi)^{\frac{1}{2}}\frac{1}{2}\begin{cases}\delta(z\cdot v_1)+\delta(z\cdot v_2) & ,\ z=0,\\ \frac{1}{i\pi(z\cdot v_1)}+\delta(z\cdot v_2) & ,\ \langle z,v_2\rangle=0,\\ \delta(z\cdot v_1)+\frac{1}{i\pi(z\cdot v_2)} & ,\ \langle z,v_1\rangle=0,\\ \frac{1}{i\pi(z\cdot v_1)}+\frac{1}{i\pi(z\cdot v_2)} & ,\ \text{else.}\end{cases}
\end{aligned}
\tag{2.47}
$$

The last step can be done because it is assumed that $v_1 \neq \alpha v_2$, $\alpha \in \mathbb{R}$. For the following equation the notation is as follows. If it is assumed that $\langle z, v\rangle = 0$, the case $z = 0$ is excluded. Using this and the definition $W := \{z \in \mathbb{R}^2 | \langle z, v_1\rangle \neq 0 \wedge \langle z, v_2\rangle \neq 0\}$ we get

$$
\begin{aligned}
\mathcal{F}(g)(z) =&e^{-iz_1\cdot x_0}e^{-iz_2\cdot x_0}\frac{k}{4(2\pi)}\\
&\cdot\begin{cases}
(\delta(z_1\cdot v_1)+\delta(z_1\cdot v_2))(\delta(z_2\cdot v_1)+\delta(z_2\cdot v_2)) & ,\ z_1=0\wedge z_2=0,\\
(\delta(z_1\cdot v_1)+\frac{1}{i\pi(z_1\cdot v_2)})(\delta(z_2\cdot v_1)+\frac{1}{i\pi(z_2\cdot v_2)}) & ,\ \langle z_1,v_1\rangle=0\wedge\langle z_2,v_1\rangle=0,\\
(\frac{1}{i\pi(z_1\cdot v_1)}+\delta(z_1\cdot v_2))(\frac{1}{i\pi(z_2\cdot v_1)}+\delta(z_2\cdot v_2)) & ,\ \langle z_1,v_2\rangle=0\wedge\langle z_2,v_2\rangle=0,\\
(\delta(z_1\cdot v_1)+\delta(z_1\cdot v_2))(\delta(z_2\cdot v_1)+\frac{1}{i\pi(z_2\cdot v_2)}) & ,\ z_1=0\wedge\langle z_2,v_1\rangle=0,\\
(\delta(z_1\cdot v_1)+\delta(z_1\cdot v_2))(\frac{1}{i\pi(z_2\cdot v_1)}+\delta(z_2\cdot v_2)) & ,\ z_1=0\wedge\langle z_2,v_2\rangle=0,\\
(\delta(z_1\cdot v_1)+\frac{1}{i\pi(z_1\cdot v_2)})(\delta(z_2\cdot v_1)+\delta(z_2\cdot v_2)) & ,\ z_2=0\wedge\langle z_1,v_1\rangle=0,\\
(\frac{1}{i\pi(z_1\cdot v_1)}+\delta(z_1\cdot v_2))(\delta(z_2\cdot v_1)+\delta(z_2\cdot v_2)) & ,\ z_2=0\wedge\langle z_1,v_2\rangle=0,\\
(\delta(z_1\cdot v_1)+\delta(z_1\cdot v_2))(\frac{1}{i\pi(z_2\cdot v_1)}+\frac{1}{i\pi(z_2\cdot v_2)}) & ,\ z_1=0\wedge z_2\in W,\\
(\frac{1}{i\pi(z_1\cdot v_1)}+\frac{1}{i\pi(z_1\cdot v_2)})(\delta(z_2\cdot v_1)+\delta(z_2\cdot v_2)) & ,\ z_2=0\wedge z_1\in W,\\
\\
(\delta(z_1\cdot v_1)+\frac{1}{i\pi(z_1\cdot v_2)})(\frac{1}{i\pi(z_2\cdot v_1)}+\delta(z_2\cdot v_2)) & ,\ \langle z_1,v_1\rangle=0\wedge\langle z_2,v_2\rangle=0,\\
(\frac{1}{i\pi(z_1\cdot v_1)}+\delta(z_1\cdot v_2))(\delta(z_2\cdot v_1)+\frac{1}{i\pi(z_2\cdot v_2)}) & ,\ \langle z_1,v_2\rangle=0\wedge\langle z_2,v_1\rangle=0,\\
(\delta(z_1\cdot v_1)+\frac{1}{i\pi(z_1\cdot v_2)})(\frac{1}{i\pi(z_2\cdot v_1)}+\frac{1}{i\pi(z_2\cdot v_2)}) & ,\ \langle z_1,v_1\rangle=0\wedge z_2\in W,\\
(\frac{1}{i\pi(z_1\cdot v_1)}+\delta(z_1\cdot v_2))(\frac{1}{i\pi(z_2\cdot v_1)}+\frac{1}{i\pi(z_2\cdot v_2)}) & ,\ \langle z_1,v_2\rangle=0\wedge z_2\in W,\\
(\frac{1}{i\pi(z_1\cdot v_1)}+\frac{1}{i\pi(z_1\cdot v_2)})(\delta(z_2\cdot v_1)+\frac{1}{i\pi(z_2\cdot v_2)}) & ,\ \langle z_2,v_1\rangle=0\wedge z_1\in W,\\
(\frac{1}{i\pi(z_1\cdot v_1)}+\frac{1}{i\pi(z_1\cdot v_2)})(\frac{1}{i\pi(z_2\cdot v_1)}+\delta(z_2\cdot v_2)) & ,\ \langle z_2,v_2\rangle=0\wedge z_1\in W,\\
(\frac{1}{i\pi(z_1\cdot v_1)}+\frac{1}{i\pi(z_1\cdot v_2)})(\frac{1}{i\pi(z_2\cdot v_1)}+\frac{1}{i\pi(z_2\cdot v_2)}) & ,\ z_1\in W\wedge z_2\in W,
\end{cases}
\end{aligned}
$$

$$= \begin{cases} * + i * & , \ (z_1, z_2) \in M, \\ \\ \infty - i \ sign(\frac{1}{z_1 \cdot v_2} + \frac{1}{z_2 \cdot v_1})\infty & , \ \langle z_1, v_1 \rangle = 0 \wedge \langle z_2, v_2 \rangle = 0, \\ \infty - i \ sign(\frac{1}{z_1 \cdot v_1} + \frac{1}{z_2 \cdot v_2})\infty & , \ \langle z_1, v_2 \rangle = 0 \wedge \langle z_2, v_1 \rangle = 0, \\ * - i \ sign(\frac{1}{z_2 \cdot v_1} + \frac{1}{z_2 \cdot v_2})\infty & , \ \langle z_1, v_1 \rangle = 0 \wedge z_2 \in W, \\ * - i \ sign(\frac{1}{z_2 \cdot v_1} + \frac{1}{z_2 \cdot v_2})\infty & , \ \langle z_1, v_2 \rangle = 0 \wedge z_2 \in W, \\ * - i \ sign(\frac{1}{z_1 \cdot v_1} + \frac{1}{z_1 \cdot v_2})\infty & , \ \langle z_2, v_1 \rangle = 0 \wedge z_1 \in W, \\ * - i \ sign(\frac{1}{z_1 \cdot v_1} + \frac{1}{z_1 \cdot v_2})\infty & , \ \langle z_2, v_2 \rangle = 0 \wedge z_1 \in W, \\ * + i \ 0 & , \ z_1 \in W \wedge z_2 \in W. \end{cases} \tag{2.48}$$

The upper bundle of equations describes the behavior of the Fourier transform in the forbidden region M. In the lower bundle of equations it can be seen that this kind of signal has high imaginary values in particular in the two-dimensional planes described by the equations $\langle z_1, v_1 \rangle = 0 \wedge \langle z_2, v_2 \rangle = 0$ and $\langle z_1, v_2 \rangle = 0 \wedge \langle z_2, v_1 \rangle = 0$. The center frequency f_{center} of a suitable bandpass should be positioned at some point defined by

$$f_{center} = s_1 \begin{pmatrix} n_1 \\ 0 \end{pmatrix} + s_2 \begin{pmatrix} 0 \\ n_2 \end{pmatrix} \quad \text{or} \quad f_{center} = s_2 \begin{pmatrix} n_2 \\ 0 \end{pmatrix} + s_1 \begin{pmatrix} 0 \\ n_1 \end{pmatrix} \tag{2.49}$$

where $s_1, s_2 \in \mathbb{R}$ with $s_1 \neq 0 \wedge s_2 \neq 0$ and $n_1, n_2 \in \mathbb{R}^2$ with $n_1 \perp v_1$ and $n_2 \perp v_2$. Note that in these regions the considered signal class also has high real function values which are similar to the first case "crossing lines" and can cause trouble in their distinction.

In summary, we derive from this analysis of $i2D$-type signals:

- "Crossing lines" can be detected by an even-symmetric real-valued filter kernel with center frequency given by Equation (2.43). This filter also detects oriented corners with the same leaving directions.
- "End-stopped lines" can be detected by an odd-symmetric imaginary-valued filter kernel with center frequency given by Equation (2.46).
- "Oriented corners" can be detected by an odd-symmetric imaginary-valued filter kernel with center frequency given by Equation (2.49).

The bandwidth of these filter kernels is considered within the following framework of generalized Gabor-filter kernels.

2.3.4 Generalized Gabor approach

In order to design $i2D$-selective second-order Volterra systems, which are able to detect $i2D$-signals, we define the following filter kernels in Fourier space analogously to the two-dimensional case in Lemma 2.7.

Definition 2.12 (Generalized Gabor filter)**.** Let $V \in \mathbb{R}^{4\times4}$ be an orthonormal coordinate transformation matrix, i.e. $V^{-1} = V^T$ and $\det(V) = 1$. Let $\Sigma \in \mathbb{R}^{4\times4}$ be a diagonal matrix and $f_{center} \in \mathbb{R}^4$. The functions $H_{even} : \mathbb{R}^4 \to \mathbb{C}$ and $H_{odd} : \mathbb{R}^4 \to \mathbb{C}$ defined by

$$H_{even}(z) = k\left(e^{-\frac{1}{2}(z-f_{center})^T V\Sigma^{-1}V^T(z-f_{center})} + e^{-\frac{1}{2}(z+f_{center})^T V\Sigma^{-1}V^T(z+f_{center})}\right) \tag{2.50}$$

$$\text{and } H_{odd}(z) = ki\left(e^{-\frac{1}{2}(z-f_{center})^T V\Sigma^{-1}V^T(z-f_{center})} - e^{-\frac{1}{2}(z+f_{center})^T V\Sigma^{-1}V^T(z+f_{center})}\right) \tag{2.51}$$

are the *generalized Gabor filter functions.*

Remark 2.13. For further investigations we define the matrix V with respect to two orientations ϕ_1 and ϕ_2 as follows

$$V = V(\phi_1, \phi_2) = \begin{pmatrix} v_1 & n_1 & 0 & 0 \\ 0 & 0 & v_2 & n_2 \end{pmatrix} = \begin{pmatrix} \cos(\phi_1) & -\sin(\phi_1) & 0 & 0 \\ \sin(\phi_1) & \cos(\phi_1) & 0 & 0 \\ 0 & 0 & \cos(\phi_2) & -\sin(\phi_2) \\ 0 & 0 & \sin(\phi_2) & \cos(\phi_2) \end{pmatrix}. \tag{2.52}$$

The matrix Σ is defined by

$$\Sigma = \begin{pmatrix} \sigma_{v_1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{n_1}^2 & 0 & 0 \\ 0 & 0 & \sigma_{v_2}^2 & 0 \\ 0 & 0 & 0 & \sigma_{n_2}^2 \end{pmatrix} \tag{2.53}$$

where σ_x determines the bandwidth in the direction of x. The center frequency f_{center} is defined by

$$f_{center} = V(\phi_1, \phi_2)\mu \tag{2.54}$$

where $\mu = (\mu_{v_1}, \mu_{n_1}, \mu_{v_2}, \mu_{n_2})^T$. μ_x determines the center frequency in the direction of x.

Similar to the classical Gabor approach, the generalized Gabor filter is determined by a low number of parameters. Given the definitions in the previous remark, ten parameters are necessary:

- Two angles ϕ_1 and ϕ_2 which define the orientations of interest.
- Four parameters μ_{v_1}, μ_{n_1}, μ_{v_2}, and μ_{n_2} to describe the center frequency where the bandpass is located. The effect of this parameters strongly depends on the choice of ϕ_1 and ϕ_2.

- Four parameters σ_{v_1}, σ_{n_1}, σ_{v_2}, and σ_{n_2} which determine the bandwidth of the bandpass filter (given a σ the full width at half maximum is $2\sqrt{2\ln(2)}\sigma$).

By using these parameters we can identify suitable parameter settings for the three cases of $i2D$-type signals "crossing lines", "end-stopped", and "oriented corner", which are analyzed in the previous section. The constraints for the parameters in each case can be found in Table 1. The center frequencies and the symmetry types result from the previous theoretical analysis.

The bandwidth parameters in the cases "crossing lines" and "oriented corner" are restricted in a way that the support of the bandpass filter does not include the origin. This results in the restriction of σ_{n_1} and σ_{n_2}. In order to be as orientation selective as possible, σ_{v_1} and σ_{v_2} should be chosen sufficiently small. The upper boundary for σ_{v_1} is determined by the intersection point of the forbidden region M_1 and the curve defined by

$$g : z = \begin{pmatrix} \mu_{n_1} n_1 \\ \mu_{n_2} n_2 \end{pmatrix} + t \begin{pmatrix} v_1 \\ 0 \end{pmatrix} \quad , \; t \in \mathbb{R}. \tag{2.55}$$

Finding an intersection point of g and M_1 is equivalent to find t, s_1, s_2, ϕ such that

$$\begin{pmatrix} \mu_{n_1} n_1 \\ \mu_{n_2} n_2 \end{pmatrix} + t \begin{pmatrix} v_1 \\ 0 \end{pmatrix} = s_1 \begin{pmatrix} n(\phi) \\ 0 \end{pmatrix} + s_2 \begin{pmatrix} 0 \\ n(\phi) \end{pmatrix}. \tag{2.56}$$

From the third and fourth dimension follows that $\phi = \phi_2 + \pi_2$. Using this in the equations defined by the first and second dimension yields

$$t = \tan(\phi_1 - \phi_2)\mu_{n_1}. \tag{2.57}$$

By definition this t is the distance between the center frequency and the point of intersection. The constraint to σ_{v_1} thus becomes $\sigma_{v_1} \leq |\tan(\phi_1 - \phi_2)\mu_{n_1}|/\sqrt{2\ln(2)}$. Analog considerations with respect to σ_{v_2} yield $\sigma_{v_2} \leq |\tan(\phi_2 - \phi_1)\mu_{n_2}|/\sqrt{2\ln(2)}$.

The bandwidth parameters σ_{n_1} and σ_{v_2} of the case "end-stopped" are also chosen in a way that the origin is excluded. The other two bandwidth parameters should be as big as possible to capture the energy of an end-stopped line, c.p. Equation (2.45). To determine these parameters, we have to find the intersection of the forbidden region M and the plane G spanned by the directions of the missing bandwidth parameters, i.e.

$$G : z = \begin{pmatrix} \mu_{n_1} n_1 \\ \mu_{v_2} v_1 \end{pmatrix} + t_v \begin{pmatrix} v_1 \\ 0 \end{pmatrix} + t_n \begin{pmatrix} 0 \\ n_1 \end{pmatrix} \quad , \; t_v, t_n \in \mathbb{R}. \tag{2.58}$$

We thus must find all $z \in \mathbb{R}^4$ with $z \in G$ and $z \in M$. This means, find ϕ, s_1, s_2, t_v, and t_n

such that

$$\begin{pmatrix}\mu_{n_1} n_1 \\ \mu_{v_2} v_1\end{pmatrix} + t_v \begin{pmatrix} v_1 \\ 0 \end{pmatrix} + t_n \begin{pmatrix} 0 \\ n_1 \end{pmatrix} = s_1 \begin{pmatrix} n(\phi) \\ 0 \end{pmatrix} + s_2 \begin{pmatrix} 0 \\ n(\phi) \end{pmatrix}. \tag{2.59}$$

From this follows

$$\mu_{n_1} n_1 + t_v v_1 = \frac{s_1}{s_2} \left(\mu_{v_2} v_1 + t_n n_1\right) \tag{2.60}$$

$$\Leftrightarrow \left(\mu_{n_1} - \frac{s_1}{s_2} t_n\right) n_1 = \left(\frac{s_1}{s_2} \mu_{v_2} - t_v\right) v_1. \tag{2.61}$$

With $v_1 \perp n_1$ follows

$$t_v = \frac{\mu_{n_1} \mu_{v_2}}{t_n}. \tag{2.62}$$

Using this equation all points within the intersection are given by

$$z(t) = \begin{pmatrix}\mu_{n_1} n_1 \\ \mu_{v_2} v_1\end{pmatrix} + \frac{\mu_{n_1} \mu_{v_2}}{t} \begin{pmatrix} v_1 \\ 0 \end{pmatrix} + t \begin{pmatrix} 0 \\ n_1 \end{pmatrix} \quad , \; t \neq 0. \tag{2.63}$$

The point with the shortest distance to the center frequency then limits the maximum bandwidth, i.e. find

$$\arg\min_t(\|f_{center} - z(t)\|_2^2) = \arg\min_t(\frac{\mu_{n_1}^2 \mu_{v_2}^2}{t^2} + t^2). \tag{2.64}$$

From this follows that the point with the shortest distance is determined by $t_n = t_v = \sqrt{\mu_{n_1} \mu_{v_2}}$. This results in the constraints $\sigma_{v_1} < \sqrt{\mu_{n_1} \mu_{v_2}}/\sqrt{2\ln(2)}$ and $\sigma_{n_2} < \sqrt{\mu_{n_1} \mu_{v_2}}/\sqrt{2\ln(2)}$ as can be found in Table 1. Note that if one parameter is close to the boundary of the constraint, the function value of the filter in a region close to the forbidden region is approximately half the amplitude. In order to avoid responses to $i0D$- and $i1D$-signals, the parameter has to be chosen sufficiently smaller than the determined boundaries.

2.3.5 Results

The three kinds of nonlinear Gabor filters, which were introduced in the previous section, are parametrized as shown in Table 2. All filters are then applied to two different datasets of stimuli which cover the cases of interest, i.e. $i1D$-signals, crossing lines, end-stopped lines, and oriented corners. The first filter H_1 is designed to detect crossing lines of specific orientations ϕ_1 and ϕ_2. As a result of the previous analysis of $i2D$-signals this filter is even-symmetric because of the real valued Fourier transform of the respective signal of crossing lines in Equation (2.42). In order to capture the characteristic of the signal and to gain

Case	"crossing lines"	"end-stopped"	"oriented corner"
Symmetry	even	odd	odd
ϕ_1	$\in [0, \pi)$	$\in [0, \pi)$	$\in [0, \pi)$
ϕ_2	$\neq \phi_1$	$= \phi_1$	$\neq \phi_1$
μ_{v_1}	$= 0$	$= 0$	$= 0$
μ_{n_1}	$\neq 0$	$\neq 0$	$\neq 0$
μ_{v_2}	$= 0$	$\neq 0$	$= 0$
μ_{n_2}	$\neq 0$	$= 0$	$\neq 0$
σ_{v_1}	$< \frac{\lvert\tan(\phi_1-\phi_2)\mu_{n_1}\rvert}{\sqrt{2\ln(2)}}$	$< \frac{\sqrt{\mu_{n_1}\mu_{v_2}}}{\sqrt{2\ln(2)}}$	$< \frac{\lvert\tan(\phi_1-\phi_2)\mu_{n_1}\rvert}{\sqrt{2\ln(2)}}$
σ_{n_1}	$< \frac{\lvert\mu_{n_1}\rvert}{\sqrt{2\ln(2)}}$	$< \frac{\lvert\mu_{n_1}\rvert}{\sqrt{2\ln(2)}}$	$< \frac{\lvert\mu_{n_1}\rvert}{\sqrt{2\ln(2)}}$
σ_{v_2}	$< \frac{\lvert\tan(\phi_2-\phi_1)\mu_{n_2}\rvert}{\sqrt{2\ln(2)}}$	$< \frac{\lvert\mu_{v_2}\rvert}{\sqrt{2\ln(2)}}$	$< \frac{\lvert\tan(\phi_2-\phi_1)\mu_{n_2}\rvert}{\sqrt{2\ln(2)}}$
σ_{n_2}	$< \frac{\lvert\mu_{n_2}\rvert}{\sqrt{2\ln(2)}}$	$< \frac{\sqrt{\mu_{n_1}\mu_{v_2}}}{\sqrt{2\ln(2)}}$	$< \frac{\lvert\mu_{n_2}\rvert}{\sqrt{2\ln(2)}}$

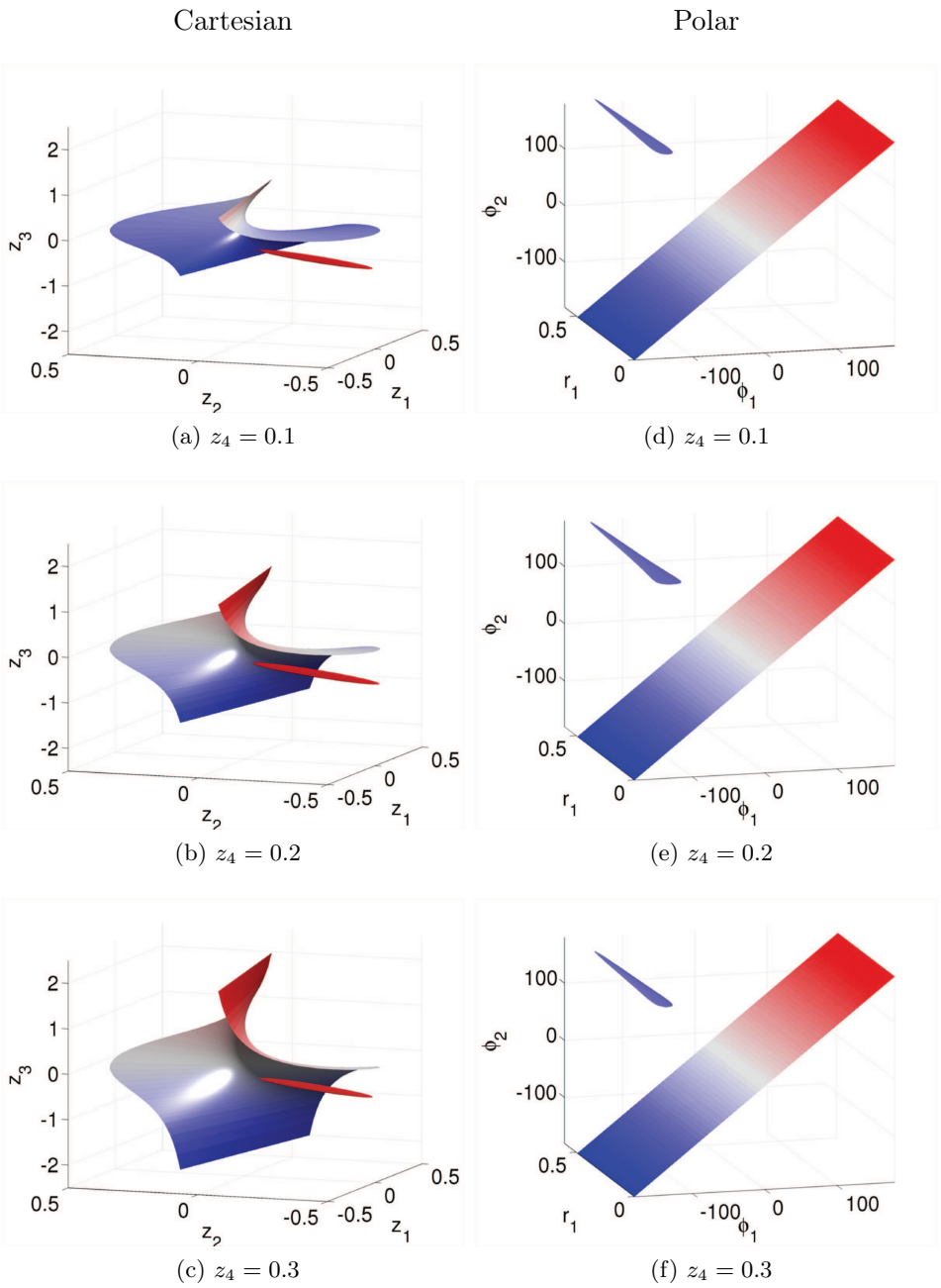
Table 1: This table shows the constraints to the parameters of the corresponding generalized Gabor filter functions for the cases "crossing lines", "end-stopped", and "oriented corner". The angles ϕ_i determine the orientations of interest and the coordinate transformation, the values μ_i specify the center frequency of the bandpass regions, and the σ_i values determine the bandwidth of the passband regions. Note that the boundaries for σ_i describe the distance to the forbidden region in the respective direction where the filter has half of its amplitude.

Filter type	H_1: "crossing lines"	H_2: "end-stopped"	H_3: "oriented corner"
Symmetry	even	odd	odd
ϕ_1	30°	30°	30°
ϕ_2	60°	30°	60°
μ_{v_1}	0	0	0
μ_{n_1}	0.4	−0.4	0.4
μ_{v_2}	0	−0.4	0
μ_{n_2}	0.4	0	0.4
σ_{v_1}	0.02	0.05	0.02
σ_{n_1}	0.1	0.3	0.1
σ_{v_2}	0.02	0.02	0.02
σ_{n_2}	0.1	0.3	0.1

Table 2: This table gives an overview over the parametrization of the three filter setups H_1, H_2, and H_3.

a high response at the crossing position, the center frequency is chosen in the direction of $(n_1, 0)^T$ and $(0, n_2)^T$. The bandwidths in these directions, i.e. the parameters σ_{n_1} and σ_{n_2}, are chosen sufficiently high such that the orientations of interest are captured by the filter. The bandwidth in the remaining two directions has to be as small as possible to obtain the desired selectivity is guaranteed. The second reason for a small bandwidth is the risk of an intersection of the filter kernel's support and the forbidden region. The third filter H_3, which is designed to be sensitive to oriented corners with leaving directions determined by the orientations ϕ_1 and ϕ_2, is parametrized like the filter H_1. The argumentation is the same with one important difference: As can be seen in Equation (2.48) the Fourier transform of a corner has significant imaginary function values in the same regions compared to the crossing lines. To capture this structure, the filter kernel is odd-symmetric and imaginary-valued. The second filter H_2 is designed to be sensitive to end-stopped lines which is a special case of the oriented corners with the same orientations ϕ_1 and ϕ_2. The parametrization has to be different because the equal orientations would cause an $i1D$-selectivity in the filter H_3. The previous investigation resulted in center frequencies in the directions $(n_1, 0)^T$ and $(0, v_2)^T$ and in an imaginary-valued odd-symmetric design. The bandwidth in the direction of n_1 should be sufficiently large to capture the desired orientation. The bandwidth in the direction of v_1 must be sufficiently small as it determines the selectivity with respect to the orientation similar to the other filters H_1 and H_3. As can be seen in Equation (2.45) the bandwidth in the direction of n_2 must be chosen sufficiently large to capture the structure of the signal. The bandwidth in the direction of v_2 is chosen smaller in order to not implement a corner-selective filter. A selectivity to right angled corners cannot be avoided completely as the center frequencies define this kind of selectivity. The support of the filter function in the four-dimensional domain and its pose with respect to the forbidden region is illustrated in Figure 2.4. From the illustration in polar coordinates where the forbidden region becomes a plane it can easily be deduced that the essential support of the filter does not intersect the forbidden region.

In order to test the selectivity properties of the filters H_1, H_2, and H_3, two kinds of signals are considered. The first kind consists of two lines intersecting in a point x_0 where each line has an orientation θ_1 or θ_2. An example stimulus is illustrated in Figure 2.5(a). In particular, this kind of stimulus includes the $i1D$-case if $\theta_1 = \theta_2$. The second kind also consists of two lines which have their origin in the point x_0 and leave in the directions determined by the angles θ_1 and θ_2. This kind of stimuli can be seen in Figure 2.5(b). For $\theta_1 = \theta_2$ a stimulus becomes an end-stopped line. The two kinds of stimuli build two data sets, "crossing" and "corner". Each set consists of 3600 images of the size 50×50. The angles θ_1 and θ_2 were varied from $0°$ to $180°$ with constant step size resulting in 60 orientations, i.e. a step of $\sim 3°$. The point x_0 is defined in pixel coordinates by $(26, 26)^T$. Each picture constructed this way is normalized.

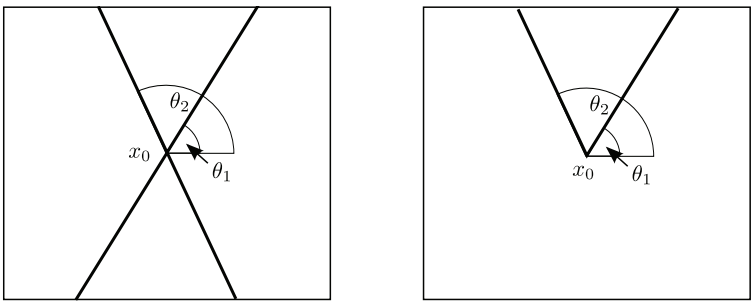


(a) $z_4 = 0.1$ (d) $z_4 = 0.1$

(b) $z_4 = 0.2$ (e) $z_4 = 0.2$

(c) $z_4 = 0.3$ (f) $z_4 = 0.3$

Figure 2.4: The support of the Gaussian filter kernel in Fourier space for the end-stopped filter H_2 is illustrated in Cartesian coordinates (left columns) and in polar coordinates (right column) for various values of z_4. The advantage of the illustration in polar coordinates is that the forbidden region M becomes a plane and it can be seen easily whether the support of the filter function intersects this region. The surface of the ellipsoid determines the region where the filter has half of its amplitude.



(a) "crossing" stimuli (b) "corner" stimuli

Figure 2.5: The two kinds of test stimuli are illustrated. The stimuli of the kind "crossing" (a) are two δ-lines which intersect in the point x_0. Each line has an orientation θ_1 and θ_2. In particular the case $\theta_1 = \theta_2$ builds an $i1D$-signal. The stimuli of the kind "corner" (b) are two δ-lines which have their origin in the point x_0. Each line has its leaving direction determined by the angle θ_1 or θ_2. In particular the case $\theta_1 = \theta_2$ builds an end-stopped line.

All filters H_1, H_2, and H_3 are applied to both datasets. The results of this analysis are illustrated in Figure 2.6 and 2.7. The filter H_1 is designed to be selective to crossing lines with orientations 30° and 60°. As can be seen in Figure 2.6(a) and Figure 2.7(a), the response of the filter has high peaks for the desired orientations within the stimulus, cf. the green circles. The response to $i1D$-signals encircled by the red ellipse is nearly perfectly inhibited. In comparison to the "corner" dataset in Figure 2.6(d) the response to corners with the orientations of interest is approximately four times smaller than the response to the crossing lines having the same orientations. The filter H_2 is designed to detect end-stopped lines with an orientation of 30°. As can be seen in Figure 2.6(e) on the diagonal and in Figure 2.7(e) in the green circle, this filter has a peak for the desired kind of stimuli. One side effect can be observed at the combination of 30° and approximately 30° + 90° in the corner stimuli. This is a result of the center frequency in the direction of $v_2 (= v_1)$. This filter also does not give a high response to $i1D$-signals as can be seen in the red ellipse of Figure 2.7(b). The third filter H_3 also shows the expected $i2D$-selectivity to corners with orientations 30° and 60° as can be seen in Figure 2.7(f) in the green circles. The results in Figure 2.6(c) and (f) show that this filter has the highest peaks for the desired corner stimuli.

Summarized, all filters show the expected behavior and can be used to implement the desired $i2D$-selectivity. The three kinds of $i2D$-selectivity can be captured by those three proposed kinds of filters. Consequently, these image features can be distinguished by their response on the datasets. The only case which causes problems is the distinction between an end-stopped line and a corner with open angle $|\theta_2 - \theta_1| = 90°$. This would require more sophisticated filter implementations. All filters have in common that they show strong orientation selectivity reported for cells in the visual cortex.

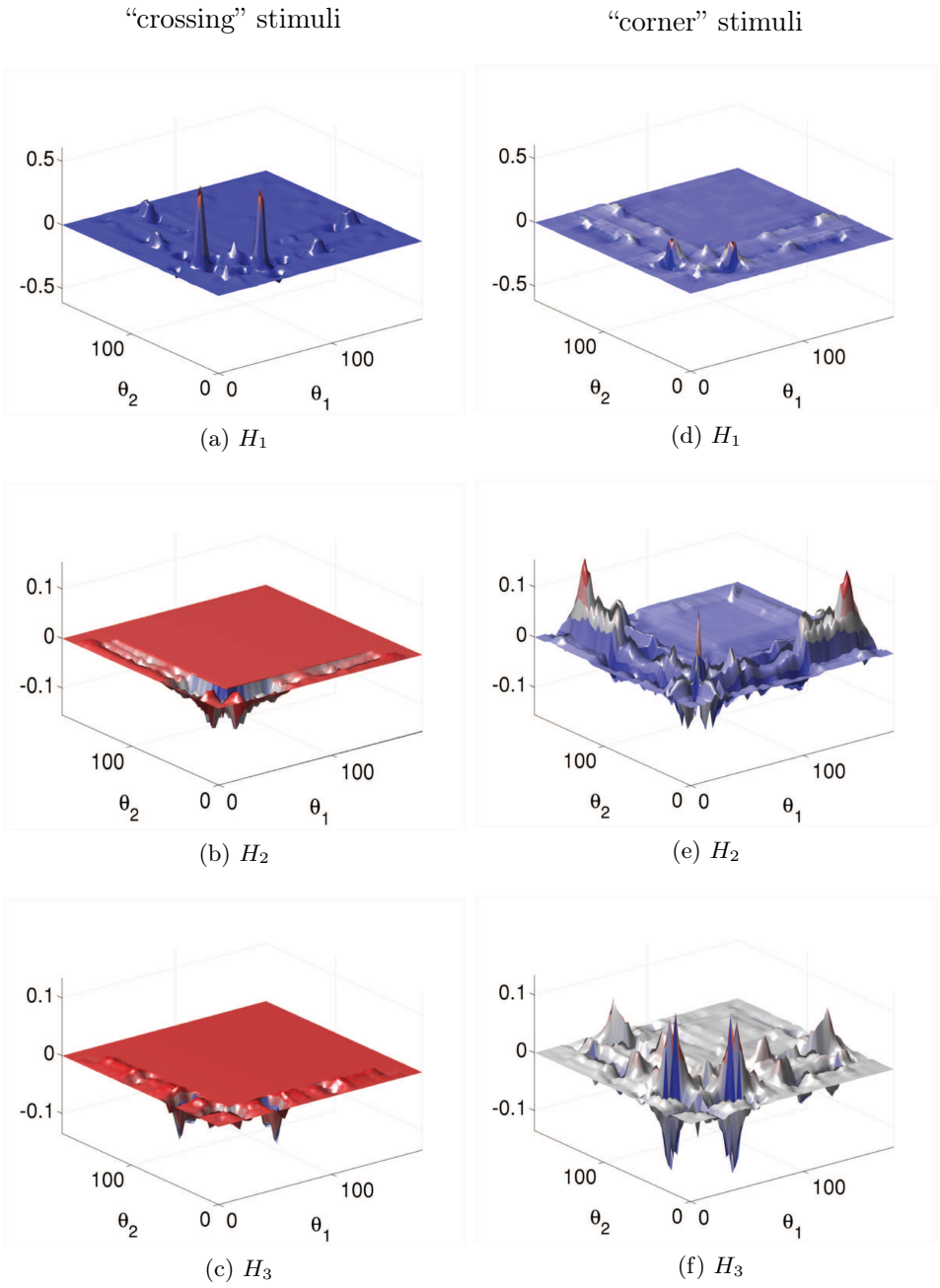
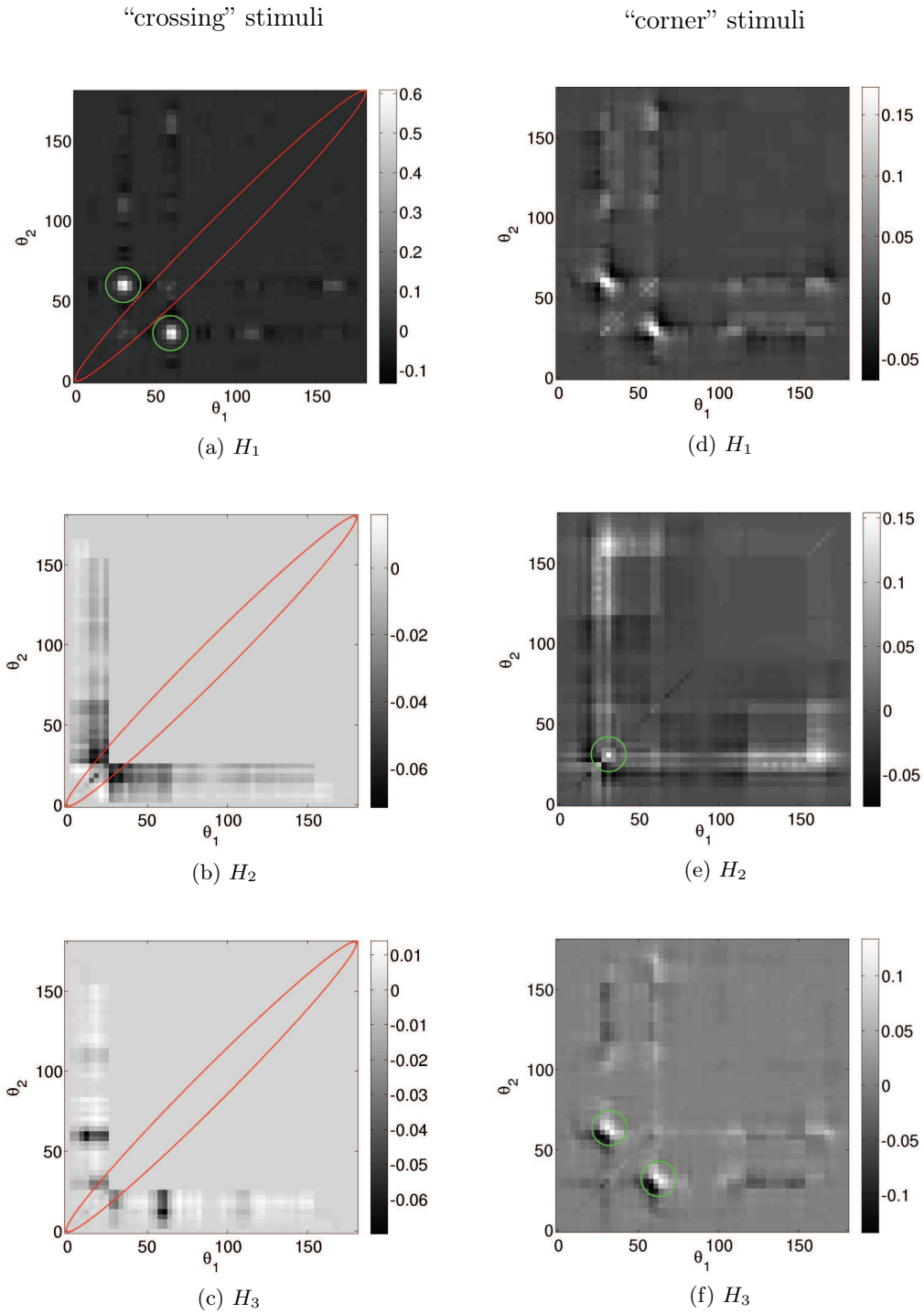


Figure 2.6: The results for the three filter types H_1, H_2, and H_3 are illustrated as three-dimensional surface plots where the first and second dimension determine the orientations in the test stimulus θ_1 and θ_2 varying from 0° to 180°. The third dimension is the response of the respective filter at position $(26, 26)^T$. The stimuli had a size of 50×50 pixels with $x_0 = (26, 26)^T$. The left column ((a)-(c)) shows the results for the "crossing" stimuli and the right columns ((d)-(f)) shows the results for the "corner" stimuli. The ith line corresponds to the results for the filter function H_i for both kinds of test stimuli, $i = 1, 2, 3$. The scaling of both results is the same such that the responses are comparable over the two kinds of test stimuli for each filter function.



(a) H_1 (d) H_1

(b) H_2 (e) H_2

(c) H_3 (f) H_3

Figure 2.7: This figure is organized in the same way and shows the same data as in Figure 2.6. The major difference is that the responses are now color encoded and scaled for each filter and kind of test stimuli such that a comparison between the kinds of test stimuli is not possible in an easy way. In the left column the responses to stimuli which belong to the class if $i1D$-signals are encircled by the red ellipse. The green circles highlight the desired regions of response for each filter H_i, $i = 1, 2, 3$.

2.4 Article: Statistical invariants of spatial form: From local AND to numerosity

Reference

C.Z., K.G., and T.K. designed research; C.Z., K.G., and T.K. performed research; K.G. and T.K. implemented the work; K.G. and T.K. tested the algorithms; C.Z., K.G., and T.K. wrote the paper.

The paper was published in *Proceedings of the Second Interdisciplinary Workshop The Shape of Things* under the following reference [89]:

C. Zetzsche, K. Gadzicki, and T. Kluth. Statistical invariants of spatial form: From local and to numerosity. In *Proceedings of the Second Interdisciplinary Workshop The Shape of Things*, Workshop Proceedings, pages 163–172. CEUR-WS.org, Apr. 2013.

Statistical Invariants of Spatial Form: From Local AND to Numerosity

Christoph ZETZSCHE [1], Konrad GADZICKI and Tobias KLUTH
Cognitive Neuroinformatics, University of Bremen, Germany

Abstract Theories of the processing and representation of spatial form have to take into account recent results on the importance of holistic properties. Numerous experiments showed the importance of "set properties", "ensemble representations" and "summary statistics", ranging from the "gist of a scene" to something like "numerosity". These results are sometimes difficult to interpret, since we do not exactly know how and on which level they can be computed by the neural machinery of the cortex. According to the standard model of a local-to-global neural hierarchy with a gradual increase of scale and complexity, the ensemble properties have to be regarded as high-level features. But empirical results indicate that many of them are primary perceptual properties and may thus be attributed to earlier processing stages. Here we investigate the prerequisites and the neurobiological plausibility for the computation of ensemble properties. We show that the cortex can easily compute common statistical functions, like a probability distribution function or an autocorrelation function, and that it can also compute abstract invariants, like the number of items in a set. These computations can be performed on fairly early levels and require only two well-accepted properties of cortical neurons, linear summation of afferent inputs and variants of nonlinear cortical gain control.

Keywords. shape invariants, peripheral vision, ensemble statistics, numerosity

Introduction

Recent evidence shows that our representation of the world is essentially determined by holistic properties [1,2,3,4,5,6]. These properties are described as "set properties", "ensemble properties", or they are characterized as "summary statistics". They reach from the average orientation of elements in a display [1] over the "gist of a scene"[7,8], to the "numerosity" of objects in a scene [9]. For many of these properties we do not exactly know by which kind of neural mechanisms and on which level of the cortex they are computed. According to the standard view of the cortical representation of shape, these properties have to be considered as high-level features because the cortex is organized in form of a local-to-global processing hierarchy in which features with increasing order of abstraction are computed in a progression of levels [10]. At the bottom, simple and locally restricted geometrical features are computed, whereas global and complex properties are represented at the top levels of the hierarchy. Across levels, invariance is system-

[1]Corresponding Author: Christoph Zetzsche, Cognitive Neuroinformatics, FB3, University of Bremen, P.O. Box 330 440, 28334 Bremen, Germany; E-mail: zetzsche@informatik.uni-bremen.de
Research supported by **DFG** (SFB/TR 8 Spatial Cognition, A5-[ActionSpace])

atically increased such that the final stages are independent of translations, rotations, size changes, and other transformations of the input. However convincing this view seems on first sight, it creates some conceptual difficulties.

The major difficulty concerns the question of what exactly is a low-level and a high-level property. Gestalt theorists already claimed that features considered high-level according to a structuralistic view are primary and basic in terms of perception. Further doubts have been raised by global precedence effects [11]. Similar problems arise with the recently discovered ensemble properties. The gist of a scene, a high-level feature according to the classical view, can be recognized in 150 msec [7,12,13,14] and can be modeled using low-level visual features [8]. In addition, categories can be shown to be faster processed than basic objects, contrary to the established view of the latter as entry-level representations [15]. A summary statistics approach, also based on low-level visual features, can explain the holistic processing properties in the periphery of the visual field [4,16,17]. What is additionally required in these models are statistical measures, like probability distributions and autocorrelation functions, from which it is not known how and on which level of the cortical hierarchy they can be realized.

One of the most abstract ensemble properties seems to be the number of elements in a spatial configuration. However, the ability to recognize this number is not restricted to humans with mature cognitive abilities but has also been found in infants and animals [9,18], recently even in invertebrates [19]. Neural reactions to numerosity are fast (100 msecs in macaques [20]). And finally there is evidence for a "direct visual sense for number" since number seems to be a primary visual property like color, orientation or motion, to which the visual system can be adapted by prolonged viewing [21].

The above observations on ensemble properties raise a number of questions, from which the following are addressed in this paper: Sect. 1: Can the cortex compute a probability distribution? Sect. 2: And also an autocorrelation function? By which kind of neural hardware can this be achieved? Sect.3: Can the shape of individual objects also be characterized by such mechanisms? Sect. 4: What is necessary to compute such an abstract property like the number of elements in a spatial configuration? Can this be achieved in early sensory stages?

1. Neural Computation of a Probability Distribution

Formally, the probability density function $p_e(e)$ of a random variable $\mathbf{e}$ is defined via the cumulative distribution function: $p_e(e) \triangleq \frac{dP_e(e)}{de}$ with $P_e(e) = Pr[\mathbf{e} \leq e]$. Their empirical counterparts, the histogram and the cumulative histogram, are defined by use of *indicator functions*. For this we divide the real line into m bins $(e^{(i)}, e^{(i+1)}]$ with bin size $\Delta e = e^{(i+1)} - e^{(i)}$. For each bin i, an indicator function is defined as

$$Q_i(e) = 1_i(e) = \begin{cases} 1, \text{ if } e^{(i)} < e \leq e^{(i+1)} \\ 0, \text{ else} \end{cases} \tag{1}$$

An illustration of such a function is shown in Fig. 1a. From N samples e_k of the random variable $\mathbf{e}$ we then obtain the histogram as $h(i) = \frac{1}{N}\sum_{k=1}^{N} Q_i(e_k)$. The cumulative histogram $H_e(e)$ can be computed by changing the bins to $(e^{(1)}, e^{(i+1)}]$ (cf. Fig. 1b), and by performing the same summation as for the normal histogram. The reverse cumulative

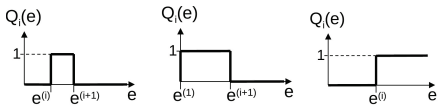


Figure 1. Indicator functions. Basic types are: (a) indicator function for computation of a classical histogram. (b) indicator function for a cumulative histogram. (c) indicator function for a reverse cumulative histogram.

histogram $\bar{H}(i)$ is simply the reversed version of the cumulative histogram. The corresponding bins are $\Delta e_i = (e^{(i)}, e^{(m+1)}]$ and the indicator functions are defined as (Fig. 1c)

$$Q_i(e) = 1_i(e) = \begin{cases} 1, \text{ if } e \geq e^{(i)} \\ 0, \text{ else} \end{cases} \tag{2}$$

The corresponding system is shown in Fig. 2.

The three types of histograms have identical information content since they are related to each other as

$$h(i) = H((i+1)) - H(i) = \bar{H}(i) - \bar{H}(i+1) \quad \text{and} \quad H(i) = 1 - \bar{H}(i) = \sum_{j=1}^{i} h(j). \tag{3}$$

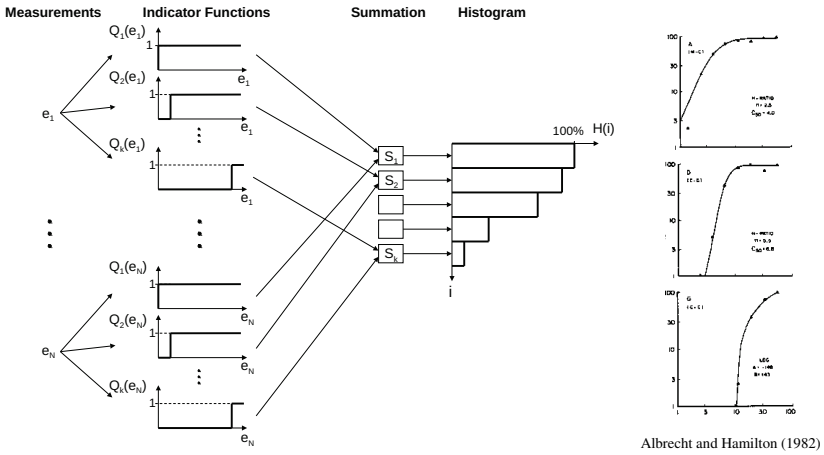


Figure 2. Computation of the reverse cumulative histogram. (a) shows the set of input variables e_1 to e_n over which the histogram should be computed. Each of these variables is input to a set of *indicator functions* $Q_i(e_k)$. For each bin of the histogram there is a summation unit S_i which sums over all indicator function outputs with index i, i.e. over all $Q_i(e_k)$.
(b) The response functions of three neurons in the visual cortex [22]. They show a remarkable similarity to the indicator functions for the reverse cumulative histogram. First, they come with different sensitivities. Second, they exhibit an independence on the input strength: once the threshold and the following transition range is exceeded the output remains constant and does no longer increase when the input level is increased.

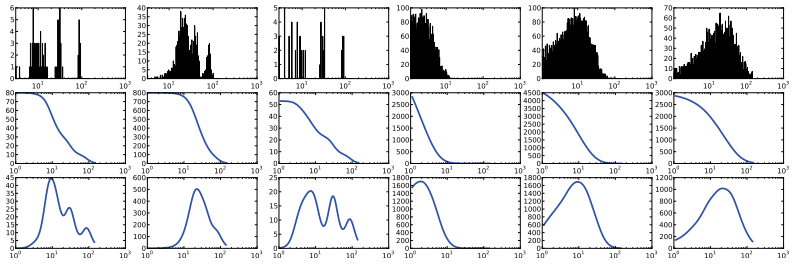


Figure 3. Neurobiological computation of a reverse cumulative histogram. The upper row shows several examples of input probability distributions. The second row shows the corresponding reverse cumulative histograms computed by a dense set of simulated neurons. The third row shows the estimated probability distributions as derived from the neural representation by use of Eq. (3).

How does all this relate to visual cortex? Has the architecture shown in Fig. 2a any neurobiological plausibility? The final summation stage is no problem since the most basic capability of neurons is computation of a linear sum of their inputs. But how about the indicator functions? They have two special properties: First, the indicator functions come with different *sensitivities*. An individual function does only generate a non-zero output if the input e exceeds a certain level, a kind of threshold, which determines the sensitivity of the element $e^{(i)}$ in Eq. (2) and Fig. 1c. To cover the complete range of values, different functions with different sensitivities are needed (Fig. 2a). Second, the indicator functions exhibit a certain *independence of the input level*. Once the input is clearly larger than the threshold, the output remains constant (Fig. 1c).

Do we know of neurons which have such properties, a range of different sensitivities, and a certain independence of the input strength? Indeed, cortical gain control (or normalization), as first described in early visual cortex (e.g. [22]) but now believed to exist throughout the brain [23], yields exactly these properties. Gain-controlled neurons (Fig. 2b) exhibit a remarkable similarity to the indicator functions used to compute the reverse cumulative histogram, since they (i) come with different sensitivities, and (ii) provide an independence of the input strength in certain response ranges.

The computation of a reverse cumulative histogram thus is well in reach of the cortex. We only have to modify the architecture of Fig. 2a by the smoother response functions of cortical neurons. The information about a probability distribution available to the visual cortex is illustrated in Fig. 3. The reconstructed distributions, as estimated from the neural reverse cumulative histograms, are a kind of Parzen-windowed (lowpass-filtered) versions of the original distributions.

2. Neural Implementation of Auto- and Cross-Correlation Functions

A key feature of the recent statistical summary approach to peripheral vision [4,6,24,16] is the usage of auto- and cross-correlation functions. These functions are defined as

$$h(i) = \frac{1}{N} \sum_{k=-N/2+1}^{N/2} e(k) \circ g(i+k), \tag{4}$$

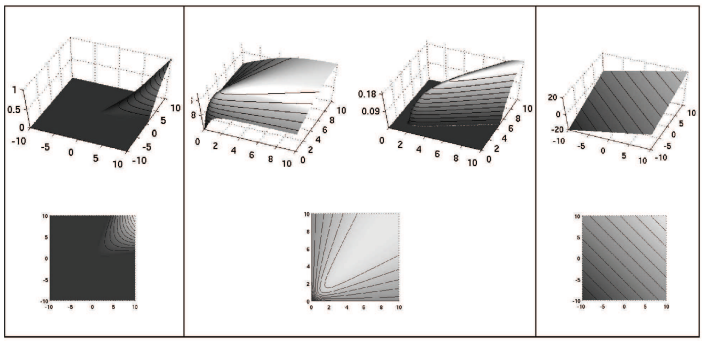


Figure 4. Different types of AND-like functions. Each function is of the type $g_k = g(s_i, s_j)$, i.e. assigns an output value to each combination of the two input values. The upper row shows the functions as surface plots, the lower row as iso-response curves. Left: Mathematical multiplication of two inputs. Center: AND-like combinations that can be obtained by use of cortical gain control (normalization). The upper left figure shows the classical gain control without additional threshold. The upper right figure shows the same mechanism with an additional threshold. This results in a full-fledged AND with a definite zero response in case that only one of the two inputs is active. Right: The linear sum of the two input values for comparison purposes.

where autocorrelation results if $e(k) = g(k)$ and where $\circ$ indicates multiplication. With respect to their neural computation, the outer summation is no problem, but the crucial function is the *nonlinear multiplicative interaction between two variables*. A neural implementation could make use of the Babylonian trick $ab = \frac{1}{4}[(a+b)^2 - (a-b)^2]$ [25,26,27], but this requires two or more neurons for the computation and thus far there is neither evidence for such a systematic pairing of neurons nor for actual multiplicative interactions in the visual cortex. However, exact multiplication is not the key factor: a reasonable statistical measure merely requires provision of a matching function such that $e(k)$ and $g(i+k)$ generate a large contribution to the autocorrelation function if they are similar, and a small contribution if they are dissimilar. For this, it is sufficient to provide a neural operation which is AND-like [27,28]. Surprisingly, such an AND-like operation can be achieved by the very same neural hardware as used before, the cortical gain control mechanism, as shown in [28]. Cortical gain control [22,29] applied to two different features $s_i(x,y)$ and $s_j(x,y)$ can be written as

$$g_k(x,y) = g(s_i(x,y), s_j(x,y)) := \max\left(0, \frac{s_i + s_j}{(\sqrt{s_i^2 + s_j^2} + \varepsilon)\sqrt{2}} - \Theta\right) \tag{5}$$

where $k = k(i,j)$, ε is a constant which controls the steepness of the response and Θ is a threshold. The resulting nonlinear combination is comparable with an AND-like operation of two features and causes a substantial nonlinear increase of the neural selectivity, as illustrated in Fig. 4.

Of course there will be differences between a formal autocorrelation function and the neurobiological version, but the essential feature, the signaling of good matches in dependence of the relative shifts will be preserved (Fig. 5).

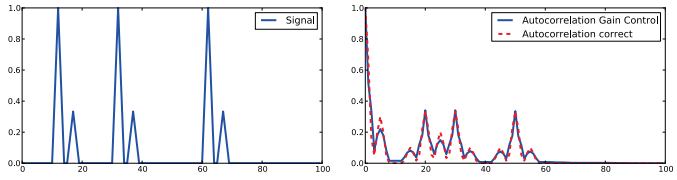


Figure 5. Mathematical and neurobiological autocorrelation functions. (a) shows a test input and (b) the corresponding mathematical (red dotted) and neurobiological (blue) autocorrelation function.

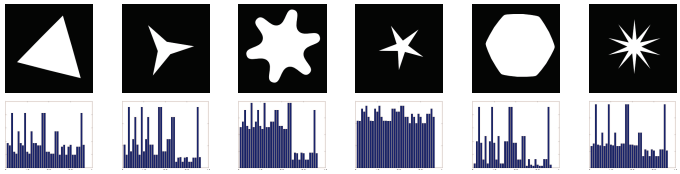


Figure 6. Different shapes and the corresponding integral features. We used parameter combinations of six different orientations $\theta_i = (i-1)\pi/6$, $i = 1,\ldots,6$, and four different scales $r_i = 2^{-i}$, $i = 1,\ldots,4$. The radial half-bandwidth was set to $f_{r,h} = \frac{1}{3}r$ and the angular half-bandwidth was constant with $f_{\theta,h} = \pi/12$. Each parameter combination creates pairs of variables for each x,y-position which are AND-combined by the gain control mechanism described in Eq. (5) as $g_k(x,y) = g(s_i(x,y), s_j(x,y))$.

3. Figural Properties from Integrals

We extracted different features $s_{r,\theta}$ from the image luminance function $l = l(x,y)$ by applying a Gabor-like filter operation $s_{r,\theta}(x,y) = (l * \mathscr{F}^{-1}(H_{r,\theta}))(x,y)$ where $\mathscr{F}^{-1}$ denotes the inverse Fourier transformation and the filter kernel $H_{r,\theta}$ is defined in the spectral space. We distinguish two cases (even and odd) which can be seen in the following definition in polar coordinates:

$$H^{even}_{r,\theta}(f_r, f_\theta) := \begin{cases} \cos^2\left(\frac{\pi}{2}\frac{f_r - r}{2f_{r,h}}\right)\cos^2\left(\frac{\pi}{2}\frac{f_\theta - \theta}{2f_{\theta,h}}\right) & ,(f_r,f_\theta) \in \Omega_{r,\theta} \\ 0 & ,\text{else,} \end{cases}$$

with $\Omega_{r,\theta} := \{(f_r,f_\theta) | f_r \in [r-2f_{r,h}, r+2f_{r,h}] \wedge f_\theta \in [\theta - 2f_{\theta,h}, \theta + 2f_{\theta,h}] \cap [\theta + \pi - 2f_{\theta,h}, \theta + \pi + 2f_{\theta,h}]\}$, where $f_{r,h}$ denotes the half-bandwidth in radial direction and $f_{\theta,h}$ denotes the half-bandwidth in angular direction. $H^{odd}_{r,\theta}$ is defined as the Hilbert transformed even symmetric filter kernel.

Various AND combinations of these oriented features (see caption Fig. 6) are obtained by the gain-control mechanism described in Eq. (5). The integration over the whole domain results in *global* features $F_k := \int_{\mathbb{R}^2} g_k(x,y)\ d(x,y)$ which capture basic shape properties (Fig. 6).

4. Numerosity and Topology

One of the most fundamental and abstract ensemble properties is the number of elements of a set. Recent evidence (see Introduction) raised the question at which cortical level

the underlying computations are performed. In this processing, a high degree of invariance has to be achieved, since numerosity can be recognized largely independent of other properties like size, shape and positioning of elements. Models which address this question in a neurobiologically plausible fashion, starting from individual pixels or neural receptors instead of an abstract type of input, are rare. To our knowledge, the first approach in this direction has been made in [30]. A widely known model [31] has a shape-invariant mapping to number which is based on linear DOG filters of different sizes, which substantially limits the invariance properties. A more recent model is based on unsupervised learning but has only employed moderate shape variations [32]. In [30] we suggested that the necessary invariance properties may be obtained by use of a theorem which connects local measurements of the differential geometry of the image surface with global topological properties [30,33]. In the following we will build upon this concept.

The key factor of our approach is a relation between surface properties and a topological invariant as described by the famous Gauss-Bonnet theorem. In order to apply this to the image luminance function $l = l(x,y)$ we interpret this function as a surface $S := \{(x,y,z) \in \mathbb{R}^3 | (x,y) \in \Omega, z = l(x,y)\}$ in three-dimensional real space. We then apply the formula for the Gaussian curvature

$$K(x,y) = \frac{l_{xx}(x,y) l_{yy}(x,y) - l_{xy}(x,y)^2}{(1 + l_x(x,y)^2 + l_y(x,y)^2)^2}, \tag{6}$$

where subscript denotes the differentiation in the respective direction (e.g. $l_{xy} = \frac{\partial^2 l}{\partial x \partial y}$). The numerator of (6) can also be written as $D = \lambda_1 \lambda_2$ where $\lambda_{1,2}$ are the eigenvalues of the Hessian matrix of the luminance function $l(x,y)$ which represent the partial second derivatives in the principal directions. The values and signs of the eigenvalues give us the information about the shape of the luminance surface S in each point, whether it is elliptic, hyperbolic, parabolic, or planar. Since Gaussian curvature results from the multiplication of the second derivatives $\lambda_{1,2}$ it is zero for the latter two cases. It has been shown that this measure can be generalized in various ways, in particular towards the use of neurophysiologically realistic Gabor-like filters instead of the derivatives [27,30]. The crucial point, however, is the need for *AND combinations of oriented features* [27,30] which can be obtained as before by the neural mechanism of cortical gain control [28].

The following corollary from the Gauss-Bonnet theorem is the basis for the invariance properties in the context of numerosity.

Corollary 4.1 *Let $S \subset \mathbb{R}^3$ be a closed two-dimensional Riemannian manifold. Then*

$$\int_S K \, dA = 4\pi(1-g) \tag{7}$$

where K is the Gaussian curvature and g is the genus of the surface S.

We consider the special case where the luminance function consists of multiple objects (polyhedra with orthogonal corners) with constant luminance level. We compare the surface of this luminance function to the surface of a cuboid with holes that are shaped like the polyhedra. The trick is that the latter surface has a genus which is determined by the number of holes in the cuboid and which can be determined by the integration of the local curvature according to Eq. (7). If we can find the corresponding contributions of

the integral on the image surface, we can use this integral to count the number of objects. We assume the corners to be locally sufficiently smooth such that the surfaces are Riemannian manifolds. The Gaussian curvature K then is zero almost everywhere except on the corners. We hence have to consider only the contributions of the corners. It turns out that these contributions can be computed from the elliptic regions only if we use different signs for upwards and downwards oriented elliptic regions. We thus introduce the following operator which distinguishes the different types of ellipticity in the luminance function. Let $\lambda_1 \geq \lambda_2$, then the operator $N(x,y) := |\min(0,\lambda_1(x,y))| - |\max(0,\lambda_2(x,y))|$ is always zero if the surface is hyperbolic and has a positive sign for positive ellipticity and a negative one for negative ellipticity. We thus can calculate the numerosity feature which has the ability of counting objects in an image by counting the holes in an imaginary cuboid as follows:

$$F = \int_\Omega \frac{N(x,y)}{(1+l_x(x,y)^2+l_y(x,y)^2)^{\frac{3}{2}}}\, d(x,y). \tag{8}$$

The crucial feature of this measure are contributions of fixed size and with appropriate signs from the corners. The denominator can thus be replaced by a neural gain control mechanism and an appropriate renormalization. For the implementation here we use a shortcut which gives us straight access to the eigenvalues. The numerator $D(x,y)$ of (6) can be rewritten as

$$D(x,y) = l_{xx}l_{yy} - \frac{1}{4}(l_{uu}-l_{vv})^2 = \frac{1}{4}[(l_{xx}+l_{yy})^2 - \underbrace{((l_{xx}-l_{yy})^2+(l_{uu}-l_{vv})^2)}_{=:\varepsilon^2}] = \frac{1}{4}(\Delta l^2 - \varepsilon^2) \tag{9}$$

with $u := x\cos(\pi/4) + y\sin(\pi/4)$ and $v := -x\sin(\pi/4) + y\cos(\pi/4)$. The eigenvalues then are $\lambda_{1,2} = \frac{1}{2}(\Delta l \pm |\varepsilon|)$ and we can directly use them to compute $N(x,y)$. Application of this computation to a number of test images is shown in Fig. 7.

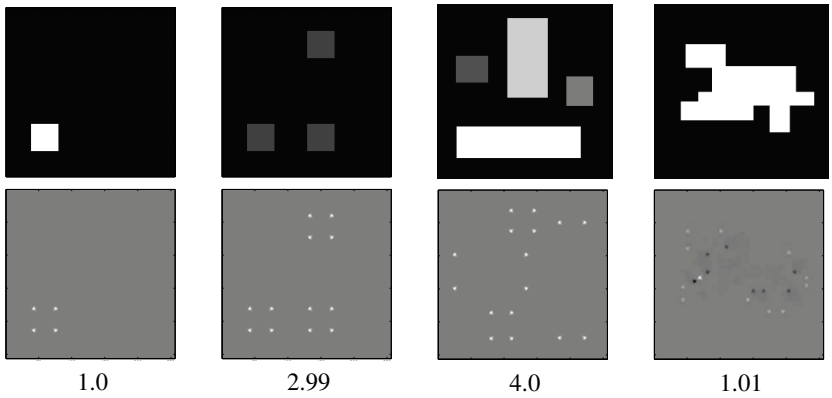


Figure 7. Based on a close relation to topological invariants the spatial integration of local curvature features can yield highly invariant numerosity estimates. The numerical values in the last row are the normalized integrals of the filter outputs (middle row).

5. Conclusion

Recent evidence shows that ensemble properties play an important role in perception and cognition. In this paper, we have investigated by which neural operations and on which processing level statistical ensemble properties can be computed by the cortex. Computation of a probability distribution requires *indicator functions* with different sensitivities, and our reinterpretation of cortical gain control suggests that this could be a basic function of this neural mechanism. The second potential of cortical gain control is the computation of *AND-like feature combinations*. Together with the linear summation capabilities of neurons this enables the computation of powerful invariants and summary features. We have repeatedly argued that AND-like feature combinations are essential for our understanding of the visual system [27,30,34,35,36,28]. The increased selectivity of nonlinear AND operators, as compared to their linear counterparts, is a prerequisite for the usefulness of integrals over the respective responses [30,28]. We have shown that such integrals of AND features are relevant for the understanding of texture perception [37], of numerosity estimation [30], and of invariance in general [28]. Recently, integrals over AND-like feature combinations in form of auto- and cross-correlation functions have been suggested for the understanding of peripheral vision [4,16,17].

A somewhat surprising point is that linear summation and cortical gain control, two widely accepted properties of cortical neurons, are the only requirements for the computation of ensemble properties. These functions are already available at early stages of the cortex, but also in other cortical areas [23]. The computation of ensemble properties may thus be an ubiquitous phenomenon in the cortex.

Acknowledgement

This work was supported by DFG, SFB/TR8 Spatial Cognition, project A5-[ActionSpace].

References

[1] S. C. Dakin and R. J. Watt. The computation of orientation statistics from visual texture. *Vision Res*, 37(22):3181–3192, 1997.
[2] D. Ariely. Seeing Sets: Representation by Statistical Properties. *Psychol Sci*, 12(2):157–162, 2001.
[3] Lin Chen. The topological approach to perceptual organization. *Visual Cognition*, 12(4):553–637, 2005.
[4] B. Balas, L. Nakano, and R. Rosenholtz. A summary-statistic representation in peripheral vision explains visual crowding. *J Vis*, 9(12):13.1–18, 2009.
[5] G. A. Alvarez. Representing multiple objects as an ensemble enhances visual cognition. *Trends Cog Sci*, 15(3):122–31, 2011.
[6] R. Rosenholtz, J. Huang, and K. Ehinger. Rethinking the role of top-down attention in vision: effects attributable to a lossy representation in peripheral vision. *Front Psychol*, 3:13, 2012.
[7] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996.
[8] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
[9] E. M. Brannon. The representation of numerical magnitude. *Curr Opin Neurobiol*, 16(2):222–9, 2006.
[10] J. Hegde and D.J. Felleman. Reappraising the Functional Implications of the Primate Visual Anatomical Hierarchy. *The Neuroscientist*, 13(5):416–421, 2007.
[11] D. Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977.
[12] M. R. Greene and A. Oliva. The Briefest of Glances. *Psychol Sci*, 20(4):464–472, 2009.

[13] J. Hegdé. Time course of visual perception: coarse-to-fine processing and beyond. *Prog Neurobiol*, 84(4):405–39, 2008.
[14] M. Fabre-Thorpe. The characteristics and limits of rapid visual categorization. *Front Psychol*, 2:243, 2011.
[15] M. J-M Macé, O. R. Joubert, J-L. Nespoulous, and M. Fabre-Thorpe. The time-course of visual categorizations: you spot the animal faster than the bird. *PloS one*, 4(6):e5927, 2009.
[16] J. Freeman and E. P. Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011.
[17] H. Strasburger, I. Rentschler, and M. Jüttner. Peripheral vision and pattern recognition: a review. *J Vis*, 11(5):13, 2011.
[18] A. Nieder, D. J. Freedman, and E. K. Miller. Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, 297(5587):1708–11, 2002.
[19] H. J. Gross, M. Pahl, A. Si, H. Zhu, J. Tautz, and S. Zhang. Number-based visual generalisation in the honeybee. *PloS one*, 4(1):e4263, 2009.
[20] J. D. Roitman, E. M. Brannon, and M. L. Platt. Monotonic coding of numerosity in macaque lateral intraparietal area. *PLoS biology*, 5(8):e208, 2007.
[21] J. Ross and D. C. Burr. Vision senses number directly. *Journal of vision*, 10(2):10.1–8, 2010.
[22] D. G. Albrecht and D. B. Hamilton. Striate cortex of monkey and cat: contrast response function. *J Neurophysiol*, 48(1):217–237, Jul 1982.
[23] M. Carandini and D. J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neurosci*, 13:51–62, Jul 2012.
[24] R. Rosenholtz, J. Huang, A. Raj, B. J. Balas, and L. Ilie. A summary statistic representation in peripheral vision explains visual search. *J Vis*, 12(4):1–17, 2012.
[25] H.L. Resnikoff and R.O. Wells. *Mathematics in Civilization*. Popular Science Series. Dover, 1984.
[26] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2):284–99, 1985.
[27] C. Zetzsche and E. Barth. Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision Res*, 30(7):1111–1117, 1990.
[28] C. Zetzsche and U. Nuding. Nonlinear and higher-order approaches to the encoding of natural scenes. *Network*, 16(2–3):191–221, 2005.
[29] D.J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neurosci*, 9(2):181–198, 1992.
[30] C. Zetzsche and E. Barth. Image surface predicates and the neural encoding of two-dimensional signal variations. In B. E. Rogowitz and Jan P. A., editors, *Proc SPIE*, volume 1249, pages 160–177, 1990.
[31] S. Dehaene and J. P. Changeux. Development of elementary numerical abilities: a neuronal model. *J. Cogn. Neurosci.*, 5(4):390–407, 1993.
[32] I. Stoianov and M. Zorzi. Emergence of a 'visual number sense' in hierarchical generative models. *Nat Neurosci*, 15(2):194–6, 2012.
[33] M. Ferraro E. Barth and C. Zetzsche. Global topological properties of images derived from local curvature features. In L. P. Cordella C. Arcelli and G. Sanniti di Baja, editors, *Visual Form 2001. Lecture Notes in Computer Science*, pages 285–294, 2001.
[34] C. Zetzsche, E. Barth, and B. Wegmann. The importance of intrinsically two-dimensional image features in biological vision and picture coding. In A. B. Watson, editor, *Digital images and human vision*, pages 109–138. MIT Press, Cambridge, MA, 1993.
[35] G. Krieger and C. Zetzsche. Nonlinear image operators for the evaluation of local intrinsic dimensionality. *IEEE Transactions Image Processing*, 5:1026–1042, 1996.
[36] C. Zetzsche and G. Krieger. Nonlinear mechanisms and higher-order statistics in biological vision and electronic image processing: review and perspectives. *J Electronic Imaging*, 10(1):56–99, 2001.
[37] E. Barth, C. Zetzsche, and I. Rentschler. Intrinsic 2D features as textons. *J. Opt. Soc. Am. A*, 15(7):1723—1732, 1998.

3 Role of curvature i2D-features in numerical cognition

The usability of $i2D$-operators to obtain desired kinds of selectivity is shown in Section 2.3. Beside the selectivity property, it has been shown in particular in Section 2.4 that the use of $i2D$-operators extended by an integration operation can be used to describe certain invariant properties. In this section, the main focus lies on the invariant property *numerosity*. In simple words, numerosity deals with the number of objects: "How many objects did you see?" is one typical question addressing the numerical cognition of humans. Or the questions "Where is more food?" or "Where are less raptors?", which are essential for survival, require the ability to differentiate between different quantities and to determine their relations, e.g. less or more. The main question considered in this chapter is how the number of objects can be estimated by vision. And furthermore, what kind of operations are required to compute an estimate. The interesting point is that there exists a mathematical relation between two kinds of $i2D$-operators and a topological invariant which can be used to give an estimate for the number of objects represented in a scene. The Gauss-Bonnet theorem relates the Gaussian curvature and the geodesic curvature of curved surfaces to the Euler characteristic of topological spaces. In Section 3.2 both operators, Gaussian curvature and geodesic curvature, are derived from differential geometry and it is shown that these operators are $i2D$-operators. Section 3.3 provides a first algorithmic solution to the problem which is then extended and related to behavioral findings in Section 3.4. Before the technical consideration starts, a brief overview about numerical cognition is given in the following Section 3.1 to extend the descriptions in the articles presented in Sections 3.3 and 3.4.

3.1 Related work

Numerical cognition goes back to first investigations by Jevons in 1871 [38]. In a self-experiment Jevons tried to estimate the number of black beans which were presented for a small time period making sequential counting impossible. The number of beans varied from 3 to 15. He observed that he did the estimation very well in the determination of numbers up to a cardinality of 4. For higher numbers a systematic error occurred. The estimation performance decreased with an increase in the number of beans.

In numerical cognition it is important to distinguish between three types of cognition. The first distinction can be done by the amount of time which is available to the subject. If one has unlimited time, one can just sequentially count where the performance is even very well for the whole number line. This is *sequential counting* which requires knowledge about what is meant by "one", i.e. a discretized representation of number. If one has strictly limited time, two kinds of cognition are reported. Up to a number of 4 the answer can be given nearly immediately without any significant errors. This phenomenon is called *subitizing* [40]. The last effect, which was also reported by Jevons, is the *numerosity estimation*. If the time of

stimulus presentation is strictly limited and the number is higher than 4, the subjects make systematic errors. These systems are commonly assumed to build two distinct subsystems. One accounts for subitzing and the other is responsible for the numerosity estimation [22].

The numerosity estimation, which is also referred to as "number sense" [17], is the determination of the cardinality of objects and is independent of the kind of visual presentation influenced by quantities like the cumulative area, object size, or object density. It also has been shown that numerosity estimation is an ability which is not restricted to humans. There exist various examples for the successful numerosity estimation performed by animals: Koehler did experiments with birds, like pigeons and jackdaws [44]. The birds were trained to pick a specific number of grains. In an experiment a larger number of grains were available to the birds. In the majority of trials they picked up the number they were trained to. The successful trials were reported up to a number of 6 grains. This shows that the estimation of cardinality does not require a symbolic number representation like humans have access to, e.g. the Arabic digits. Other species like rhesus monkeys are also able to estimate the numerosity and even more they are able to distinguish between different cardinalities [32]. Gray parrots are also able to perform simple arithmetic tasks with respect to a total number of 6 [61].

The estimation of numerosity was not only reported for human adults [83] but it was also shown that even six month old infants were able to do a number distinction task [85]. Furthermore, infants are also able to perform simple arithmetic tasks like $1+1=2$ or $2-1=1$ [84]. The development of children was also investigated with respect to the development of mathematical competence and their ability in numerosity estimation. In experiments with 14-year-old children it was found out that the mathematical ability is correlated with the acuity in numerosity estimation [29].

The numerosity estimation is not only restricted to the estimation of the cardinality of objects. It was also discovered that the quantity of physical properties like sound volume, space, and time shows similar characteristics [8]. This is in line with the opinion that there must be a generalized system for magnitude estimation [24, 79].

Two important phenomena which characterize the numerosity estimation are reported in the literature. The distance effect and the size effect [18]. The distance effect describes the error behavior in the comparison of two cardinalities. The smaller the distance between the two cardinalities, the more errors are done by the observer. Or equivalently, the larger the numerical distance, the easier two cardinalities can be distinguished. This effect does not only occur in the comparison of visually presented cardinalities. It also occurs in the comparison of Arabic digits which allows the conclusion that the Arabic digits are transformed into the approximative representation for comparison. The size effect states that with constant numerical distance the distinction between two cardinalities becomes more difficult with higher absolute cardinalities. For example, the numbers 4 and 5 are easier to distinguish than the numbers 8 and 9. Both effects can be explained by the Weber-Fechner law.

Weber reported that two heavy weights have to differ more than two lighter weights to observe a difference in weight [81]. This corresponds to the reported size effect in numerosity estimation. Weber formulated the following law: In order to be able to distinguish two sensory stimuli, they must differ at least by a fraction k of the stimulus intensity I. The difference in stimulus intensity ΔI thus becomes

$$\Delta I = kI. \tag{3.1}$$

The fraction $k = \Delta I/I$ is also referred to as the *Weber fraction*.

In 1860, Fechner extended Weber's law that it is related to the perceived stimulus intensity explicitly [21]. The larger the stimulus intensity, the larger the difference in stimulus intensity must be to cause equal differences in perceived stimulus intensity. Fechner observed that the perceived stimulus intensity S is a logarithmic function of the stimulus intensity I with a constant factor k, i.e.

$$S = k \ln(I). \tag{3.2}$$

The relation extended by a constant summand is also referred to as *Weber-Fechner law*. This law corresponds to the distance effect in numerosity estimation. It states that an exponential increase in stimulus intensity is perceived linearly only.

The reported effects allow two possible explanations for the mental representation of number [22]. On the one hand, the numerosity can be represented linearly. Then the uncertainty in the belief in a specific numerosity increases with the absolute cardinality. On the other hand, the number line is represented logarithmically with constant uncertainty at all numerosities. Neural findings give rising evidence for the logarithmic representation [55].

In order to avoid repetition the reader is referred to Sections 3.3 and 3.4 as further related works and background information are considered within the context of the presented articles.

3.2 Mathematical preliminaries

A coarse introduction of the concepts of differential geometry is given. These concepts are applied to derive the curvature operators used to develop a model for numerosity estimation. And finally it is shown that these operators belong to the class of $i2D$-operators. The intention of this section is not to present the complete theory. For this purpose, the reader is referred to standard differential geometry textbooks, e.g. [5, 26, 48]. The parts which are important for later considerations are presented and applied to one "case of interest". The studied case is a curved surface in the three-dimensional space which is defined by a function in its z-coordinate and parametrized by its x- and y-coordinates. Let $l : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ be a function of the class $C^3(\Omega)$, i.e. three times continuously differentiable. The surface of interest is then defined by $(x, y, l(x, y))^T$, $(x, y)^T \in \Omega$. But before we consider surfaces in the three dimensional

space, we start with the consideration of space curves. A space curve p is a subset of the three-dimensional space which is defined by a parametrization, e.g. $p(t) = (x(t), y(t), z(t))^T$, $t \in I \subset \mathbb{R}$. A particular case, which is of major interest for later consideration with respect to numerosity estimation, is a space curve with constant height which lies on the surface $(x, y, l(x, y))^T$. All presented theoretical quantities are derived for this case. In each point on a space curve three important characteristic vectors can be determined: the tangent vector, the principal normal vector, and the binormal vector.

Definition 3.1 (Tangent vector)**.** Let $X : I \subset \mathbb{R} \to \mathbb{R}^3$ be a differentiable space curve. Then the *tangent vector* T is defined by

$$T(t) = \frac{X'(t)}{\|X'(t)\|}. \tag{3.3}$$

Corollary 3.2. *If a space curve is given implicitly by $l(x(t), y(t)) = c$ with the parametrization $X(t) = (x(t), y(t), l(x(t), y(t)))^T$, then the tangent vector is given by*

$$T(t) = \begin{pmatrix} l_y \\ -l_x \\ 0 \end{pmatrix} (l_x^2 + l_y^2)^{-1/2}. \tag{3.4}$$

Proof. Using the constraint $l(x(t), y(t)) = c$, we get

$$X'(t) = \begin{pmatrix} x' \\ y' \\ x'l_x(x, y) + y'l_y(x, y) \end{pmatrix}. \tag{3.5}$$

The definition of the curve yields

$$x'l_x(x, y) + y'l_y(x, y) = 0. \tag{3.6}$$

We thus get

$$T(t) = \begin{pmatrix} x' \\ -\frac{l_x(x,y)}{l_y(x,y)}x' \\ 0 \end{pmatrix} (x'^2 + \frac{l_x(x, y)^2}{l_y(x, y)^2}x'^2)^{-1/2} = \begin{pmatrix} l_y \\ -l_x \\ 0 \end{pmatrix} (l_x^2 + l_y^2)^{-1/2}. \tag{3.7}$$

□

Definition 3.3 (Principal normal vector)**.** At any point with $T(t) \neq 0$ the *principal normal*

vector is defined by

$$P(t) := \frac{T'}{\|T'\|}. \tag{3.8}$$

Corollary 3.4. *For the implicitly given curve ($l(x(t), y(t)) = c$) the principal normal is given by*

$$P(t) = -\frac{1}{(l_x^2 + l_y^2)^{1/2}} \begin{pmatrix} l_x \\ l_y \\ 0 \end{pmatrix}. \tag{3.9}$$

Proof. The tangent vector for a general space curve $(x(t), y(t), c)^T$ is given by

$$T(t) = \frac{1}{(x'^2 + y'^2)^{1/2}} \begin{pmatrix} x' \\ y' \\ 0 \end{pmatrix}. \tag{3.10}$$

The derivative of the tangent vector thus becomes

$$\begin{aligned} T'(t) &= \frac{1}{(x'^2 + y'^2)^{1/2}} \begin{pmatrix} x'' \\ y'' \\ 0 \end{pmatrix} - \frac{x'x'' + y'y''}{(x'^2 + y'^2)^{3/2}} \begin{pmatrix} x' \\ y' \\ 0 \end{pmatrix} \\ &= \frac{1}{(x'^2 + y'^2)^{3/2}} \left[\begin{pmatrix} x''x'^2 + x''y'^2 \\ y''x'^2 + y''y'^2 \\ 0 \end{pmatrix} - \begin{pmatrix} x''x'^2 + y''x'y' \\ x''x'y' + y''y'^2 \\ 0 \end{pmatrix} \right] \\ &= \frac{1}{(x'^2 + y'^2)^{3/2}} \begin{pmatrix} x''y'^2 - y''x'y' \\ y''x'^2 - x''x'y' \\ 0 \end{pmatrix} \\ &= \frac{x''y' - y''x'}{(x'^2 + y'^2)^{3/2}} \begin{pmatrix} y' \\ -x' \\ 0 \end{pmatrix}. \end{aligned} \tag{3.11}$$

With this result the principal normal vector is given by

$$P(t) = \frac{1}{(x'^2 + y'^2)^{1/2}} \begin{pmatrix} y' \\ -x' \\ 0 \end{pmatrix}. \tag{3.12}$$

With $x' = l_y$ and $y' = -l_x$ follows

$$P(t) = -\frac{1}{(l_x^2 + l_y^2)^{1/2}} \begin{pmatrix} l_x \\ l_y \\ 0 \end{pmatrix}.$$

□

Definition 3.5 (Binormal vector)**.** The *binormal vector* B is defined by the vector product of th orthogonal vectors T and P, i.e.

$$B = T \times P. \tag{3.13}$$

Remark 3.6. For the implicitly defined curve X the binormal vector is

$$B = -\frac{1}{l_x^2 + l_y^2} \left[\begin{pmatrix} l_y \\ -l_x \\ 0 \end{pmatrix} \times \begin{pmatrix} l_x \\ l_y \\ 0 \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}. \tag{3.14}$$

Definition 3.7. The triple of vectors (T, P, B) associated to each point of a continuous space curve is called the *Frenet frame*.

With the knowledge of the Frenet frame specific properties of the curve can be determined. One example is the curvature of a space curve which is used subsequently to determine the desired operators.

Definition 3.8 (Curvature)**.** Let $X \in C(I)^2$ be a parametrized curve. The curvature $\kappa : I \to \mathbb{R}^+$ is defined as

$$\kappa(t) = \frac{\|T'(t)\|}{s'(t)} \tag{3.15}$$

where $s'(t) = \|X'(t)\|$. By multiplication both sides with $T'(t)$, it follows

$$T'(t) = s'(t)\kappa(t)P(t). \tag{3.16}$$

Corollary 3.9. *If a space curve is given implicitly by $l(x(t), y(t)) = c$ with the parametrization $X(t) = (x(t), y(t), l(x(t), y(t)))^T$, then the curvature is given by*

$$\kappa(t) = \frac{l_x^2 l_{yy} + l_y^2 l_{xx} - 2 l_x l_y l_{xy}}{(l_x^2 + l_y^2)^{3/2}}. \tag{3.17}$$

Proof. Using T' from Equation (3.11) and using P from Equation (3.12) in Equation (3.16)

yield

$$\frac{x''y' - y''x'}{(x'^2 + y'^2)^{3/2}} \begin{pmatrix} y' \\ -x' \\ 0 \end{pmatrix} = (x'^2 + y'^2)^{1/2} \kappa \frac{1}{(x'^2 + y'^2)^{1/2}} \begin{pmatrix} y' \\ -x' \\ 0 \end{pmatrix}.$$

With $x' = l_y$, $y' = -l_x$, $x'' = l_{yx}l_y - l_{yy}l_x$, and $y'' = -(l_{xx}l_y - l_{xy}l_x)$ follows

$$\begin{aligned} \kappa(t) &= \frac{x''y' - y''x'}{(x'^2 + y'^2)^{3/2}} \\ &= \frac{1}{(l_x^2 + l_y^2)^{3/2}}(x''y' - y''x') \\ &= \frac{1}{(l_x^2 + l_y^2)^{3/2}} \left(l_x(l_{yx}l_y - l_{yy}l_x) - l_y(l_{xx}l_y - l_{xy}l_x)\right) \\ &= \frac{l_x^2 l_{yy} + l_y^2 l_{xx} - 2l_x l_y l_{xy}}{(l_x^2 + l_y^2)^{3/2}}. \end{aligned}$$

□

The curvature is a property of the space curve but if the the space curve is assumed to lie on a parametrized surface, further quantities can be determined with the aid of the curvature of the space curve. In order to deal with parametrized surfaces, this term is specified in the following definition.

Definition 3.10. $S \subset \mathbb{R}^3$ is called a *parametrized surface* if

$$\forall p \in S : \exists U \subset \mathbb{R}^2, V(p) \subset \mathbb{R}^3, X \in C(U, \mathbb{R}^3) : X(U) = V \cap S \tag{3.18}$$

where $V(p)$ is an open neighborhood of p. X is called a parametrization.

The parametrization plays an important role as it enables one to determine properties of the described objects, e.g. surfaces or space curves on a surface. The previously described case of interest makes use of a specific kind of parametrization. The described surface is given by the parametrization $X(u, v) = (u, v, l(u, v))^T$. An important space in the study of curved surfaces is the tangent space which is defined with the aid of tangent vectors as follows.

Definition 3.11 (Tangent space)**.** Let $S \subset \mathbb{R}^3$ be a parametrized surface, $p \in S$, and M is the set of tangent vectors to S in p. If $\dim(M) = 2$, M is called the *tangent space* and it is denoted by T_pS. Then the set $\{p + v | v \in T_pS\}$ is called the *tangent plane*.

This formal definition of the tangent space does not take a specific parametrization into account. How the tangent space and the tangent plane is determined by a given parametrization and more importantly whether the tangent space exists, is considered in the following proposition.

Proposition 3.12. *Let $S \subset \mathbb{R}^3$ be a parametrized surface with the parametrization $X : U \subset \mathbb{R}^2 \to \mathbb{R}^3$. Assume X is injective and $p = X(u_0, v_0)$. Then the set of tangent vectors forms a subspace of $\mathbb{R}^3$ if and only if*

$$X_u(u_0, v_0) \times X_v(u_0, v_0) \neq 0. \tag{3.19}$$

The corresponding tangent plane is then defined by

$$\{x \in \mathbb{R}^3 | (x - p) \cdot (X_u(u_0, v_0) \times X_v(u_0, v_0)) = 0\}. \tag{3.20}$$

Further considerations require more regularity with respect to the parametrization which is provided by a regular surface as defined in the following.

Definition 3.13. $S \subset \mathbb{R}^3$ is a *regular surface* if for all $p \in S$ exists an open set $U \subset \mathbb{R}^2$, an open neighborhood $V(p) \subset \mathbb{R}^3$, and a surjective continuous function $X : U \to S \cap V(p)$ such that

(i) X is differentiable,
(ii) X is a homeomorphism, i.e. X is continuous and X^{-1} exists and is also continuous,
(iii) X satisfies the regularity condition ($\ker(DX(u, v)) = \{0\}$).

The following proposition is an important result and guarantees in the "case of interest" that the surface is a regular surface.

Proposition 3.14. *Let $U \subset \mathbb{R}^2$ be open. Then if a function $f : U \to \mathbb{R}$ is differentiable, the set*

$$S = \{(u, v, f(u, v)) \in \mathbb{R}^3 | (u, v) \in U\} \tag{3.21}$$

is a regular surface.

Similar to the principal normal vector of a space curve, the unit normal vector of a surface is defined by the vector which is orthogonal to the tangent plane.

Definition 3.15 (Unit normal vector)**.** Let $S \subset \mathbb{R}^3$ be a regular surface, $p \in S$, X a parametrization of a neighborhood of p, and $p = X(q)$. Then the *unit normal vector* is defined by

$$N(q) = \frac{X_u \times X_v}{\|X_u \times X_v\|}(q). \tag{3.22}$$

The quantities of the previous definitions are determined for the studied case in the following remark. These findings play an important role in the subsequent derivation of the geodesic curvature and the Gaussian curvature.

Remark 3.16. Let $l : U \subset \mathbb{R}^2 \to \mathbb{R}$ be a differentiable function. The surface $S \subset \mathbb{R}^3$ defined by $S := \{X(x,y) = (x,y,l(x,y))^T | (x,y) \in U\}$ is a regular surface and $X : U \to \mathbb{R}^3$ is the parametrization. The tangent and unit normal vectors in $p = X(x_0, y_0)$, $p \in S$, then become

$$X_x(x,y) = \begin{pmatrix} 1 \\ 0 \\ l_x \end{pmatrix}, \qquad X_y(x,y) = \begin{pmatrix} 0 \\ 1 \\ l_y \end{pmatrix}, \quad \text{and} \quad N(x,y) = \frac{1}{(1 + l_x^2 + l_y^2)^{1/2}} \begin{pmatrix} -l_x \\ l_y \\ 1 \end{pmatrix}. \tag{3.23}$$

The first fundamental form is an important mapping defined with respect to the tangent space. With the aid of the first fundamental form, quantities of the intrinsic geometry of the surface can be computed. This means quantities with respect to the surface itself, e.g. length of curves on the surface or the area of regions on the surface. But more importantly in the context of this work is the relation to the Gaussian curvature.

Definition 3.17 (First fundamental form)**.** Let S be a regular surface. The first fundamental form $I_p(\cdot,\cdot)$ is the restriction of the usual dot product in $\mathbb{R}^3$ to the tangent plane T_pS . That means $I_p(x,y) = x \cdot y$ for all $x, y \in T_pS$ with respect to the standard basis of $\mathbb{R}^3$.

The first fundamental form is defined on the tangent space of the surface. But it remains unclear how the parametrization of the surface influences it. The following proposition gives an answer to this question.

Proposition 3.18. *Let $I_p : T_pS \times T_pS \to \mathbb{R}$ be the first fundamental form at a point p on a regular surface S. Given a regular parametrization $X : U \to \mathbb{R}^3$, the matrix associated with the first fundamental form I_p with respect to the basis $\mathcal{B} = \{X_u, X_v\}$ is*

$$g = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} = \begin{pmatrix} X_u \cdot X_u & X_u \cdot X_v \\ X_v \cdot X_u & X_v \cdot X_v \end{pmatrix}. \tag{3.24}$$

The first fundamental form can be expressed by the quadratic form $I_p(x,y) = x^T g y$ for x and y being points on the tangent plane and expressed in the basis $\mathcal{B}$.

In literature the coefficients of the matrix associated with the first fundamental form are often denoted by E, F, and G such that

$$g = \begin{pmatrix} X_u \cdot X_u & X_u \cdot X_v \\ X_v \cdot X_u & X_v \cdot X_v \end{pmatrix} =: \begin{pmatrix} E & F \\ F & G \end{pmatrix}. \tag{3.25}$$

The following formula by Brioschi relates the first and second derivatives of the coefficients of the first fundamental form to the Gaussian curvature. In particular it gives an explicit formula which enables one to determine the Gaussian curvature. The theorem and a proof can be found in [26].

Theorem 3.19 (Brioschi's formula)**.** *Let S be a regular surface with its parametrization $X : U \to \mathbb{R}^3$. Then the Gaussian curvature of X is given by*

$$K = \frac{1}{(EG - F^2)^2} \left(\begin{vmatrix} -\frac{1}{2}E_{vv} + F_{uv} - \frac{1}{2}G_{uu} & \frac{1}{2}E_u & F_u - \frac{1}{2}E_v \\ F_v - \frac{1}{2}G_u & E & F \\ \frac{1}{2}G_v & F & G \end{vmatrix} - \begin{vmatrix} 0 & \frac{1}{2}E_v & \frac{1}{2}G_u \\ \frac{1}{2}E_v & E & F \\ \frac{1}{2}G_u & F & G \end{vmatrix} \right). \tag{3.26}$$

Corollary 3.20. *Let $l : U \subset \mathbb{R}^2 \to \mathbb{R}$ be a differentiable function. The surface $S \subset \mathbb{R}^3$ defined by $S := \{X(x, y) = (x, y, l(x, y))^T | (x, y) \in U\}$ is a regular surface and $X : U \to \mathbb{R}^3$ is the parametrization. Then the Gaussian curvature is given by*

$$K = \frac{l_{xx}l_{yy} - l_{xy}^2}{(1 + l_x^2 + l_y^2)^2}. \tag{3.27}$$

Proof. By using Equation (3.23) the quantities E, F, and G become

$$E = 1 + l_x^2,\ F = l_x l_y,\ \text{and}\ G = 1 + l_y^2$$

with

$$\begin{aligned} &E_x = 2l_x l_{xx},\ E_y = 2l_x l_{xy},\ E_{yy} = 2(l_{xy}^2 + l_x l_{xyy}) \\ &F_x = l_{xx}l_y + l_x l_{yx},\ F_y = l_{xy}l_y + l_x l_{yy},\ F_{xy} = l_{xxy}l_y + l_{xx}l_{yy} + l_x l_{yxy} + l_{xy}^2 \\ &G_x = 2l_y l_{yx},\ G_y = 2l_y l_{yy},\ G_{xx} = 2(l_{xy}^2 + l_y l_{yxx}). \end{aligned}$$

With Equation (3.26) follows

$$\begin{aligned} K &= \frac{1}{(1 + l_x^2 + l_y^2)^2} \left(\begin{vmatrix} l_{xx}l_{yy} - l_{xy}^2 & l_x l_{xx} & l_y l_{xx} \\ l_x l_{yy} & 1 + l_x^2 & l_x l_y \\ l_y l_{yy} & l_x l_y & 1 + l_y^2 \end{vmatrix} - \begin{vmatrix} 0 & l_x l_{xy} & l_y l_{xy} \\ l_x l_{xy} & 1 + l_x^2 & l_x l_y \\ l_y l_{xy} & l_x l_y & 1 + l_y^2 \end{vmatrix} \right) \\ &= \frac{1}{(1 + l_x^2 + l_y^2)^2} \big[(l_{xx}l_{yy} - l_{xy}^2)(1 + l_x^2)(1 + l_y^2) + 2l_x^2 l_y^2 l_{xx}l_{yy} \\ &- l_y^2 l_{xx}l_{yy}(1 + l_x^2) - l_x^2 l_{xx}l_{yy}(1 + l_y^2) - l_x^2 l_y^2 (l_{xx}l_{yy} - l_{xy}^2) \\ &-2l_x^2 l_y^2 l_{xy}^2 + l_x^2 l_{xy}^2 (1 + l_y^2) + l_y^2 l_{xy}^2 (1 + l_x^2) \big] \\ &= \frac{l_{xx}l_{yy} - l_{xy}^2}{(1 + l_x^2 + l_y^2)^2} \underbrace{\big[(1 + l_x^2)(1 + l_y^2) - l_y^2(1 + l_x^2) - l_x^2(1 + l_y^2) + l_x^2 l_y^2 \big]}_{=1}. \end{aligned}$$

□

Remark 3.21. The definition of the Gaussian curvature is not given in this section as it re-

quires knowledge about the second fundamental form. The shortcut without this knowledge would not be possible without Brioschi's formula and Gauss' prominent *Theorema Egregium*, which states that the Gaussian curvature can be computed without knowledge of the second fundamental form, by using first and second derivatives of the coefficients of the first fundamental form only. The interested reader is referred to standard differential geometry literature [5, 26, 48]. The following is a constructive description of what is meant by Gaussian curvature. From the definition it can be derived that the Gaussian curvature K is the product of the two principal curvatures k_1 and k_2 in each point on a regular surface S, i.e.

$$K = k_1 k_2. \tag{3.28}$$

Let T_pS denote the tangent plane and N_1 and N_2 are two orthogonal planes which are both orthogonal to T_pS. The curvature k_{n_i} of the intersection curves in N_i is termed the normal curvature. If N_1 and N_2 are chosen so that the corresponding k_{n_i} become extreme values, the normal curvatures with respect to N_1 and N_2 are the principal curvatures k_1 and k_2. The product of the principal curvatures is then the Gaussian curvature.

The last quantity which is considered describes a specific type of curvature of curves on a surface. This is the geodesic curvature. Given a curved surface and a curve on the surface. Then the geodesic curvature in one point is the curvature of the curve which results from the projection of the curve to the tangent plane. The following lemma describes how this curvature can be determined.

Lemma 3.22 (Geodesic curvature)**.** *Let S be a regular surface with parametrization $X : U \subset \mathbb{R}^2 \to \mathbb{R}^3$. And let $Y : I \subset \mathbb{R} \to S$ be a space curve on S. Given the Frenet frame (T, P, B) and the unit normal vector N, the geodesic curvature can be calculated by*

$$\kappa_g = \kappa P \cdot U \tag{3.29}$$

where

$$U = N \times T. \tag{3.30}$$

Corollary 3.23. *Let $l : U \subset \mathbb{R}^2 \to \mathbb{R}$ be a differentiable function. The surface $S \subset \mathbb{R}^3$ defined by $S := \{X(x, y) = (x, y, l(x, y))^T | (x, y) \in U\}$ is a regular surface and $X : U \to \mathbb{R}^3$ is the parametrization. Furthermore a space curve on S is given implicitly by $l(x(t), y(t)) = c$ with $Y(t) = (x(t), y(t), l(x(t), y(t)))^T$. Then the geodesic curvature of Y is given by*

$$\kappa_g = \frac{l_x^2 l_{yy} + l_y^2 l_{xx} - 2 l_x l_y l_{xy}}{(l_x^2 + l_y^2)^{3/2}(1 + l_x^2 + l_y^2)^{1/2}}. \tag{3.31}$$

Proof. Given the vectors P, N, and T from Equations (3.9), (3.23), and (3.4), we get

$$U = \frac{1}{(l_x^2 + l_y^2)^{1/2}(1 + l_x^2 + l_y^2)^{1/2}} \begin{pmatrix} l_x \\ l_y \\ l_x^2 - l_y^2 \end{pmatrix}. \tag{3.32}$$

With

$$P \cdot U = \frac{1}{(1 + l_x^2 + l_y^2)^{1/2}} \tag{3.33}$$

and the curvature κ from Equation (3.17) it follows

$$\kappa_g = \frac{l_x^2 l_{yy} + l_y^2 l_{xx} - 2 l_x l_y l_{xy}}{(l_x^2 + l_y^2)^{3/2}(1 + l_x^2 + l_y^2)^{1/2}}. \tag{3.34}$$

□

We now have the Gaussian curvature and the geodesic curvature, the elementary quantities in the Gauss-Bonnet theorem, which play the key role in the subsequent model for numerosity estimation. But the relation to the concept of intrinsic dimensionality is still unclear. The following theorem makes the relation explicit as it shows that both, the Gaussian curvature and the geodesic curvature, define *i2D*-operators.

Theorem 3.24. *The operators $T_i : C^2(\Omega) \to C(\Omega)$, $i = 1, 2$, with compact $\Omega \subset \mathbb{R}^2$ defined by*

$$T_1(u)(x) := \frac{\frac{\partial^2}{\partial x_1^2} u \frac{\partial^2}{\partial x_2^2} u - (\frac{\partial^2}{\partial x_1 \partial x_2} u)^2}{(1 + (\frac{\partial}{\partial x_1} u)^2 + (\frac{\partial}{\partial x_2} u)^2)^2} \tag{3.35}$$

and

$$T_2(u)(x) := \frac{(\frac{\partial}{\partial x_1} u)^2 \frac{\partial^2}{\partial x_2^2} u + (\frac{\partial}{\partial x_2} u)^2 \frac{\partial^2}{\partial x_1^2} u - 2 \frac{\partial}{\partial x_1} u \frac{\partial}{\partial x_2} u \frac{\partial^2}{\partial x_1 \partial x_2} u}{((\frac{\partial}{\partial x_1} u)^2 + (\frac{\partial}{\partial x_2} u)^2)^{3/2}(1 + (\frac{\partial}{\partial x_1} u)^2 + (\frac{\partial}{\partial x_2} u)^2)^{1/2}} \tag{3.36}$$

are i2D-operators.

Proof. Let $u \in C^2(\Omega)$ and $x \in I_0(u) \cup I_1(u)$ with respect to $B_\epsilon = \{x \in \mathbb{R}^2 | \|x\|_2 < \epsilon\}$ for an arbitrary $\epsilon > 0$. Then there exists a direction $v \in S^1$ such that for all $t \in \mathbb{R}$ with $tv \in B_\epsilon$ the following holds

$$u(x) = u(x + tv). \tag{3.37}$$

The directional derivative in the direction of v then becomes

$$D_v u(x) = \lim_{t \to 0} \frac{u(x+tv) - u(x)}{t} = 0. \tag{3.38}$$

u is continuosly differentable by definition such that it is also totally differentiable. For totally differentiable functions the following relation to the directional derivate holds

$$D_v u = \nabla_x u \cdot v = v_1 \frac{\partial}{\partial x_1} u + v_2 \frac{\partial}{\partial x_2} u. \tag{3.39}$$

With Equation (3.38) we obtain the relation

$$\frac{\partial}{\partial x_1} u = k \frac{\partial}{\partial x_2} u \tag{3.40}$$

where $k := -v_2/v_1$. In the following it is shown that the nominator of the operators T_1 and T_2 becomes zero within this setup. The nominator of T_1 becomes

$$\frac{\partial^2}{\partial x_1^2} u \frac{\partial^2}{\partial x_2^2} u - (\frac{\partial^2}{\partial x_1 \partial x_2} u)^2 = k \frac{\partial^2}{\partial x_1 \partial x_2} u \frac{1}{k} \frac{\partial^2}{\partial x_2 \partial x_1} u - (\frac{\partial^2}{\partial x_1 \partial x_2} u)^2 = 0.$$

The nominator of T_2 becomes

$$\begin{aligned} &(\frac{\partial}{\partial x_1} u)^2 \frac{\partial^2}{\partial x_2^2} u + (\frac{\partial}{\partial x_2} u)^2 \frac{\partial^2}{\partial x_1^2} u - 2 \frac{\partial}{\partial x_1} u \frac{\partial}{\partial x_2} u \frac{\partial^2}{\partial x_1 \partial x_2} u \\ =&(k \frac{\partial}{\partial x_2} u)^2 \frac{\partial^2}{\partial x_2 \partial x_1} u + k (\frac{\partial}{\partial x_2} u)^2 \frac{\partial^2}{\partial x_1 \partial x_2} u - 2k \frac{\partial}{\partial x_2} u \frac{\partial}{\partial x_2} u \frac{\partial^2}{\partial x_1 \partial x_2} u = 0. \end{aligned}$$

Both operators thus fulfill the requirements to be an $i2D$-operator. □

3.3 Article: Spatial numerosity: A computational model based on a topological invariant

Reference

This work was carried out under the supervision of Christoph Zetzsche;
T.K. and C.Z. designed research; T.K. and C.Z. performed research; T.K. implemented the work; T.K. tested the algorithm; T.K. and C.Z. wrote the paper.

The paper was published in *Spatial Cognition IX* under the following reference [43]:

T. Kluth and C. Zetzsche. Spatial numerosity: A computational model based on a topological invariant. In C. Freksa, B. Nebel, M. Hegarty, and T. Barkowsky, editors, *Spatial Cognition IX*, volume 8684 of *Lecture Notes in Computer Science*, pages 237–252. Springer International Publishing, 2014.

Spatial Numerosity: A Computational Model Based on a Topological Invariant

Tobias Kluth and Christoph Zetzsche

Cognitive Neuroinformatics, University of Bremen,
Enrique-Schmidt-Straße 5, 28359 Bremen, Germany
tkluth@math.uni-bremen.de
http://www.informatik.uni-bremen.de/cog_neuroinf/en

Abstract. The estimation of the cardinality of objects in a spatial environment requires a high degree of invariance. Numerous experiments showed the immense abstraction ability of the numerical cognition system in humans and other species. It eliminates almost all structures of the objects and determines the number of objects in a scene. Based on concepts and quantities like connectedness and Gaussian curvature, we provide a general solution to this problem and apply it to the numerosity estimation from visual stimuli.

Keywords: numerosity, curvature, connectedness, Gauss-Bonnet.

1 Introduction

A fundamental ability humans and other species share is that they can interact efficiently with a spatial environment. This requires knowledge about the location of objects and their spatial arrangement. For example, the spatial distribution of food sources is an evolutionary crucial factor. To know which contains more fruits can decide on survival. The number of objects or its approximation, "numerosity" [3], thus is an important feature of spatial perception. But there is also evidence for a more extensive relation between number and space because it is assumed that the number representation of human adults is translated into corresponding spatial extensions and positions [11,15], also referred to as number-space mapping. It has also been reported that numerical processing modulates spatial representation according to a cognitive illusion [10,31]. For example, the error in the reproduction of a spatial extension is strongly dependent on the numbers delimiting the space [9], or bisecting a line flanked by two numbers is biased by the larger one [27]. Studies regarding the development of cognitive abilities in children also suggest a close relation between spatial and number sense [34]. Furthermore, with the "Theory of Magnitude" [35] exists an approach which suggests that a more general class of magnitudes, including number, are closely connected to space. This is also supported by evidence from neuroanatomical findings showing that both numerical and spatial tasks cause a similar activation in common parietal structures [13].

C. Freksa et al. (Eds.): Spatial Cognition 2014, LNAI 8684, pp. 237–252, 2014.

Qualitative spatial reasoning frameworks based on topological information, e.g. the Region Connection Calculus [26], rely on Whitehead's development of a theory of extensive abstraction based on a two-place predicate describing connection. The important point in Whitehead's theory is that formal individuals can be interpreted as spatially connected (for further information and a short historical overview we refer to [7]). Connectedness thus is not only an important concept in the description of formal individuals and in qualitative spatial reasoning but also in the estimation of numbers as we propose in this paper.

Number estimation, which includes approximate number recognition but not sequential counting, is not restricted to humans with mature cognitive abilities but has also been found in infants and animals [3,22], recently even in invertebrates [17]. Humans recognize numbers rapidly and exactly up to four, which is named subitizing [20]. They also rapidly estimate larger numbers but an error according to the Weber-Fechner law arises [16]. It is still an open question whether the two different observations rely on the same processing system [25,28]. The number is one of the most abstract properties of elements in a spatial configuration as it requires an invariance property which is not affected by spatial attributes like the orientation [1] and the shape [32] of the objects or by the modality [29]. Numerosity is an important element of a larger class of holistic properties like summary statistics [2,6], the gist of a scene [23], or just the average orientation of elements in a display [3].

The standard view of cortical organization as a local-to-global processing hierarchy [19] identifies numerosity as a high-level feature which is computed in a progression of levels. But there is evidence for a "direct visual sense for number" since number seems to be a primary visual property like color, orientation or motion, to which the visual system can be adapted by prolonged viewing [28].

Models which address the numerosity recognition in a neurobiologically plausible fashion, starting from individual pixels or neural receptors instead of an abstract type of input, are rare. A widely known model by Dehaene and Changeux [12] is based on linear Difference-Of-Gaussian filters of different sizes and an accumulation system. The linear filter operation restricts the model to number estimation from blob-like stimuli, which substantially limits the invariance property. A more recent model by Stoianov and Zorzi [30] is based on unsupervised learning in a deep recurrent network. The training images were binary and the objects presented were only rectangular areas such that moderate shape variations were investigated.

In contrary, our proposed approach starts with a general formulation of the problem of number recognition. We address the more fundamental question of how the number is interpreted as an invariance property of perceived scenes. This builds the basis to deal with the computational issues and the related neural requirements. To our knowledge, the first approach in this direction has been made in [37]. The paper is structured as follows: A clear definition of what is an object and how the number can be determined is considered in Section 2. Qualitative results for the number extraction from images are presented in Section 3. Finally, the paper ends with a discussion in Section 4.

2 Methods

As mentioned, modeling numerical cognition of humans was addressed in a few works, e.g. [12,30,37]. The majority of approaches has in common that they address the question of how numerosity within a stimulus is computed using a given algorithm. We address the same problem from a different point of view. Before we give a solution to the computational part of the problem using a specified modality, we focus on a proper formulation and its solution. We thus will begin with a characterization of the mathematical basis for the term *number*.

In set theory the definition of cardinality seems to have the same properties as what we understand as *number*. Generally spoken, the cardinality of a set describes the number of elements within the set. But if we have a closer look at this definition, the choice of what represents a number seems to be arbitrary because the numbers itself are just a formal construction to provide a label for the equivalence classes of the relation "having the same cardinality". We thus must differentiate the understanding of numbers from the definition of the natural numbers and their Arabic digits. However, set theory alone does not provide the tools and basics to properly describe how the number of objects in the real world is related to basic properties of objects. We thus address the three following issues under a topological point of view. First, we must find a formal definition of the "real world". Second, we have to specify objects and their properties. And third, the meaning of *number* in this context and its invariant properties are defined. The following considerations are based on first investigations in a similar direction in [37].

The world in which we live can be assumed to be four-dimensional if we consider the time as an additional dimension. Within the context of this paper we restrict the real world to be static such that it can be represented by a three-dimensional vector space, the real valued space $\mathbb{R}^3$. This space becomes a topological space if it is equipped with a suitable topology, e.g. the usual topology of the $\mathbb{R}^n$.

This directly leads to the second question, what is an object in the real world and how can it be specified. An object is three-dimensional which means that points, lines and planes in the common geometrical sense are no objects in the real world, rather they are theoretical constructions to describe geometrical quantities. For example a sheet of paper can be as thin as it is possible to produce, nevertheless it will always have an extension in the third dimension. We thus restrict an object to be a connected subset of the real world, which means that it cannot be represented by a union of disjoint subsets. Furthermore we make the technically caused assumption that objects are not only connected subsets but also simply connected subsets. This implies that subsets with holes, e.g. a donut or a pretzel, independent of whether they exist in nature or not, are not taken into account. This case will be considered in more detail in the discussion part. In Figure 1 the considered setup is illustrated in three examples. From a perceptional point of view, the interior of the objects is quite uninteresting, thus it is reasonable to represent the objects by their surface. As the kind of connectedness is an important feature of the objects, we need a unique

relation between the connectedness of an object and its surface. It can be shown that if two objects share the same kind of connectedness, their surfaces also have the same kind of connectedness [14] (the two-dimensional invariant Euler characteristic is twice the three-dimensional one).

We now have a definition of the real world and of what is meant by an object and we thus can address the third and most important question concerning the *number* of objects. To find formal mechanisms which are suitable to measure the number of objects, we formulate the requirements for this mechanisms. The mechanism must be able to obtain the number in a one-shot manner, which means that the number should not be computed sequentially, e.g. like counting. It should be a mechanism which maps one invariant quantity to each object which is ideally the same such that it can be normalized easily in order to sum it up afterwards. The summation then results directly in the number of objects.

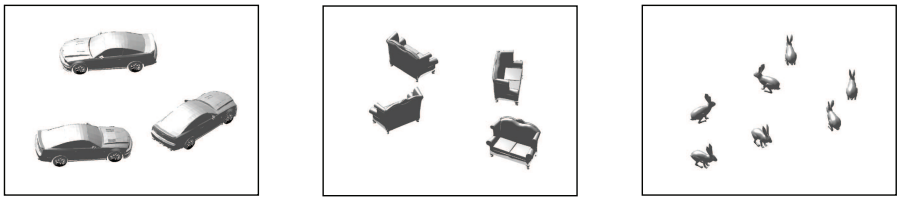


Fig. 1. All illustrations show a variety of three-dimensional simply connected objects which build a scene. They are all (a)-(c) differing in the spatial arrangement and the kind of objects they show.

The invariance property *number* we expect needs more clarification in the sense that an invariant is always defined on a set of elements and a set of operations. In our case the elements are the previously defined objects and the operations have yet to be defined. The set of operations which map one object to another object must be number-conserving, or the other way around the invariant property must be independent of the operation. Assuming the *number* can be represented by a topological invariant, it is obvious that the operations are restricted to homeomorphisms, i.e. continuous and bijective mappings whose inverse is also continuous. This kind of mapping is structure-conserving in the sense that the kind of connectedness does not change. Cutting, tearing and gluing of objects, for example, are no homeomorphisms in general and would cause a change of the invariant.

The additivity we expect also requires further considerations of the object properties. In order to just sum up over all surfaces without a priori knowledge we must find a feature of the surface which then results in an invariant complying with the previously described requirements. The Gauss-Bonnet theorem provides a connection between topology and differential geometry and relates the Gaussian curvature of two-dimensional surfaces to their Euler characteristic.

Theorem 1 (Gauss-Bonnet). *Let $S \subset \mathbb{R}^3$ be a regular oriented surface (of class C^3), and let R be a compact region of S. Suppose that the boundary ∂R is a simple, closed, piecewise regular, positively oriented curve. Assume ∂R consists of k regular arcs ∂R_i (of class C^2), and let θ_i be the external angles of the vertices of ∂R. Then*

$$\int_R K \; dS + \sum_{i=1}^{k} \int_{\partial R_i} \kappa_g \; ds + \sum_{i=1}^{k} \theta_i = 2\pi\chi(R) \tag{1}$$

where K is the Gaussian curvature, κ_g is the geodesic curvature, and χ is the Euler characteristic.

The presented theorem is more general as it also considers the case that the surface has a boundary. It becomes important in a later part dealing with the problems which arise from the perception via sensors. For the moment, the objects we deal with do not have a one-dimensional boundary such that we can consider the following corollary of the theorem. Proofs of the theorem and the corollary can be found in almost every differential geometry textbook.

Corollary 1. *Let S be an orientable, compact, regular surface of class C^3. Then*

$$\int_S K \; dS = 2\pi\chi(S).$$

The Euler characteristic is a topological invariant which maps a number to any subset within a topological space. This number then characterizes the kind of connectivity of the subset. For example, the surface of a sphere (no holes) has a different characteristic number than the surface of a torus (one hole). Being a topological invariant, the Euler characteristic of an object remains constant if a homeomorphism is applied to it. Given the Gauss-Bonnet theorem, the question is what that means with respect to the Gaussian curvature of the object's surface. We first give a short description of what is meant by Gaussian curvature and we then consider the influence of homeomorphisms and why the integration over this quantity is constant. Though the proof of the theorem provides a technical solution, it gives no idea of the interplay of different kinds of curvature.

The Gaussian curvature of a regular surface is defined as the product of its principal curvatures. These are the minimal and maximal normal curvatures of two orthogonal planes which are both orthogonal to the tangent plane of the surface. The local shape of a surface can be distinguished by its Gaussian curvature in elliptic ($K > 0$), hyperbolic ($K < 0$), and parabolic parts ($K = 0$). For example, the surface S of a sphere, as shown in Figure 2, is completely elliptic (blue) and has the Euler characteristic $\chi(S) = 2$. In this case, the Gaussian curvature is constant and depends only on the radius of the sphere. As increasing the radius of a sphere is a homeomorphism, Corollary 1 states that the integral over the Gaussian curvature is constant. In this example the interplay between the surface area and the Gaussian curvature is easy to conceptualize. While increasing the radius, the surface area increases and the Gaussian curvature decreases such that the surface integral, i.e. the product of Gaussian curvature

and surface area in this special case, is 4π constantly. We want to emphasize that homeomorphic deformations of the sphere are exactly the objects resulting from the previously formulated assumptions. The surface integral also does not change if the sphere is dented, as can be seen in the middle row of Figure 2. Here a hyperbolic (red) curvature emerges at the boundary of the dent. Under the assumption that the curvature does not change except in the dent, curvature with a negative sign always implies the emergence of another elliptic part with a higher absolute value. We can find this elliptic part in the middle of the dent, which can be seen in Figure 2c. The Gaussian curvature of a regular surface is a continuous function and thus there exists at least a curve between hyperbolic and elliptic parts on the surface where the curvature is parabolic. The influence of a homeomorphism producing a bulge on the surface of a simply connected object and its Gaussian curvature is also shown in Figure 2. The connection of Gaussian curvature and the invariant Euler characteristic fulfills the first requirement we formulated.

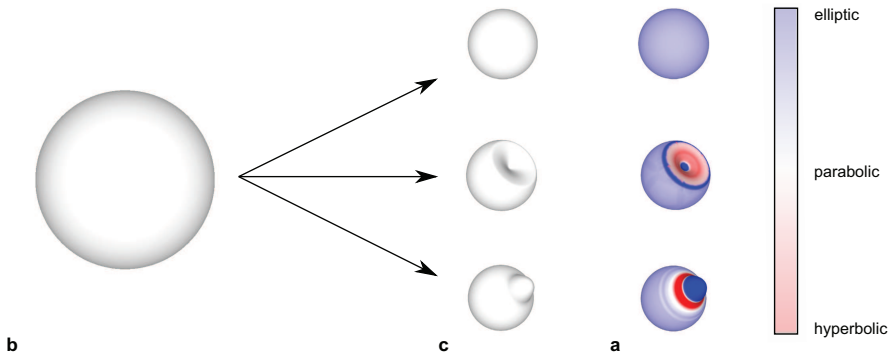


Fig. 2. In (a) the surface of a three-dimensional sphere is shown which is transformed via homeomorphisms into (b) a smaller sphere (top), a smaller sphere with a dent (middle), and a smaller sphere with a bulge (bottom). In (c) the Gaussian curvature of the 3D objects in (b) is shown qualitatively by coloring the surface respectively its kind of curvature. The elliptic parts of the surface are blue, the hyperbolic parts are red, and the parabolic parts are white. The color saturation shows the amount of curvature contributed to the integral in the Gauss-Bonnet theorem.

The second requirement, the additivity, follows directly from the linearity of the integral operator. Assume $S := \bigcup_{i=1}^{n} \partial O_i$ to be the union of surfaces of pairwise disjoint objects O_i. With $\chi(\partial O_i) = 2$, $i = 1, \ldots, n$, the integral becomes

$$\int_S K \; dS = \sum_{i=1}^{n} \int_{\partial O_i} K \; dS = 2\pi \sum_{i=1}^{n} \chi(\partial O_i) = 4\pi n.$$

In principle the problem of estimating the number of objects in a scene is hereby solved. If we assume measuring the local property, Gaussian curvature, of the

surface is directly possible, the integration of local measurements over all simply connected object's surfaces results in a representation of the number.

Sensors

Estimating the number of objects requires access to the Gauss-Bonnet quantities of the surface of the object. But if we think about the modalities from which we can gather the required information, it is not obvious how an estimate of these quantities can be obtained. The abstract formulation of the problem is a comfortable starting position for further investigations. Sequential tactile scanning of the object's surface or visual sensing are imaginable sensor strategies for human number estimation. Multisensory approaches would also be possible and with additional technical sensor devices, as they are used in robots, the number of possibilities for number estimation becomes quite high. In the following we consider the human visual system as the modality to obtain an estimate of the Gaussian curvature yielding the number of objects in a perceived scene. Here we restrict the visual stimulus to be luminance only such that we have a sensory representation of the scene by the luminance function $l = l(x, y)$. An example for the luminance function of a three-dimensional sphere is illustrated in Figure 3a. But it remains unclear how this is related to the real world. The interplay between lighting and reflectance properties of the object's surface result in a mapping from the real world to the sensed luminance function. For further considerations in this directions, we refer to [21]. In the following, the real world is equipped with a binary function $g : \mathbb{R}^3 \to \{0, 1\}$ describing the physical properties at a position $x \in \mathbb{R}^3$. In the considered setup the function g is just an indicator function whether a position is occupied by an object or not. The sensor operator F maps this physical properties to the luminance function $l : \mathbb{R}^2 \to \mathbb{R}_0^+$ such that the sensory process becomes

$$F(g) = l \quad , \; g(x) := \chi_S(x) = \begin{cases} 1 \, , & x \in S, \\ 0 \, , & \text{else}, \end{cases} \tag{2}$$

where S is a disjunction of sufficiently smooth objects according to the previous definitions and χ_S is the characteristic function. Note that χ without an index is the Euler characteristic. Assuming F is an orthogonal projection not incorporating lighting the resulting luminance level is always constant for the objects. In particular, this case for objects resulting in right-angled polygons as projections was considered in previous works [37,38]. If we assume planar surfaces to be projected this way, the resulting setup matches common visual stimuli in psychological studies, e.g. [18]. However, we assume that F takes lighting of the objects into account such that it does not result in a constant luminance level on objects. Additionally, we assume the operator F to preserve the differentiability on the objects surface. The piecewise constant function g is thus mapped to an almost everywhere sufficiently smooth function l. The background is assumed to be uniform and clearly separable from the objects. The discontinuity between objects and the environment is thus projected to the luminance function. In the following we apply similar concepts of the previously presented general

solution to the luminance function in order to estimate the number of perceived objects. The different approaches all have in common that we investigate the luminance surface $\Omega \subset \mathbb{R}^3$ which is defined by the perceived luminance function l, i.e. $\Omega := \{(x, y, z) \in \mathbb{R}^3 | x, y \in \mathbb{R}, z = l(x, y)\}$. For example, in Figure 3b the luminance surface of the luminance function in Figure 3a is shown.

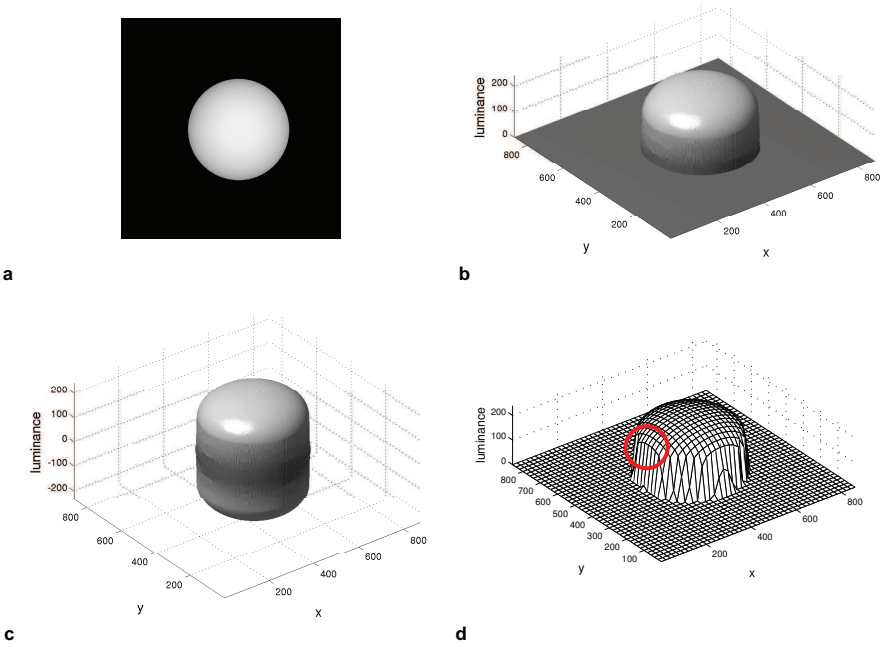


Fig. 3. In (a) the luminance function resulting from the projection of a slightly lightened three-dimensional sphere is shown. (b) shows the luminance function of the image in (a) interpreted as a two-dimensional surface embedded in the three-dimensional space. In (c) the background of the luminance surface in (b) was removed, the remaining surface was mirrored at the x-y-plane, and both surface pieces are then connected in z-direction. Thus, (c) shows a simply connected surface of a 3D object. In (d) a subsampled version of the underlying surface mesh is shown. The red circle highlights the projected discontinuity between the object and the environment. If the surface is continued in z-direction, an arc can emerge in this region, which has to be taken into account additionally while doing the curvature computation.

First, a simplified basic concept is investigated. Assume that the background, which is constantly zero, is cut off from the luminance surface. Then the rest of the surface is mirrored at the x-y-plane and it is connected in z-direction. We thus can construct closed surfaces representing three-dimensional objects from the embedded luminance surface. In Figure 3c the resulting surface corresponding

to Figure 3a is illustrated. This is exactly the starting situation of the general problem in the previous section and the problem seems to be solved for luminance images. We can estimate the Gaussian curvature using the luminance function and it becomes

$$K(x,y) = \frac{l_{xx}(x,y)l_{yy}(x,y) - l_{xy}(x,y)^2}{(1 + l_x(x,y)^2 + l_y(x,y)^2)^2}, \tag{3}$$

where subscript denotes the differentiation in the respective direction (e.g. $l_{xy} = \frac{\partial^2 l}{\partial x \partial y}$). Integration over this quantity should result in the number of objects times a characteristic constant. If we have a closer look at the constructed closed three-dimensional surface, there is a critical region at the boundary of the luminance surface without the background, see Figure 3d. In general we cannot guarantee the differentiability in this region. The projected discontinuity between objects and background causes errors because an additional arc at the critical region has to be taken into account. If the tangent planes in all points at the boundary region are orthogonal to the x-y-plane, this problem does not occur. For example, this is the case if the luminance function without background looks like a hemisphere. Then the differentiable extension in z-direction is possible and no error terms would occur. But this is not the general case.

The problem can be solved by only taking the boundary region into account. The boundary curve can emerge from a discontinuity detection on the luminance function. Given an appropriate detection function, possibly a linear differential operation or a threshold function, the general three-dimensional case is projected to two dimensions. The luminance is thus used for the detection only. The resulting objects are bounded two-dimensional subsets of the x-y-plane such that they can be represented by their one-dimensional boundary curve. In this case the one-dimensional counterpart of the Gauss-Bonnet theorem is the following standard corollary in differential geometry.

Corollary 2. *Let C be a closed, regular, plane curve. Then the quantity*

$$\int_C \kappa_g \, ds = 2\pi n, \tag{4}$$

where κ_g is the curvature of the plane curve and n is an integer called the rotation index of the curve.

As the rotation index is always one for the objects we consider and the integral operator is linear, the integral over the disjunction of multiple boundary curves results in their number. In Figure 4 the two-dimensional case of Figure 2 is shown. Assuming a counterclockwise parametrization, the curvature in a point is positive if the tangent is on the right side of the curve and respectively on the left side if the curvature has a negative sign. The circle in Figure 4b has always a positive curvature such that its integral becomes 2π. The second shape is a circle with a dent. Negative curvature emerges in the middle part of the dent but if we have a closer look to the positive curvature at the boundary of the dent, the absolute of this contribution to the integral increases such that it equalizes the negative contribution of the dent. Note that the curvature contribution is not mapped equivalently from the three-dimensional case to the one

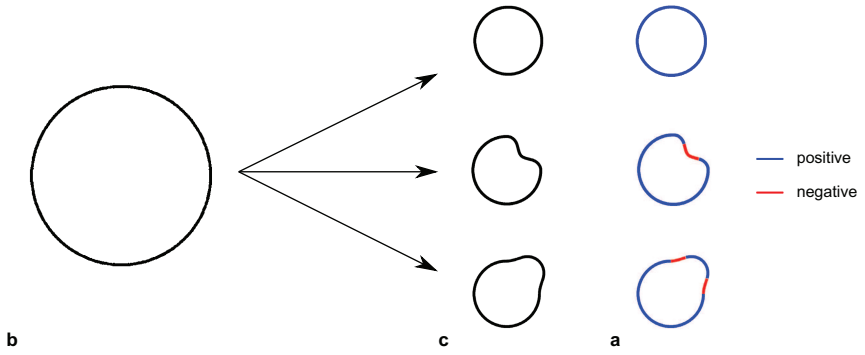


Fig. 4. The two-dimensional case of Figure 2 is illustrated. In (a) the contour of a circle is shown which is transformed into (b) a smaller circle (top), a smaller circle with a dent (middle), and a smaller circle with a bulge (bottom). In (c) the curvature of the plane boundary curves in (b) is shown qualitatively. The color shows whether the curvature has a positive or negative contribution to the curvature integral.

with lower dimensionality as the dent only produces one negative contribution and no positive contribution emerges in the middle of the dent, compare Figures 2c and 4c.

Assume a smooth boundary curve C is given by an arbitrary parametrization $c : I \to \mathbb{R}^2$ with $c(t) = (x(t), y(t))^T$, where $I \subset \mathbb{R}$. The curvature κ_g then becomes

$$\kappa_g(t) = \frac{x''(t)y'(t) - y''(t)x'(t)}{(x'(t)^2 + y'(t)^2)^{3/2}}, \tag{5}$$

where the primes denote differentiation with respect to t, e.g. $x' = \frac{\partial}{\partial t}x$ and $x'' = \frac{\partial^2}{\partial t^2}x$. This approach corresponds to the Gestalt grouping principle of connectedness [24], which assumes that points are grouped into objects based on connected regions. In this case the objects could be represented by the outlines of the connected regions. But, in order to estimate the number of objects from the one-dimensional line integral the first and second derivatives of a parametrization of a curve are necessary. Regarding a neural representation of the curve which can be interpreted as a point set, the first derivative is plausible as it can be obtained from simultaneously firing neighbored units. Second derivatives of the luminance functions are also plausible, as an neuronal approximation can be obtained via linear filtering with second derivatives of filter kernels [36,37]. But the second derivative of the parametrization seems to be an additional effort which is unnecessary. Thus it is reasonable for the system to compute a solution to the number problem using these quantities. Further investigation of curved surfaces results in another computational approach which does not contradict the Gestalt grouping principle of connectedness as both approaches are formally equivalent.

In the following we address the previously arisen question and derive a solution to avoid the computation of second derivatives of parametrized curves. Theorem 1 provides a solution which incorporates the luminance surface properties directly. Assuming a smooth space curve C which is the boundary of a closed region S on the luminance surface Ω, i.e. $C = \partial S$, it results

$$\int_S K \; dS + \int_C \kappa_g \; ds = 2\pi\chi(S), \tag{6}$$

where κ_g denotes the geodesic curvature. Using the parametrization $\phi(x, y) := (x, y, l(x, y))^T$ of the surface and the Gaussian curvature from equation 3, the first integral can be computed by

$$\int_S K \; dS = \int_{\mathbb{R}^2} \underbrace{\frac{l_{xx}(x,y)l_{yy}(x,y) - l_{xy}(x,y)^2}{(1 + l_x(x,y)^2 + l_y(x,y)^2)^{3/2}}}_{=:\tilde{K}(x,y)} \chi_S(\phi(x,y)) \; d(x,y), \tag{7}$$

where χ_S is the characteristic function with respect to the set S. In order to calculate the second integral, the geodesic curvature of the boundary curve C must be estimated. Using the parametrization $c : I \to \mathbb{R}^3$ with $c(t) := (x(t), y(t), l(x(t), y(t)))^T$ of the boundary curve and the assumption $l(x(t), y(t)) = const.$, $\forall t \in I$, we can determine the geodesic curvature by

$$\kappa_g(t) = \tilde{\kappa}_g(x(t), y(t)) = \frac{l_x^2 l_{yy} + l_y^2 l_{xx} - 2 l_x l_y l_{xy}}{(l_x^2 + l_y^2)^{3/2}(1 + l_x^2 + l_y^2)^{1/2}}. \tag{8}$$

The additional assumption of constant height is made to eliminate the second derivatives of the parametrization. The second integral in equation 6 thus becomes

$$\int_C \kappa_g \; ds = \int_{\mathbb{R}} \frac{l_x^2 l_{yy} + l_y^2 l_{xx} - 2 l_x l_y l_{xy}}{(l_x^2 + l_y^2)^{3/2}(1 + l_x^2 + l_y^2)^{1/2}} (x'^2 + y'^2)^{1/2} \chi_C(c(t)) \; dt, \tag{9}$$

where χ_C is the characteristic function with respect to the set C. Assuming constant height, the first derivative of the parametrization yields $x' = -(l_y/l_x)y'$. Thus the geodesic curvature depends only on differential operators applied to the luminance function and one first derivative x' or y' of the parametrization of the boundary curve. Finally, the number of objects n from a union S of pairwise disjoint regions of luminance surfaces can be determined by

$$2\pi n = \int_{\mathbb{R}^2} \tilde{K}(x,y)\tilde{\chi}_S(x,y) \; d(x,y) + \int_{\mathbb{R}} \tilde{\kappa}_g(x(t), y(t))\tilde{\chi}_C(x(t), y(t)) \; dt, \tag{10}$$

where $\tilde{\chi}_S(x, y) := \chi_S(\phi(x, y))$ and $\tilde{\chi}_C(x(t), y(t)) := (x'^2 + y'^2)^{1/2}\chi_C(c(t))$. The computation now depends on three quantities $\tilde{K}$, $\tilde{\chi}_S$, $\tilde{\kappa}_g$ depending on surface properties and the quantity $\tilde{\chi}_C$ which has to extract curve properties. All quantities have in common that they represent local properties of the luminance

surface. Assuming a neural hardware, the characteristic functions correspond to the well known threshold functionality of neurons. The curvature operators $\tilde{K}$ and $\tilde{\kappa}_g$ depend only on derivatives of the luminance function which can be realized by neurophysiologically realistic Gabor-like filters [36,37]. The computation also requires a multiplicative "AND"-like combination of these features which can be obtained by the neural mechanism of cortical gain control [39]. A neural hardware thus is able to estimate the number of objects using the neural correlates of the operations in equation 10.

3 Results

We implemented and tested the algorithm implied by the operator defined in equation 10 to estimate the number of objects in an image. The images are represented by positive real-valued matrices. The stimuli are assumed to be sufficiently smooth which can cause high numerical errors if it is not satisfied. In order to obtain the differentiability in the discrete representation, each stimulus is lowpass filtered using a Gaussian kernel. In the following we present qualitative results for the images presented in Figure 1.

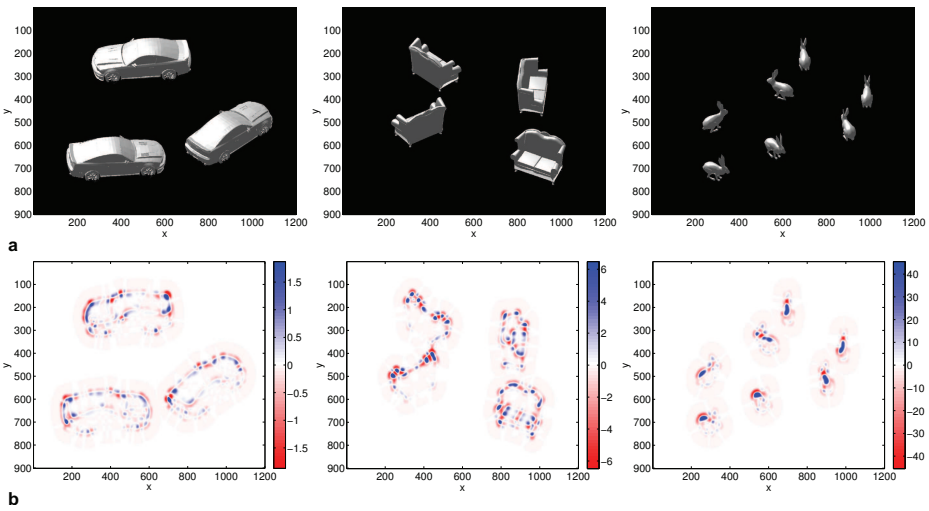


Fig. 5. In (a) the visual stimuli from Figure 1 with their original background are shown. In (b) the corresponding Gaussian curvatures (elliptic: blue, parabolic: white, hyperbolic: red) of the luminance surfaces belonging to the images in (a) are shown.

The Gaussian curvature is a prominent quantity for the extraction of the number of objects. In Figure 5 the test images and their Gaussian curvature are shown. They all show high variations in the kind of curvature over the luminance surface and especially on objects there is a noticeable interplay between

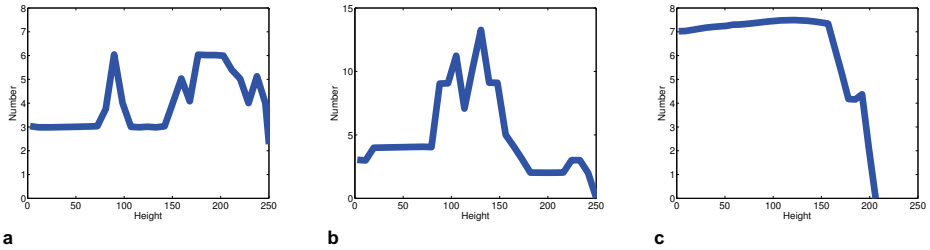


Fig. 6. In (a)-(c) the estimated number of objects from the corresponding stimuli in Figure 1a-c are shown. The number was estimated by the operator defined in equation 10.

hyperbolic and elliptic parts of curvature. Especially the boundary region of the objects has at least one change of sign in Gaussian curvature. This is caused by the zero background and the Gaussian filtering in advance which is necessary to obtain acceptable results. There is a relation between the geodesic curvature of the boundary curve and the shape of the surface if we assume the boundary curve to be a height line. The steeper the luminance surface in a point, i.e. the smaller the angle between the tangent normal vector and the x-y-plane, the less is the contribution of the geodesic curvature integral to the number. In order to get rid of the geodesic curvature integral, scaling the stimulus would be one option to control this relation but can cause numerical instability. However, the interplay of Gaussian curvature and geodesic curvature results in an estimate of the number of objects. The only parameter which must be chosen is the threshold of luminance (height) at which the boundary curves are defined. Everything above this threshold belongs to the integration domain for the Gaussian curvature. The implementation then results in a number estimate which depends on the height, as can be seen in Figure 6. The number of objects is estimated correctly for height values up to 50. Heights higher than 50 cause seemingly unsystematical errors. But it is obvious, they must have a close relation to the shape and absolute values of the luminance surface. In Figure 6b an underestimation is done for low heights. It has its origin in two objects which are close together such that the Gaussian filter operation connects both objects which means they are not separated by a background region anymore. Figure 6c has a noticeable increase of number with increased height. All graphs have in common that they suddenly descend to zero as the height is higher than the maximum value of the luminance surface.

4 Discussion

Our approach for computing the number of objects from a scene differs from other approaches in several respects.

First, our approach can deal with arbitrary simply connected objects. On a certain level the model by Dehaene et al. [12] can be seen to have a similar

structure: It also sums up the output of normalization units to obtain the number of objects within a stimulus. In this case the objects are assumed to be blob-like which is necessary for the object detection in the normalization step. Our approach is restricted to the case of simply connected objects. We do not know a similar invariant which maps the same value to each kind of connectedness, e.g. to tori or other objects with more holes. And if an invariant could comply with these requirements, it remains unclear whether it could be computed by local properties of the object, like the Gaussian curvature in our approach. As there is evidence that the human perception is sensible to topological quantities [6] and as the number estimation method we proposed is based on topological concepts, the role of topology in perception should be rethought.

Second, the underlying principles to obtain the invariance are clearly defined and the system is no black box. Recent results by Stoianov et al. [30] show that an unsupervised training of a recurrent multi-layered network is able to learn a representation of number but the underlying invariance principle is not known. This raises the question, which kind of system is able to learn the full invariance principles in an unsupervised fashion and whether it extracts local properties, like the Gaussian curvature.

In digital topology there exist connected component labeling algorithms, e.g. [33], which are also able to estimate the number of labels they distribute. In contrary to our fully parallel approach these algorithms are sequentially and need multiple passes through the image which is quite unplausible for the neurobiological system. But the formulation of our approach in terms of digital topology is desirable as the neural system consists of a countable and finite set of neurons.

Other approaches [8] suggest a strong interrelation between density, covered area, and the number of objects. It is assumed that the number is obtained approximately from the product of area and density. However, the crucial properties in our view are the extraordinary invariance properties which would also be required for generally responding computations of density.

The accuracy of the proposed approach is independent of the number presented in the stimulus. The structure of the model, which includes a multiplicative gain control, is a promising basis for the emergence of a log-normally distributed output [4] if noise is taken into account. As a result the model would explain the errors in the rapid estimation for the whole range of numbers, i.e. for subitizing and for the estimation of higher numbers, but this remains a question for future research.

As we have shown, linear summation and cortical gain control, two widely accepted properties of cortical neurons, are the only requirements for the computation of the number of objects. These functions are already available at early stages of the cortex, but also in other areas [5].

Acknowledgements. This work was supported by DFG, SFB/TR8 Spatial Cognition, project A5-[ActionSpace].

References

1. Allik, J., Tuulmets, T., Vos, P.G.: Size invariance in visual number discrimination. Psychological Research 53(4), 290–295 (1991)
2. Alvarez, G.A.: Representing multiple objects as an ensemble enhances visual cognition. Trends in Cognitive Science 15(3), 122–131 (2011)
3. Brannon, E.M.: The representation of numerical magnitude. Current Opinion in Neurobiology 16(2), 222–229 (2006)
4. Buzsáki, G., Mizuseki, K.: The log-dynamic brain: how skewed distributions affect network operations. Nature Reviews Neuroscience (2014)
5. Carandini, M., Heeger, D.J.: Normalization as a canonical neural computation. Nature Reviews Neuroscience 13, 51–62 (2012)
6. Chen, L.: The topological approach to perceptual organization. Visual Cognition 12(4), 553–637 (2005)
7. Clarke, B.L.: A calculus of individuals based on "connection". Notre Dame Journal of Formal Logic 22(3), 204–218 (1981)
8. Dakin, S.C., Tibber, M.S., Greenwood, J.A., Kingdom, F.A.A., Morgan, M.J.: A common visual metric for approximate number and density. Proceedings of the National Academy of Sciences 108(49), 19552–19557 (2011)
9. De Hevia, M.-D., Girelli, L., Bricolo, E., Vallar, G.: The representational space of numerical magnitude: Illusions of length. The Quarterly Journal of Experimental Psychology 61(10), 1496–1514 (2008)
10. de Hevia, M.D., Girelli, L., Vallar, G.: Numbers and space: a cognitive illusion? Experimental Brain Research 168(1-2), 254–264 (2006)
11. Dehaene, S., Bossini, S., Giraux, P.: The mental representation of parity and number magnitude. Journal of Experimental Psychology: General 122(3), 371 (1993)
12. Dehaene, S., Changeux, J.P.: Development of elementary numerical abilities: a neuronal model. Journal of Cognitive Neuroscience 5(4), 390–407 (1993)
13. Dehaene, S., Piazza, M., Pinel, P., Cohen, L.: Three parietal circuits for number processing. Cognitive Neuropsychology 20(3-6), 487–506 (2003)
14. Dillen, F., Kühnel, W.: Total curvature of complete submanifolds of euclidean space. Tohoku Mathematical Journal 57(2), 171–200 (2005)
15. Fias, W.: The importance of magnitude information in numerical processing: Evidence from the snarc effect. Mathematical Cognition 2(1), 95–110 (1996)
16. Gallistel, C.R., Gelman, R.: Preverbal and verbal counting and computation. Cognition 44(1), 43–74 (1992)
17. Gross, H.J., Pahl, M., Si, A., Zhu, H., Tautz, J., Zhang, S.: Number-based visual generalisation in the honeybee. PloS one, 4(1), e4263 (2009)
18. He, L., Zhang, J., Zhou, T., Chen, L.: Connectedness affects dot numerosity judgment: Implications for configural processing. Psychonomic Bulletin & Review 16(3), 509–517 (2009)
19. Hegde, J., Felleman, D.: Reappraising the Functional Implications of the Primate Visual Anatomical Hierarchy. The Neuroscientist 13(5), 416–421 (2007)
20. Kaufman, E.L., Lord, M., Reese, T., Volkmann, J.: The discrimination of visual number. The American Journal of Psychology, 498–525 (1949)
21. Koenderink, J.J., van Doorn, A.: Shape and shading. The visual neurosciences, 1090–1105 (2003)
22. Nieder, A., Freedman, D.J., Miller, E.K.: Representation of the quantity of visual items in the primate prefrontal cortex. Science 297(5587), 1708–1711 (2002)

23. Oliva, A., Torralba, A.: Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
24. Palmer, S., Rock, I.: Rethinking perceptual organization: The role of uniform connectedness. Psychonomic Bulletin & Review 1(1), 29–55 (1994)
25. Piazza, M., Mechelli, A., Butterworth, B., Price, C.J.: Are subitizing and counting implemented as separate or functionally overlapping processes? Neuroimage 15(2), 435–446 (2002)
26. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. KR 92, 165–176 (1992)
27. Ranzini, M., Girelli, L.: Exploiting illusory effects to disclose similarities in numerical and luminance processing. Attention, Perception, & Psychophysics 74(5), 1001–1008 (2012)
28. Ross, J., Burr, D.C.: Vision senses number directly. Journal of Vision 10(2), 1–8 (2010)
29. Starkey, P., Spelke, E.S., Gelman, R.: Numerical abstraction by human infants. Cognition 36(2), 97–127 (1990)
30. Stoianov, I., Zorzi, M.: Emergence of a 'visual number sense' in hierarchical generative models. Nature Neuroscience 15(2), 194–196 (2012)
31. Stöttinger, E., Anderson, B., Danckert, J., Frühholz, B., Wood, G.: Spatial biases in number line bisection tasks are due to a cognitive illusion of length. Experimental Brain Research 220(2), 147–152 (2012)
32. Strauss, M.S., Curtis, L.E.: Infant perception of numerosity. Child Development 52(4), 1146–1152 (1981)
33. Suzuki, K., Horiba, I., Sugie, N.: Linear-time connected-component labeling based on sequential local operations. Computer Vision and Image Understanding 89(1), 1–23 (2003)
34. van Nes, F., van Eerde, D.: Spatial structuring and the development of number sense: A case study of young children working with blocks. The Journal of Mathematical Behavior 29(3), 145–159 (2010)
35. Walsh, V.: A theory of magnitude: common cortical metrics of time, space and quantity. Trends in Cognitive Sciences 7(11), 483–488 (2003)
36. Zetzsche, C., Barth, E.: Fundamental limits of linear filters in the visual processing of two-dimensional signals. Vision Research 30(7), 1111–1117 (1990)
37. Zetzsche, C., Barth, E.: Image surface predicates and the neural encoding of two-dimensional signal variations. In: Rogowitz, B.E., Allebach, J.P. (eds.) Proceedings SPIE, Human, Vision and Electronic Imaging: Models, Methods, and Applications, vol. 1249, pp. 160–177 (1990)
38. Zetzsche, C., Gadzicki, K., Kluth, T.: Statistical invariants of spatial form: From local and to numerosity. In: Proc. of the 2nd Interdisciplinary Workshop The Shape of Things, pp. 163–172. CEUR-WS.org (April 2013)
39. Zetzsche, C., Nuding, U.: Nonlinear and higher-order approaches to the encoding of natural scenes. Network: Computation in Neural Systems 16(2-3), 191–221 (2005)

3.4 Article: Numerosity as a topological invariant

Reference

This work was carried out under the supervision of Christoph Zetzsche;
T.K. and C.Z. designed research; T.K. and C.Z. performed research; T.K. implemented the work; T.K. conducted the computational experiments; T.K. and C.Z. wrote the paper.

The paper is still under review. It was submitted to the *Journal of Vision* under the title "A topological view on numerical cognition". The first submission was part of the submitted thesis. The revised manuscript, which is part of this work, was submitted to *Journal of Vision* under the following title:

T. Kluth and C. Zetzsche. Numerosity as a topological invariant.

Numerosity as a Topological Invariant

Tobias Kluth Cognitive Neuroinformatics
University of Bremen, Germany

Christoph Zetzsche Cognitive Neuroinformatics
University of Bremen, Germany

The ability for a fast recognition of the number of objects in our environment is a fundamental cognitive function. However, it is far from clear which computations and which actual neural processing mechanisms are used to provide us with such a skill. Here we try to provide a detailed and comprehensive analysis of this issue, which comprises both the basic mathematical foundations and the peculiarities imposed by the structure of the visual system and by the neural computations provided by the visual cortex. We suggest that numerosity should be considered as a mathematical invariant. Making use of concepts from mathematical topology, like connectedness, the Betti numbers, and the Gauss-Bonnet theorem, we derive the basic computations which are suited for the computation of this invariant. We show that the computation of numerosity is possible in a neurophysiologicaly plausible fashion, using only computational elements which are known to exist in the visual cortex. We further show that a fundamental feature of numerosity perception, its Weber property, arises naturally assuming noise in the basic neural operations. It is hoped that our results can provide a general framework for the future research on the invariance properties of the numerosity system.

Keywords: numerosity, topological invariance, Betti numbers, Euler characteristic, Gaussian curvature, functional model

Introduction

The information about the number of objects in the environment can be extracted in a fast and effortless fashion by the visual systems of humans and other species. Access to this information is a crucial factor in evolution: How many predators am I confronted with? On which tree can I get the larger amount of food? Questions like these show that survival can essentially depend on having access to this type of knowledge. It is therefore not surprising that the estimation of number, or of its approximation "numerosity" (Brannon, 2006), is also considered to be a fundamental property of cognition.

Humans recognize numbers rapidly and precisely up to four ("subitizing" (Kaufman, Lord, Reese, & Volkmann, 1949)) but they also estimate it rapidly for larger numbers, although there with increasing errors. A typical characteristic of the numerosity system is that the errors depend on the set size in accordance with the Weber-Fechner law (Gallistel & Gelman, 1992). Whether these characteristics support the idea of two clearly distinct subsystems or reflect different operation modes of a general number system is still under debate (Feigenson,

Dehaene, & Spelke, 2004; Piazza, Mechelli, Butterworth, & Price, 2002; Ross & Burr, 2010). The estimation of numerosity was not only reported for human adults (Whalen, Gallistel, & Gelman, 1999) but it was also shown that even six-month-old infants were able to do a number distinction task (Xu & Spelke, 2000). Investigation of the development of mathematical competence and the ability for numerosity estimation in children suggested that mathematical ability is correlated with the acuity in numerosity estimation (Halberda, Mazzocco, & Feigenson, 2008). But number estimation is not restricted to humans with mature cognitive abilities, it has also been found in infants and animals (Brannon, 2006; Nieder, Freedman, & Miller, 2002), recently even in invertebrates (Gross et al., 2009).

In general, numerosity is one of the most abstract properties in the environment and its perception is almost independent of spatial attributes like orientation (Allik, Tuulmets, & Vos, 1991) and of the shape of the objects (Strauss & Curtis, 1981). It is also not confined to a specific sensory modality (Starkey, Spelke, & Gelman, 1990). And finally, numerosity estimation extends beyond the estimation of the cardinality of objects. The quantity of physical properties like sound volume, space, and time shows similar characteristics (Bonn & Cantlon, 2012). This suggests that there might exist a generalized system for magnitude estimation (Gallistel & Gelman, 2000; Walsh, 2003).

The standard view of cortical organization as a local-to-global processing hierarchy (Hegde & Felleman, 2007) which goes from basic sensory properties towards the representation of the most abstract properties on top of the hierarchy would suggest that numerosity has to be considered a very high-level feature. On the other hand, single cell recordings show that neural reactions to numerosity are quite fast, approximately 100 msecs in macaques (Roitman, Brannon, & Platt, 2007). Likewise, human evoked potentials show number-specific responses as early as 75 msecs (Park, DeWind, Woldorff, & Brannon, 2015). This indicates that number processing starts at a relatively early level. The reaction times in numerosity estimation tasks are independent of the number of elements, suggesting that numerosity is processed in parallel (Mandler & Shebo, 1982). Physiological results also argue for a parallel extraction of numerosity (Nieder et al., 2002). In addition, there is evidence for a "direct visual sense for number" since number seems to be a primary visual property like color, orientation, or motion, to which the visual system can be adapted by prolonged viewing (Ross & Burr, 2010). And finally, there is an ongoing debate about whether we have a true sense of number (Ross & Burr, 2010; Anobile, Cicchini, & Burr, 2014) or whether our apparent number sense is in fact just a variant of texture perception, namely the perception of texture density (Durgin, 2008; Dakin, Tibber, Greenwood, Kingdom, & Morgan, 2011; Raphael, Dillenburger, & Morgan, 2013).

All this shows that the understanding of numerosity should be regarded as a constitutive element for our understanding of perception and cognition. And for this it is of obvious importance to understand how numerosity is computed by the visual system. However, as yet there is no agreement upon a canonical structure for models which address the problem of numerosity perception. Rather, there exists a variety of different model approaches (e.g., (Dakin et al., 2011; Dehaene & Changeux, 1993; Meck & Church, 1983; Stoianov & Zorzi, 2012; Verguts & Fias, 2004; Zetzsche & Barth, 1990b)), and it is unclear how they are exactly related to each other, and whether they all share some common basis. It is also important to note that some of these models cannot be considered as full computational models that can be realized in a neurobiologically realistic fashion. Such models have to account for the complete processing chain, from the retinal image, over neurobiologically plausible transformation stages, to the final number assignment.

The first model that matched these criteria has to our knowledge been suggested by (Zetzsche & Barth,

1990b), and their earlier results constitute an important basis for our present analysis. A very influential model that also matches the criteria (up to the point that it is only a 1-D model) has been suggested by Dehaene and Changeux (Dehaene & Changeux, 1993). This model is based on different processing stages, with the essential components being Difference-Of-Gaussian (DOG) filters of different sizes and an accumulation system. The DOG filters restrict the model to represent all stimuli as blob-like features, which limits its invariance properties with respect to the shape of the elements in a stimulus. Whether human observers show related deviations from invariance which can be attributed to a blob-like representation remains to be determined.

A more recent model by Dakin et al. (Dakin et al., 2011) uses texture density computed by a ratio of pooled high and low spatial frequency filter outputs. Assuming that the high spatial frequencies are largely determined by the number of objects, an estimate for numerosity is determined by the product of area and density. By definition the pooled high frequency output depends on the length of the object contours presented in the stimulus. The model tests so far used stimuli consisting of similar objects such that numerosity is approximately proportional to accumulated contour length. The degree of invariance of the model with respect to comparisons involving elements of very different shape, and therefore substantially different contour length, has not yet been systematically tested.

Another model by Stoianov and Zorzi (Stoianov & Zorzi, 2012), which can also be seen to match the criteria, is based on unsupervised learning in a deep recurrent network. Neural network models are very valuable in so far as they demonstrate that a capability like numerosity perception can indeed be learned by a biological system (Hornik, Stinchcombe, & White, 1989). The training images were binary and the elements presented had rectangular shapes, so that only moderate shape variations were investigated. Whether this model and its abstract mathematical counterpart (Stoianov & Zorzi, 2012; Cappelletti, Didino, Stoianov, & Zorzi, 2014) can provide the desired invariance properties for arbitrarily shaped elements thus remains to be determined.

In our view, this situation suggests that a systematic account of the logical, mathematical, and computational requirements for a ”sense of number” is highly desirable. A formal analysis through a hierarchy of abstraction similar to Marr’s approach (Marr, 1982) and a discussion of the relation to the philosophy of mathematics can be found in (Kluth & Zetzsche, 2015). Simplified early variants of the model which represent objects as ”‘rectangular”’ polygons and make use of a generalized eigenvalue operation are described in (Zetzsche & Barth, 1990b; Zetzsche, Gadzicki, & Kluth, 2013). First results on the present version of the model have been presented in a conference paper (Kluth & Zetzsche, 2014). In this paper the focus is on the invariance properties of the model, on its neural realization, and on its quantitative predictions regarding human behavior.

The paper is organized as follows. We start with a specification of the preconditions and of the general problem statement, and present then the specific questions that have to be answered. In the following section we describe our approach to the problem: numerosity should be considered as a *mathematical invariant*, and the zeroth Betti number, a specific topological invariant, should be considered as the ideal solution. We then derive a realistic approximation of the ideal solution by considering the image luminance function as a curved surface and by making use of the Gauss-Bonnet theorem to compute the number of simply connected components in the image. This solution is then tested for a variety of differently shaped elements being arranged in varying configurations. In the next step we consider the neurobiological plausibility of the suggested computations and show that required hardware can be assumed to be available in the visual cortex. The relation of the model to human behavior is investigated in the following section. Here we investigate how reasonable assumptions about the neural noise lead to predictions about the behavior in different tasks and about the corresponding Weber

fractions. The paper is closed with a discussion in which we compare our model to other suggested models, identify the different invariance properties of the models, and consider what testable predictions can be deduced form this comparison.

Mathematical Principles and Models

Definition of the Problem

The common mathematical basis for numbers is given in set theory. The cardinality of a set is closely related to our understanding of what is meant by *numerosity* as it describes how many elements the set contains. But representations of numbers seem to be somewhat arbitrarily chosen because numbers are just a formal construction to provide a label for the equivalence classes of the relation "having the same cardinality". The definitions of the natural numbers and their Arabic digits are both concepts which can have an effect on numerical cognition. However, they do not provide a basis for the mental representation of numbers as species without access to these concepts are also able to estimate the number of objects in a configuration. Another problem is that the concept of cardinality in set theory is based on a clear distinction between the elements within a set, and this cannot be assumed as a general property of the perceptual process. As a consequence, set theory does not provide the tools and concepts to properly describe how the number of objects in the real world is derived by perceptual processes from basic properties of these objects. The biological representation of number a derived from perceptual processes is also referred to as numerosity.

In the context of the *perception* of numerosity we thus have to consider a perspective which is different from the abstract mathematical realm.. Here we are confronted with the problem of inferring numerosity from the physical world. Constitutional aspects of this problem are the notions of space and object. We envisage space here as the continuous real valued space $\mathbb{R}^3$. We further assume that this 3-D space is not completely empty but contains a configuration of matter, i.e. at each position (x, y, z) we have either some matter of type m_j (with multiple assignments not allowed) or empty space (vacuum). Is it possible to assign in a meaningful way a *numerosity* to this spatial configuration of matter? And what are the requirements for such an assignement? The critical concept here is the notion of an object. We think here of ordinary objects like dogs, trees, tables, etc.. But what makes up an object formally? At least since Descartes a very common conception of a physical object (body) is that of a contiguous bounded region of matter in 3-dimensional space. This region is distinguished from its surround. The approach of establishing objects by such a contiguous region of matter is also supported from a perceptual point of view by evidence from Gestalt theory. The connectedness of pieces strongly affects grouping into one object (Palmer & Rock, 1994). Moreover, it has been shown that this kind of connectedness also has an effect on numerosity estimation (Franconeri, Bemis, & Alvarez, 2009; He, Zhang, Zhou, & Chen, 2009). One critical issue here is how the distinction between object and surround is established. For the context of this paper we will not further pursue this question and will just assume that this is achieved in some reasonable fashion. An object is thus restricted to a connected subset of the real world which means that it cannot be represented by a union of elements, i.e. disjoint subsets.

The world in which we live can be assumed to be four-dimensional if we consider the time as an additional dimension. Within the context of this paper we restrict the world and the objects to be static such that they can be represented by and within a three-dimensional vector space, the real valued space $\mathbb{R}^3$. We then have our configuration of one or more contiguous regions of 3-D space and we can assign a label to each point (x, y, z)

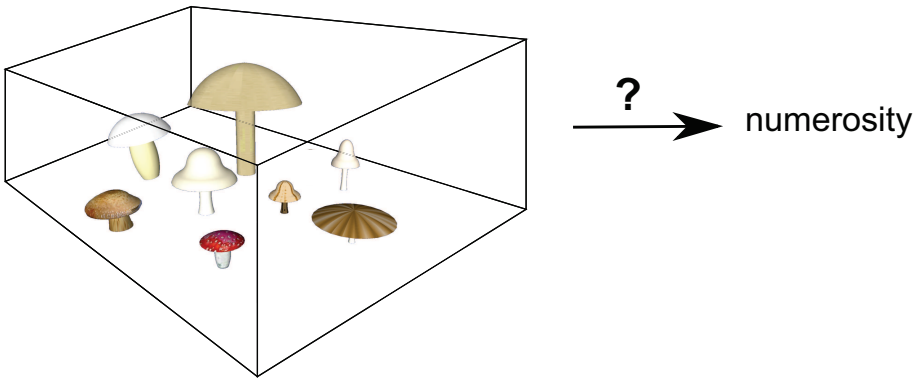


Figure 1: A variety of three-dimensional simply connected objects are illustrated and build a scene. The main question is, how the numerosity of objects can be determined from such a scene. In a strict manner this illustration gives only an idea about the three-dimensional case because they are not able to show the general three-dimensional case as mapping the setup to a two-dimensional image is automatically the application to estimation via sensors. The shown objects (mushrooms) differ in their position, size, and shape. There exist multiple possible approaches of how the numerosity can be estimated. If time is not limited one may sequentially count the objects. For the case of up to four objects the number can be determined immediately, named subitizing. We want to deal with the full range of numbers and we therefore investigate the approximate number sense which estimates the numerosity under a strict time limitation.

of the interior of the region belonging to an object, the objecthood label. We further assume that there exists something like a "background" in form of the complement of the union of all object regions (in physical terms this could be air, water, vacuum, etc.), and we can assign the background label to all points of this background. For simplicity, let us also assume that any two objects can come arbitrarily close together but can never have full contact so that there is always at least a tiny bit of background in between them. A configuration $c(x, y, z)$ can then be defined by the associated objecthood function. In mathematical terms this means, if multiple objects O_i are given as subsets of the $\mathbb{R}^3$ building a configuration set $C = \cup_i O_i$, the objecthood function defining a configuration becomes the characteristic function with respect to C. Figure 1 shows an example for a possible configuration.

Having specified the formal conditions so far, we now address the following questions:

1. How can we assign a *numerosity* to this configuration in a mathematically reasonable way?
2. How can we compute this numerosity from visual measurements?
3. How can this computation be realized with the available neural hardware of the human visual cortex?
4. How does all this relate to the human perception of numerosity?

Numerosity is an *Invariant*

As a first step we consider the mathematical problem behind the number assignment problem addressed by the first question. Let us consider the set $\mathcal{C}$ which has all possible configurations $c(x, y, z)$ as elements, i.e. all the configurations of objects in the above specified sense. The relation $\sim$ defines the *equinumerosity* relation, i.e. two configurations are equinumerous if each configuration has the same numerosity. The set $\mathcal{C}$ is then structured into subsets according to the equivalence relation $\sim$ which means that each subset is an equivalence class with respect to $\sim$. Let us define the property $N(c(x, y, z))$ as the numerosity of objects in the configuration

c. Since whenever $c_i \sim c_j$ then $N(c_i) = N(c_j)$ for two configurations c_i and c_j, the property N, respectively the numerosity, is an *invariant* under the relation $\sim$. Note that the fundamental difference to the classical set theoretical approach is as follows. Given a configuration c with respect to a union of objects O_i, we only have access to $C = \cup_i O_i$. But determination of cardinality in a set theoretical manner would require access to $\{O_1, \ldots, O_n\}$, i.e. an explicit distinction between the objects. So let us look somewhat closer on the nature of the invariant N.

Invariant properties are usually specified with respect to transformations T which map an object configuration to another one. Which transformations T have to be considered? The simplest class are obviously changes of positions (Figure 2a), and of course the numerosity $N(c(T[x, y, z]))$ should not be influenced by such transformations T. The same is true for other geometric properties, like size and orientation changes (Figure 2b,c), and in general every affine geometric transformation T should be admissible while leaving the number value invariant. However, geometric transformations are not sufficient since in addition to changes of the geometric properties of the objects it should also be possible to change their *shapes*. Is there a class of transformations which enables an arbitrary change of the shapes of the objects? And what is the appropriate mathematical formalism that can provide suitable invariants with respect to this class of transformations? We suggest that the appropriate formalism is provided by *topology*. Loosely stated, topology is the mathematical discipline that deals with those problems that do **not** depend on the exact shape of the objects and configurations involved. The topological structure of the support of a configuration or more generally a topological space is described by the series of Betti numbers. The k-th Betti number is the rank of the k-th simplicial co-homology group. A more intuitive interpretation of this number is that it measures the number of k-dimensional holes of the space. The zeroth Betti number is the most interesting one with respect to numerosity estimation as it is the number of connected components. Each configuration consists of multiple contiguous objects so that the zeroth Betti number of the support C is equivalent to the number of objects.

It is interesting to ask whether the proposed invariant is a topological invariant. Then it is obvious that the operations are restricted to homeomorphisms, i.e. continuous and bijective mappings whose inverse is also continuous. This class of operations not only includes the geometric transformations illustrated in Figure 2a,b,c but also allows the complete change of shape appearance of the objects as illustrated in Figure 2d. Homeomorphisms are structure conserving in the sense that the kind of connectedness does not change. Cutting, tearing, and gluing of objects, for example, are no homeomorphisms in general and could cause a change of a topological invariant. The proposed invariant has the desired property as it can be identified by the zeroth Betti number, which is a topological invariant. But the zeroth Betti number can be seen as a "stronger" invariant as it does not change its value for a broader class of operations. This class of operations allows non-homeomorphic transformations like the mapping from a sphere to a torus without changing the invariant's value. It is only important that each object remains somehow connected.

There is a class of closely related algorithms in digital topology, i.e. the research area dealing with the computation of topological properties in image processing. This class of algorithms is usually referred to as "connected component labeling". As the name states, the algorithms are based on a labeling strategy, and the number of connected components results as a by-product from the number of different labels having been assigned on termination of the algorithm. The main problem with these algorithms is that current knowledge of numerosity perception assumes a parallel, almost instantaneous process. To our knowledge, there are no hints on a specific temporal dependence on reaction times for numerosity estimation. The only exception is sequential

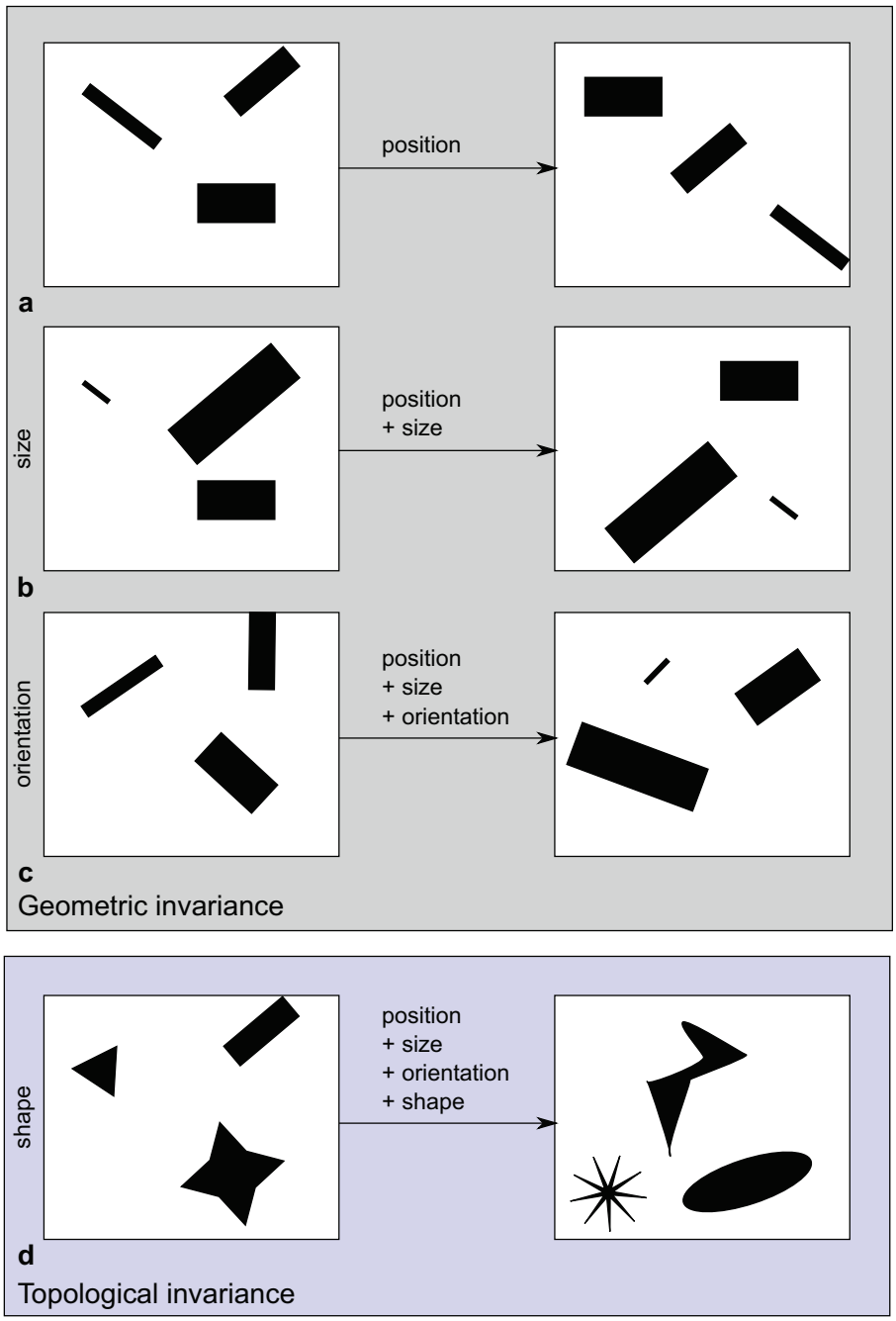


Figure 2: The different kinds of invariance properties required for numerosity estimation are illustrated in the two-dimensional case. The left column shows the different operations to which the numerosity should be invariant in isolation. The right column shows the combination of different operations with increasing complexity from top to bottom. The kind of operations are partitioned in two groups, geometric invariances and topological invariances. Geometric invariances are defined with respect to operations which change position, size and orientation of the objects. The topological invariances are additionally defined with respect to operations which change the shape of the object (formally defined with respect to homeomorphisms). Each topological invariance is automatically a geometric invariance.

counting, respectively mental counting, which is assumed to rely on different mechanisms (Mandler & Shebo, 1982; Trick & Pylyshyn, 1994). However, the known algorithms from digital topology for computation of the number of connected components have all a substantial serial component. Many of them show a direct runtime dependence on the number of objects (see, e.g., (He, Chao, Suzuki, & Wu, 2009)), whereas it is a hallmark of numerosity perception that the reaction times are independent of the number of objects in the stimulus. In the moment it is not possible to determine the number by a connected component labeling algorithm which is compatible with our current knowledge on numerosity perception and the neural architecture of the visual cortex. However, this does not at all rule out the possibility that new insights on the former or the latter, or both, will turn up in the future. This issue should not be removed from the research agenda.

An important special case results from the assumption that all objects are approximately simply connected, i.e. objects have no holes. In this case the zeroth Betti number equals another important topological invariant, the Euler characteristic. The important point to note is that whereas the former formal argumentation enables a precise mathematical specification of number for a given configuration of physical objects (provided the assumptions are valid), the following steps are prone to approximations and thereby inevitably entail *deviations* and *errors*, as compared to the ideal solution. In particular, we derive the proposed computational model, which is applicable to images, from the Euler characteristic.

Sensor Implementation

Estimating the numerosity of objects requires access to the Euler characteristic of the object. The challenging part of the information gathering process is, how this quantity can be estimated by a given modality. Given the abstract formulation of the problem it is a comfortable starting position. The main focus is to identify the relation between the necessary information and the information provided by a sensor. Sequential tactile scanning of the object's surface or visual sensing are imaginable sensor strategies for human number estimation. Multisensory approaches would also be possible and with additional technical sensor devices, as they are used in robots, the number of possibilities for number estimation becomes quite high. In this article we restrict the investigation to the human visual system. Here we restrict the visual stimulus to be luminance only such that we have a sensory representation of the scene by the luminance function $l = l(x, y)$. An example for the luminance function of a slightly lightened three-dimensional cube is illustrated in Figure 3a. But it remains unclear how this is related to the real world. The interplay between lighting and reflectance properties of the object's surface result in a mapping from the real world to the sensed luminance function. For further considerations in this direction see (Phong, 1975; Schlick, 1993) and for its relation to human perception, we refer to (Koenderink & Doorn, 2003). However, for our investigations we make a very simplified assumption on the real world to the luminance mapping. The sensor operator G maps this physical properties of a configuration c to the luminance function $l : \mathbb{R}^2 \to \mathbb{R}_0^+$ such that the sensory process becomes $G(c) = l$. Assuming G is an orthogonal projection not incorporating lighting the resulting luminance level is always constant for the objects. In particular, this case for objects resulting in right-angled polygons as projections was considered in previous works (Zetzsche & Barth, 1990b; Zetzsche et al., 2013). If we assume planar surfaces to be projected this way, the resulting setup matches common visual stimuli in psychological studies, e.g. (He, Zhang, et al., 2009; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004). However, we assume that G takes lighting of the objects into account such that it does not result in a constant luminance level on objects. Additionally, we assume the operator G to preserve the differentiability on the objects surface. The configuration c is thus mapped to an almost every-

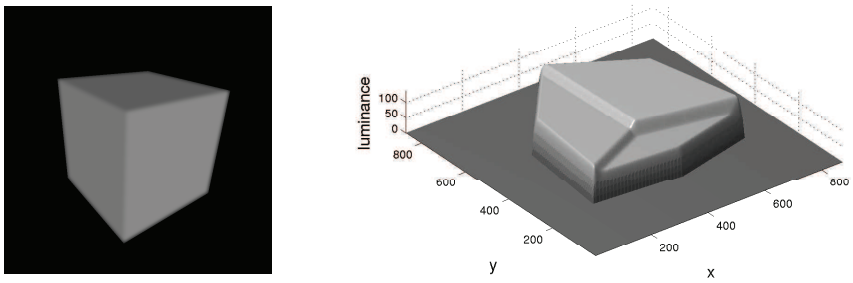


Figure 3: In (a) the image of a slightly lightened three-dimensional cube is shown. (b) shows the luminance function of the image in (a) as a two-dimensional surface embedded in the three-dimensional space. Applying differential operators to the luminance surface (b) becomes numerically instable because of emerging discontinuities which require the incorporation of additional arcs in the general solution.

where sufficiently smooth function l. The background is assumed to be uniform and clearly separated from the objects. The discontinuity between objects and the environment is thus projected to the luminance function, cf. Figure 3a. In the following we apply the Gauss-Bonnet theorem to the luminance function in order to estimate the numerosity of perceived objects, i.e. the Euler characteristic is related to local features of the luminance function. We investigate the luminance surface $\Omega \subset \mathbb{R}^3$ which is defined by the perceived luminance function l, i.e. $\Omega := \{(x, y, z) \in \mathbb{R}^3 | x, y \in \mathbb{R}, z = l(x, y)\}$. For example, in Figure 3b the luminance surface of the luminance function in Figure 3a is shown.

The Gauss-Bonnet theorem provides a connection between topology and differential geometry and relates the Gaussian curvature of two-dimensional surfaces to their Euler characteristic. It thus becomes reasonable to represent the number in terms of the Euler characteristic with the aid of the following theorem.

Theorem 1 (Gauss-Bonnet) *Let $S \subset \mathbb{R}^3$ be a regular oriented surface (of class C^3), and let R be a compact region of S with the boundary ∂R. Suppose that ∂R is a simple, closed, piecewise regular, positively oriented curve. Assume ∂R consists of k regular arcs ∂R_i (of class C^2), and let θ_i be the external angles of the vertices of ∂R. Then*

$$\int_R K \, dS + \sum_{i=1}^{k} \int_{\partial R_i} \kappa_g \, ds + \sum_{i=1}^{k} \theta_i = 2\pi\chi(R) \tag{1}$$

where K is the Gaussian curvature, κ_g is the geodesic curvature, and χ is the Euler characteristic.

The Euler characteristic is a topological invariant which maps a number to any subset within a topological space. This number then characterizes the kind of connectivity of the subset. For example, the surface of a sphere (no holes) has a different characteristic number than the surface of a torus (one hole). However, being a topological invariant, the Euler characteristic of an object stays constant if a homeomorphism is applied to it.

Theorem 1 provides a solution which incorporates the luminance surface properties directly. Assuming a smooth space curve Γ which is the boundary of a closed region S on the luminance surface Ω, i.e. $\Gamma = \partial S$, it

results

$$\int_S K\, dS + \int_\Gamma \kappa_g\, ds = 2\pi\chi(S), \tag{2}$$

where κ_g denotes the geodesic curvature and the Gaussian curvature K is given by

$$K(x,y) = \frac{l_{xx}(x,y)l_{yy}(x,y) - l_{xy}(x,y)^2}{(1 + l_x(x,y)^2 + l_y(x,y)^2)^2}, \tag{3}$$

where the subscript letters denote the derivative in the respective direction. The consideration of the space curve Γ is necessary because the integral of the Gaussian curvature over the whole luminance surface of arbitrary luminance functions would result in the same quantity. This means that in this case all images have the same Euler characteristic (Barth, Ferraro, & Zetzsche, 2001). Using the parametrization $\phi(x,y) := (x, y, l(x,y))^T$ of the surface $S \subset \Omega$ and the Gaussian curvature given by equation 3, the first integral can be computed by

$$\int_S K\, dS = \int_{\mathbb{R}^2} \underbrace{\frac{l_{xx}(x,y)l_{yy}(x,y) - l_{xy}(x,y)^2}{(1 + l_x(x,y)^2 + l_y(x,y)^2)^{3/2}}}_{=:\tilde{K}(x,y)} \chi_S(\phi(x,y))\, d(x,y), \tag{4}$$

where χ_S is the characteristic function with respect to the set S. In order to calculate the second integral, the geodesic curvature of the boundary curve Γ must be estimated. Using the parametrization $\theta : I \to \mathbb{R}^3$ with $\theta(t) := (x(t), y(t), l(x(t), y(t)))^T$ of the boundary curve and the assumption $l(x(t), y(t)) = const.,\ \forall t \in I$, we can determine the geodesic curvature by

$$\kappa_g(t) = \tilde{\kappa}_g(x(t), y(t)) = \frac{l_x^2 l_{yy} + l_y^2 l_{xx} - 2 l_x l_y l_{xy}}{(l_x^2 + l_y^2)^{3/2}(1 + l_x^2 + l_y^2)^{1/2}}. \tag{5}$$

The additional assumption of constant height allows to eliminate the second derivatives of the parametrization. The second integral in equation 2 thus becomes

$$\int_\Gamma \kappa_g\, ds = \int_{\mathbb{R}} \frac{l_x^2 l_{yy} + l_y^2 l_{xx} - 2 l_x l_y l_{xy}}{(l_x^2 + l_y^2)^{3/2}(1 + l_x^2 + l_y^2)^{1/2}} (x'^2 + y'^2)^{1/2} \chi_\Gamma(\theta(t))\, dt, \tag{6}$$

where χ_Γ is the characteristic function with respect to the set Γ. Assuming constant height, the first derivative of the parametrization yields $x' = -(l_y/l_x)y'$. Consequently, the geodesic curvature depends only on differential operators applied to the luminance function and one first derivative x' or y' of the parametrization of the boundary curve. Finally, the numerosity of objects N from a union S of pairwise disjoint regions of luminance surfaces can be determined by

$$2\pi N = \int_{\mathbb{R}^2} \tilde{K}(x,y)\tilde{\chi}_S(x,y)\, d(x,y) + \int_{\mathbb{R}} \tilde{\kappa}_g(x(t), y(t))\tilde{\chi}_\Gamma(x(t), y(t))\, dt, \tag{7}$$

where $\tilde{\chi}_S(x,y) := \chi_S(\phi(x,y))$ and $\tilde{\chi}_\Gamma(x(t), y(t)) := (x'^2 + y'^2)^{1/2}\chi_\Gamma(\theta(t))$. The computation now depends on three quantities $\tilde{K}$, $\tilde{\chi}_S$, $\tilde{\kappa}_g$ depending on surface properties and the quantity $\tilde{\chi}_C$ which has to extract curve properties. All quantities have in common that they represent local properties of the luminance surface. Regarding the implementation, images are assumed to be positive real-valued matrices. In order to replace the continuous operators by their discretized approximations we have to guarantee sufficient smoothness on the

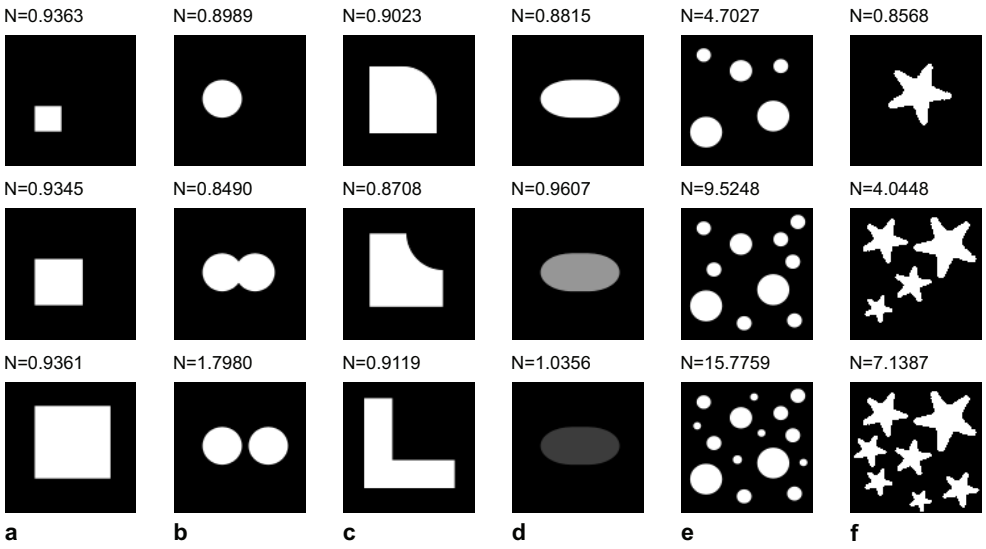


Figure 4: Various stimuli (size 100 × 100 pixels) and the corresponding responses of the model defined by Equation 7 are illustrated. Stimuli vary with respect to different criteria like cumulative area (a), number change by morphing (b), convex and concave shape (c), contrast of a single object (d), large numbers of convex (e) and concave (f) objects.

luminance surface. If this is not satisfied, high numerical errors can emerge. In order to obtain the differentiability in the discrete representation, each stimulus is additionally low pass filtered using a Gaussian kernel $g : [-1,1] \times [-1,1] \to \mathbb{R}$ with fixed standard deviation $\sigma = 0.04$ for all computations considered within this article. The images are assumed to be defined by functions $l : [-1,1] \times [-1,1] \to [0,1]$. The discretization assumed to be done with a step size of $\Delta x = \Delta y = \frac{1}{50}$ results in image matrices with 100×100 pixels. The threshold height is determined by half the maximum luminance value within an image.

The model's behavior for stimuli commonly used in vision research is illustrated in Figure 4. More complex stimuli are shown in Figure 5. The model shows a strong invariance with respect to cumulative area of the region covered by objects as can be seen in Figure 4a. Furthermore the model is also invariant to changes in object shape from convex to concave objects as can be seen in Figure 4c. The second column, i.e. Figure 4b, shows a circle morphed into two circles. The slight change from the top stimulus to the middle one is due to numerical instability at the connecting points of the contours. The model shows a strong change in the response when the two circles are not connected anymore, cf. the bottom stimulus in Figure 4b. This selectivity with respect to numerosity is one of the most important properties of the model. It is also present for larger numbers of objects as can be seen in Figure 4e and f.

By definition the proposed model is invariant to contrast to some degree. The threshold which is defined with respect to maximum contrast causes the invariance illustrated in Figure 4d. The slight changes are caused by numerical differences in the curvature computation which highly depends on the absolute level of luminance. The contrast invariance is also present for multiple objects as can be seen in the top and middle stimuli of Figure 5f. The bottom stimulus shows objects with different contrast. In this case the threshold to determine the integration domain was chosen smaller, i.e. 30% of maximum luminance (denoted by N_t). The invariance with respect to contrast is preserved as the model output changes only slightly. Note that the model can only be contrast invariant to some degree. The lower the threshold the higher is the influence of the standard deviation

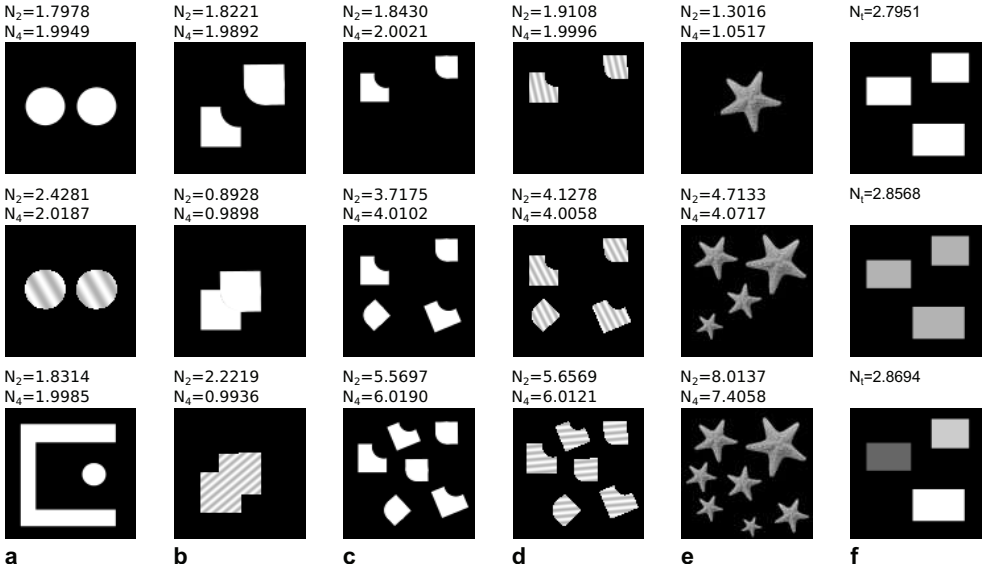


Figure 5: Various stimuli (size 100 × 100 pixels) and the corresponding responses of the model defined by Equation 7 are illustrated (N_2: $\sigma = 2\Delta x$; N_4: $\sigma = 4\Delta x$; N_t: $\sigma = 2\Delta x$ and the threshold is 0.3 times the maximum luminance value). Stimuli vary with respect to different criteria like grating texture (a), convex and concave object morphing (b), increasing number of convex and concave binary (c) and grating-textured objects (d), textured concave objects (e), and contrast variations of multiple rectangles (f).

σ of the Gaussian filter. This means that the minimum spatial distance between two objects which is necessary to distinguish them in terms of numerosity is increased. One should find a systematic underestimation in the numerosity estimation when using stimuli which include a mixture of low and high contrast objects spatially placed with the critical distance for constant contrast. For further illustrations of illuminated 3D-objects and threshold dependent responses we refer to (Kluth & Zetzsche, 2014).

We also tested the model on more complex stimuli including various numerosities of convex and concave objects with different luminance patterns (constant, sinusoidal gratings, and texture), cf. Figure 5a-e. In all cases with piecewise constant luminance patterns the model ($N = N_2$), which uses the parameters for further investigations within this article, works well and the desired selectivity and invariance properties are present. In a few cases the model fails. In particular the introduction of sinusoidal gratings (cf. Figure 5a (middle), b (bottom), and d) or texture (cf. Figure 5e) on the objects results in larger deviations from the expected model output. More sophisticated luminance patterns have a larger spatial region which requires a certain degree of regularity for the differentiation. If the regularity is not given, the size of this spatial region strongly determines the amount of error within the computation of the differentials. We tested this error cause by increasing the standard deviation of the Gaussian filter to increase the regularity of the differentiated signal. By using twice the standard deviation (N_4; $\sigma = 4\Delta x$) of the previously used one ($N = N_2$; $\sigma = 2\Delta x$) the model (N_4) becomes a perfect enumerator (deviations smaller than 0.02) for constant and sinusoidal luminance patterns as can be seen in all cases of Figure 5a-d. The deviations in the texture case are slightly higher which is due to the higher degree of regularity which is required on the objects. Further increasing the standard deviation solves this issue but could cause a spatial pooling of close objects or an exclusion of small objects. For $\sigma = 6\Delta x$ and a threshold of 30% maximum luminance the outputs for the middle and the bottom stimulus of Figure 5e become

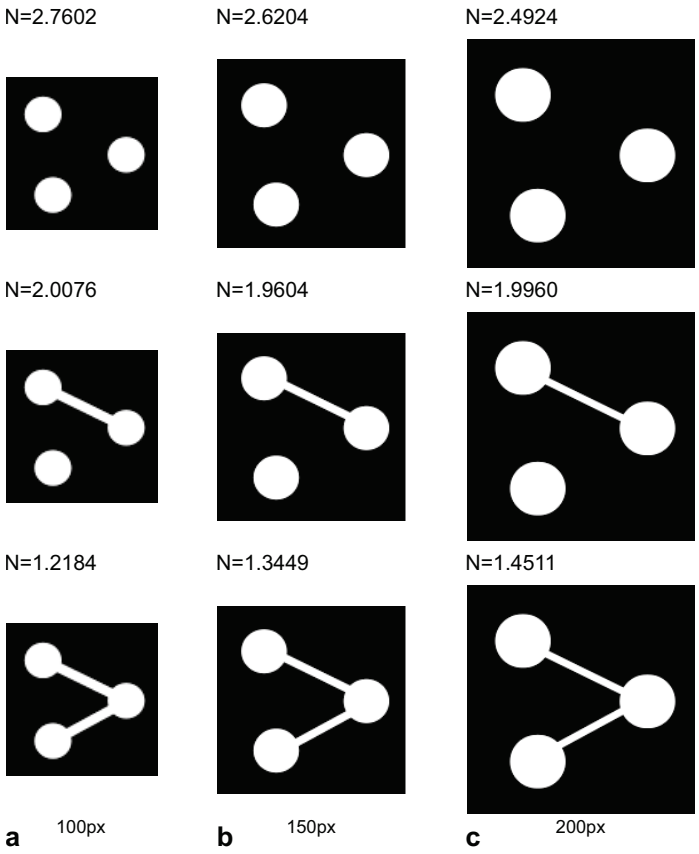


Figure 6: Various stimuli of three circles with different number of connecting lines and the corresponding model responses are illustrated. The image size is varied from 100×100 pixels (a) to 200×200 pixels (c). The relative size of the circles is constant. The absolute width of the connecting lines is constant resulting in a decrease in relative width to circle size.

4.0145 and 7.0336. Thus spatial pooling has no effect in this case such that the model output is improved by changing these parameters.

That connectedness matters was also shown in (Franconeri et al., 2009; He, Zhang, et al., 2009). In Figure 6a the kind of stimuli used in this study and the corresponding model responses are illustrated. A line connecting the circles strongly affects the response of the model such that the response is reduced with every inserted line from the top to the bottom stimulus. This implies that the perceived numerosity should be underestimated by the introduction of connecting lines. This is in line with the findings reported in (He, Zhang, et al., 2009). The stimuli are compared to stimuli with a higher image size, cf. Figure 6a-c. The relative size of the circles is constant for all stimuli but the relative width of the connecting lines varies from column a to c. The model responses in each line of the figure change which is due to numerical issues caused by a combination of the constant absolute standard deviation of the Gaussian filter g and the change of the size of the integration domain. In each column we can observe an approximately constant difference in response from top to bottom which illustrates the desired selectivity of numerosity estimation with respect to the number of objects present in the stimulus. We can also observe that the constant difference decreases from the left column to the right one. This effect could additionally be caused by the decreased line width resulting in higher numerical errors in the gradient computation at the connecting points of the lines and circles. This implies that the sensitivity to

the connecting lines emerges up to a certain line width. The thinner the connecting lines, the less influence they have.

Neural Implementation

A crucial question which immediately arises for any model being based on heavily theoretical arguments is whether such a model is still neurobiologically plausible. In order to investigate how the computations required by our numerosity model can be provided by the neural hardware available in the visual we derived a possible architecture for the implementation of Equation 7 which is illustrated in Figure 7. The processing direction is from left to right. First the input is filtered with a 2D-Gaussian kernel. The filtered input is then used to compute the first- and second-order derivatives of the luminance function, as well as the threshold value used to control the boundary curve. The computation of the derivatives builds the second path (blue block "Linear filter") with its origin in the filtered input image. The threshold value in the model depends on the maximum of the whole Gaussian-filtered luminance function. Given the derivatives of the luminance function, a bunch of multiplications must be realized, compare the red block "GC-Product" in Figure 7. The resulting products then must be summed (blue block "GC-Sum"). The output of this stage is fed into a ratio computation stage (block "GC-Norm"). After this stage, we finally have two local nonlinear features available, $\tilde{K}(x, y)$ and $\tilde{\kappa}_g(x, y)$. These are assumed to be computed across the whole visual field but are selectively gated by the threshold mechanism such that $\tilde{\kappa}_g$ values are only passed at the boundary locations, and the $\tilde{K}$ values are only passed within the interior regions to the final global summation stage (blue block "Integration"). This global summation provides the estimate of the numerosity of the input pattern.

Now let us consider the neurophysiological plausibility of these operations in a step-by-step fashion: The Gaussian filtering operation is commonly assumed to be available from the earliest stage of the visual system, in particular it could be realized in the Ganglion cells of the retina (Kuffler, 1953; Marr & Hildreth, 1980). The key operations in the model are directional derivatives of first and second order (combined with the Gaussian filter for regularization). Translated in receptive field properties these are orientation-selective mechanisms with odd and even symmetry. It is well known that Gaussian derivatives are well suited for the description of neurons in the primary visual cortex , e.g. (Koenderink & Van Doorn, 1990; Lindeberg, 2013; Marr & Ullman, 1981; Martens, 1990; Young, 1987; Young & Lesperance, 2001). In particular, it has been argued that they can be approximated in a plausible fashion as DOOG filters (Difference Of Offset Gaussians, (Young, 1987; Young & Lesperance, 2001)). And it has been suggested that curvature-selective operators can thus be easily realized by the available cortical hardware (Zetzsche & Barth, 1990a, 1990b).

In addition to the linear filtering operation this requires the nonlinear AND-like multiplication of two signals. One possibility that this could be directly realized in the dendritic tree of a neuron (Mel, 1993; Koch & Segev, 2000). Alternatively, combinations of subunits could be used to realize the Babylonian trick and compute the product by the sum of squared spatial filter outputs (Adelson & Bergen, 1985; Resnikoff & Wells Jr, 2015; Zetzsche & Barth, 1990a), i.e. $ab = 1/4[(a + b)^2 - (a - b)^2]$. Finally, it can be shown that the nonlinear mechanism of cortical gain control can be combined with subsequent nonlinear transducer functions to realize an AND-like interaction (Zetzsche & Nuding, 2005). Thus there exists no principle obstacle to prevent the neural implementation of a multiplicative interaction. Furthermore, it has been shown that certain forms of end-stopping, of the hypercomplex property, and of extra-classical receptive field properties show a close relation to the computation of curvature (Zetzsche & Barth, 1990a, 1990b; Zetzsche & Roehrbein, 2001; Zetzsche &

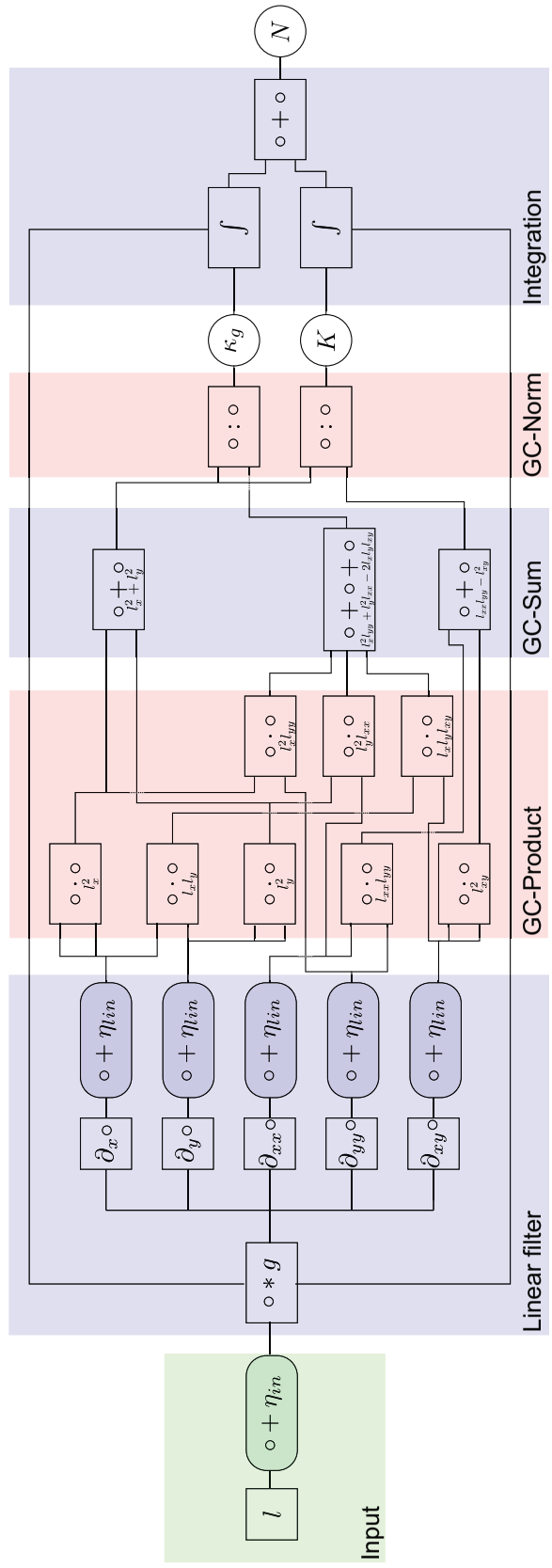


Figure 7: This figure illustrates a possible architecture of the proposed model and how the error is introduced in the system. The information precessing direction is from left to right and each box represents one processing step. The processing steps in which noise is included have a slightly darker background color. The $\circ$ in each box is a placeholder for the input coming from the left direction. Linear operation stages are blue and nonlinear ones are red. The stimulus is fed in the system as a luminance function l in the input stage (green box). Then the additive normally distributed noise $\eta_{in} \sim N(0, \sigma_{in})$ is added. The convolution with a Gaussian kernel g and applying the differential operator can be interpreted as one linear filter processing stage, i.e. the convolution with the respective derivative of the Gaussian kernel. The results of the linear filter operation are fed into the next noise-adding-unit where different samples of $\eta_{lin} \sim N(0, \sigma_{lin})$ are added to each input. The following multiplication (GC-Product), summation (GC-Sum), and division (GC-Norm) build the cortical gain control stage which results in an estimate for the required curvature quantities. Finally the curvature quantities are integrated (blue box) resulting in an estimate for the numerosity N.

Nuding, 2005).

The resulting products then must be summed again (blue block "GC-Sum") which is an undisputed ability of neurons. The output of this operation is fed into a divisive operation. That neurons can implement such an operation is shown by the established role of the divisive normalization mechanism in models of the visual cortex (Abbott, Varela, Sen, & Nelson, 1997; Carandini & Heeger, 2012; Geisler & Albrecht, 1992; Heeger, 1992; Schwartz & Simoncelli, 2001). It has even been argued that such an operation should be regarded as a "'canonical"' neural computation (Carandini & Heeger, 2012). The gain control layer, which is illustrated by the rightmost red box in Figure 7, results directly in an estimate for the required curvature quantities $\tilde{K}(x, y)$ and $\tilde{\kappa}_g(x, y)$.

Although it can be assumed that these features are computed throughout the visual field, the implementation of equation 7 requires a selective gating before feeding them into the global summation stage. The first part of this process is the determination of a level for the boundary curve. For simplicity, we have here assumed that this is achieved by selecting the 50% level of the maximum luminance value. Neural computation of the maximum is a well-known principle discussed in the context of winner-take-all networks (Kaski & Kohonen, 1994; Mead & Ismail, 2012) electing the boundary which determines the region where the surface curvature operations setting the value for the Maximum operations are also routinely used in models of the visual cortex (Lampl, Ferster, Poggio, & Riesenhuber, 2004; Serre, Wolf, & Poggio, 2005). However, we do not put special weight to the use of this principle since for the successful application of equation 7 the specific method used to determine a boundary curve plays no crucial role. It might thus as well be based on some other mechanism, e.g. on the determination of some equilibrium level (Dayan & Abbott, 2001; Grossberg, 1988). Once the boundary is selected, what remains to be achieved is the gating operation, i.e., it has to be controlled which values are passed to the global integration mechanism. This type of operation is a special case of the general principle of neural gain modulation which is an essential neural mechanism in sensorimotor processing (Salinas & Thier, 2000), and which is also used in the visual cortex (Reynolds & Chelazzi, 2004).

To summarize: the curvature operators $\tilde{K}$ and $\tilde{\kappa}_g$ depend only on derivatives of the luminance function which can be realized by neurophysiologically realistic filters. The multiplicative "AND"-like combination of these features can be computed by a variety of plausible neural mechanisms. And the ratio operations are closely related to the well established principle of cortical gain control. The outputs of the curvature operators are globally integrated in a gated fashion, which is also an established neural principle known as gain modulation. The computation of equation 7 can thus easily be achieved by the available neural hardware in the visual cortex.

Simulation Experiments on Weber Fraction

In the following we extend the proposed model such that noise is taken into account. We modeled additive normally distributed noise $\eta_{in} \sim \mathcal{N}(0, \sigma_{in})$ at the input and neural noise $\eta_{lin} \sim \mathcal{N}(0, \sigma_{lin})$ at the linear filter outputs. Both curvature operators, $\tilde{K}$ and $\tilde{\kappa}_g$, are influenced by both types of noise. The structure of the curvature operators and their resulting quantities as products of noise affected quantities is an optimal basis for a log-normally distributed resulting quantity (Buzsáki & Mizuseki, 2014). This noise behavior was reported in several studies regarding approximate number estimation. The influence of noise in the system which computes the Gaussian curvature as well as the geodesic curvature is illustrated in Figure 7. The digital quantity described by the perfect model thus becomes an analog quantity from which the desired information must be extracted. In the following we focus on two different standard tasks to analyze the proposed model and to compare its

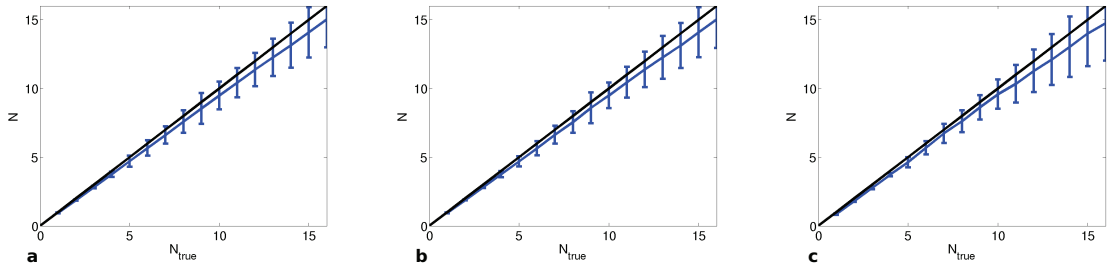


Figure 8: Mean model output with standard deviation plotted against the true number in the stimuli for the zero noise case ($(\sigma_{in}, \sigma_{lin}) = (0, 0)$) on (a) dataset (TR), (b) dataset (R), and (c) dataset (C).

performance with the behavioral results from experiments with humans. The first task is the discrimination whether it is a specific number or not (same-different), and the second task is to decide whether a stimulus is larger than a reference stimulus (larger-smaller) (Piazza et al., 2004).

All investigations are based on three synthetic datasets of binary images generated in a similar fashion as described in (Stoianov & Zorzi, 2012). The 30×30 pixel binary images contained up to 32 randomly-placed non-overlapping objects with variable surface area and shape. The cumulative area of the objects was varied from 32 to 256 pixels with a step of 32 resulting in 8 levels of cumulative area. For each tupel of number and cumulative area 200 images of size 30×30 pixel were computed and then resized to 100×100 pixel. The datasets thus consist of 51,200 100×100 binary images with 32 different numbers and 8 levels of cumulative area. The datasets are distinguished by the kind of shape and whether they are for training or testing. One dataset (R) has rectangular objects, as used for the analysis in (Stoianov & Zorzi, 2012), and the other one (C) consists of circular objects, which are commonly used in behavioral experiments. Both are used as test sets. We generated a third dataset (TR) consisting of rectangles to obtain the parameters for the optimal estimators. We trained optimal estimators tuned to a specific numerosity N_{tuned} and a fixed task. Detailed information about the optimal estimators can be found in appendix A1. See Table 1 for an overview of the trained parameters given the subsequently motivated noise levels. The joint relative frequencies of the true numerosities and the response of the estimator were obtained from two distinct datasets (rectangles (R), circles (C)). We then fitted a continuous function to the conditional relative frequencies of the estimator's response given the true numerosity to obtain the internal Weber fraction. Further technical details can be found in appendix A2.

Before we start considering noise in the system, we investigate the output of the noise-free model on each dataset. The mean model output and the corresponding standard deviation for each true number of objects within the stimulus are illustrated in Figure 8. All datasets have in common that the model is nearly the identity (black curve) and has very small standard deviations for small numbers. In this case the mathematical model is approximated well by the discretization. For larger numbers the standard deviation increases which can have two reasons. On the one hand, the mean object size decreases with a higher number of objects such that the Gaussian filter could just smooth them below the threshold. And on the other hand, with a higher number of objects the probability increases that two objects are too close such that they could be counted as one. In both cases, we expect a smaller model output which is confirmed by Figure 8 where the mean model output (blue line) falls below the identity (black line) for larger numbers of objects in all datasets. The increase of the standard deviations is in line with the increased probability that one of the previously described cases occurs for

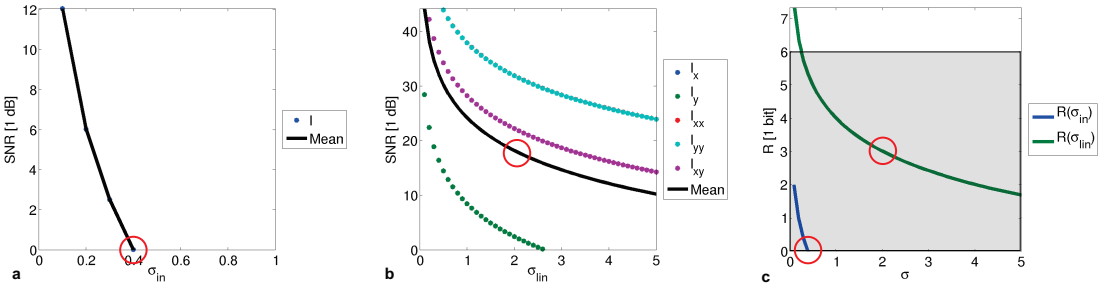


Figure 9: The SNR curves with respect to the standard deviation of the noise for the dataset (R) are illustrated for the input stage (a) and the linear filter stage (b). The plotted information of (a) and (b) can be found in Table 3. The required mean bit rates to encode the signals of each stage are shown in (c). The noise parameters chosen for the empirical investigation of numerosity judgment tasks are highlighted with red circles in (a)-(c).

Table 1: **Parameters of optimal estimators.** In all cases the noise levels were $\sigma_{in} = 0.4$ and $\sigma_{lin} = 2$.

		same-different		*smaller-larger*
$\mathbf{N}_{tuned}$	**Dataset**	**lb**	**d**	**t**
8	TR	4.6633	5.6566	9.0050
	R	4.6633	5.3333	8.6834
	C	5.6281	4.6868	8.8442
12	TR	8.2010	5.6566	11.7387
	R	8.3618	5.9798	12.0603
	C	8.8442	5.6566	12.3819
16	TR	11.0955	7.2727	14.9548
	R	11.0955	6.7879	14.7940
	C	12.2211	6.4646	15.1156

larger numbers.

The remaining free parameters to control the system behavior are the standard deviations σ_{in} and σ_{lin} of the additive noise in the system. In order to make reasonable assumptions on the noise parameters, we analyzed the datasets with respect to their signal-to-noise values (SNR) at different stages of the architecture, see Table 3. The technical details can be found in appendix A3. The SNR values at the input stage plotted against the the noise parameter σ_{in} can be found in Figure 9a. Analogously the SNR curves for the derivatives of the luminance function with respect to σ_{lin} are illustrated in Figure 9b. For further consideration with respect to the information encoded by a single unit the mean value of the curves was computed. Note that the mean curve and the curve for luminance l cover each other in Figure 9a. The same holds true for the pairs l_x/l_y and l_{xx}/l_{yy} in Figure 9b. Figure 9c shows the rate which is required to encode the signals at the different stages with the corresponding SNR without loss of signal quality. Details about the relation between SNR and the rate can be found in the appendix. This rate can now be related to findings in the literature regarding single neurons and their encoding rate of transmitted signals. The transmitted information is reported with a range from 5 bits to 30 bits per second for a single neuron (Reich, Mechler, Purpura, & Victor, 2000). Assuming a mean firing rate code and a time window of approximately 200 ms we get a rate interval of 1-6 bits which can be compared with the rates in Figure 9c (gray window). At the linear filter we thus chose the standard deviation which corresponds to half the maximum of the reported rate interval, i.e. $\sigma_{lin} = 2$. At the input stage we chose the worst case

with $\sigma_{in} = 0.4$. These two cases are highlighted with a red circles in Figure 9. These parameters were used to simulate the model behavior in both numerosity judgment tasks. Note that the mean peak signal-to-noise ratio ($PSNR$) of the first layer is approximately 40 dB higher than the respective SNR, cf Table 3. A higher rate of approximate 6-7 bits is required to guarantee the same signal quality for the whole signal amplitude. Multiple units encoding the same quantity could be used to overcome this issue.

The results of the rectangle dataset (R) for the specified noise parameter configuration are shown in Figure 10. We find that in both tasks the performance of the artificial estimators shows a very similar behavior compared to a human estimator (Piazza et al., 2004). On the linear axis shown in the first row of the figure, an increase of the variance with increasing tuned numerosity can be seen in the same-different task and a decrease in steepness with increase of tuned numerosity can be observed in the smaller-larger task. If we consider the same data on a logarithmic scale of the true numerosity, the effects are not observable anymore as can be seen in the second row. The third row shows the same data plotted on a logarithmic scale of the ratio between the true numerosity and the tuned numerosity of the estimator. For all tuned numerosities the curves are nearly identical which implies a behavior relying on the Weber-Fechner law. In the last row of Figure 10 all data points on the logarithmic scale of the ratio were also used to fit one continuous function. The tuning curves were used to obtain the internal Weber fraction w, an index to describe the discriminability between two numbers. In both tasks we find internal Weber fractions in the dimension of human behavior, 0.167 in the same-different task and 0.169 in the smaller-larger task on the rectangle dataset. In comparison Piazza et al. (Piazza et al., 2004) reported corresponding Weber fractions of 0.17 and 0.174. In summary the computational results agree excellent with behavioral results (Piazza et al., 2004, 2010; Halberda et al., 2008) and can compete with recent computational results (Stoianov & Zorzi, 2012) ($w = 0.15$ in larger-smaller task).

We found slightly different but similar Weber fractions on the circle dataset, see Table 2. In the same-different task we observe a better numerosity discrimination in the circles dataset (C) for the numbers 8 and 16 and worse discrimination for the number 16. These differences can may be explained by the optimal estimator parameters in Table 1. The parameters for the evaluation were obtained from the dataset (TR). Comparing these parameters with the parameters obtained from the datasets (R) and (C) should provide explanations for the observed differences. The parameters are determined by maximizing the true positive rate and minimizing the false negative rate, cf. appendix. For number 8 in the same-different task the parameters of (R) nearly match the parameters of the training set such that the Weber fraction of (R) is a good baseline. Compared to the interval of the training set the interval of (C) is a subset with a smaller length. This directly implies that the used parameter setting differs from the optimum of the receiver operator characteristic in the dataset (C). As the interval is a subset, the true positive rate is at least the same or higher. This can be interpreted as a constant or a lower valley in the curve in Figure 10c, for example. If the true positive rate is constant, a better Weber fraction would emerge only if the false positive rate is smaller. For the curve in Figure 10c this would imply that the values next to the tuned numerosity 8 must be higher, i.e. the curve becomes steeper. If the true positive rate is higher, the false positive rate can stay constant or decrease to explain the observed effect of a smaller Weber fraction. But a smaller false positive rate is not reasonable in this setup because the interval of the estimator is larger than the optimal interval for dataset (C). This implies that in the worst case more different stimuli are classified as the same. The most reasonable case is that the true positive rate increased and the false positive rate is constant. For number 12 the intervals are similar for all datasets resulting in similar Weber fractions. For number 16 we find a similar parameter setup compared to number 8 (optimal interval of (C) is subset of estimator interval (TR))

Table 2: **Parameters of the continuous data fitting functions.** In all cases the noise levels were $\sigma_{in} = 0.4$ and $\sigma_{lin} = 2$. The parameters of the optimal estimators were obtained from the dataset TR, see Table 1.

		same-different			*smaller-larger*	
$\mathbf{N}_{tuned}$	**Dataset**	**w**	**c**	δ	**w**	**c**
8	R	0.170	0.0968	0.412	0.158	-0.172
	C	0.14	0.137	0.408	0.150	-0.128
12	R	0.151	0.0551	0.277	0.151	-0.0502
	C	0.149	0.107	0.270	0.172	-0.015
16	R	0.150	0.0352	0.263	0.152	-0.024
	C	0.175	0.0799	0.267	0.203	0.00166
{8, 12, 16}	R	0.167	0.054	0.299	0.169	-0.0655
	C	0.176	0.104	0.303	0.197	-0.0353

but the Weber fraction is worse for (C). Here it is reasonable that the false positive rate increased resulting in a worse Weber fraction.

In the larger-smaller task the same arguments hold true for number 8 but number 12 and 16 are more complicated. The threshold of (C) does not differ much from the estimators threshold (TR) for number 16 but the difference in Weber fraction is quite high. In this task the similar threshold also implies a constant or higher true positive rate but here the distinction is done between two sets both consisting of multiple numbers. Even if the true positive rate stays constant, which is reasonable for number 16, the distribution over the whole set of larger numbers can change dramatically. This could result in a worse Weber fraction.

Finally, we went one step further and analyzed various noise parameter combinations. The resulting Weber fractions on the rectangle dataset (R) are illustrated in Figure 11. In both tasks we find a similar distribution of Weber fractions. The only difference can be observed for high noise levels in the upper right corner of the right illustrations of Figure 11. This could be caused by numerical instability of the fitting algorithm which relies on the resulting error in the numerosity quantity obtained from the model. However, the Weber fraction for these noise parameters is beyond the reported Weber fractions for humans (children ~ 0.3). In both tasks the distribution in vertical direction seems to have a discontinuity for lower linear noise σ_{lin}. This discontinuity is a result of the non continuous luminance function in the dataset. The standard deviation of the Gaussian filter is not sufficient to compensate the discontinuity in the stimulus signal function. The regularity required for the model computation (i.e. differentiability) is not guaranteed. This implies that the input noise stabilizes the system to a certain degree. Changing the standard deviation of the Gaussian filter should move the discontinuity in the Weber fraction distribution in Figure 11, in particular an increase of the standard deviation shifts this line towards the bottom of the illustartion, i.e. towards a smaller σ_{in}-value.

We also highlighted a height line for a constant Weber fraction $w = 0.15$ to demonstrate that various parameter combinations can result in the same fraction value. The qualitative difference between two cases (red and blue circle) is illustrated for the larger-smaller task in Figure 12. A smaller noise parameter at the linear filters in combination with higher noise at the input has quantitatively the same overall Weber fraction but qualitatively it does not show the same reported log-normally distributed noise behavior for all numbers, see Figure 12a. This supports the previous suggestion that the noise at the linear filter outputs which are fed into the multiplication is the essential reason for the log-normal behavior of the model, which is in line with the literature (Buzsáki & Mizuseki, 2014).

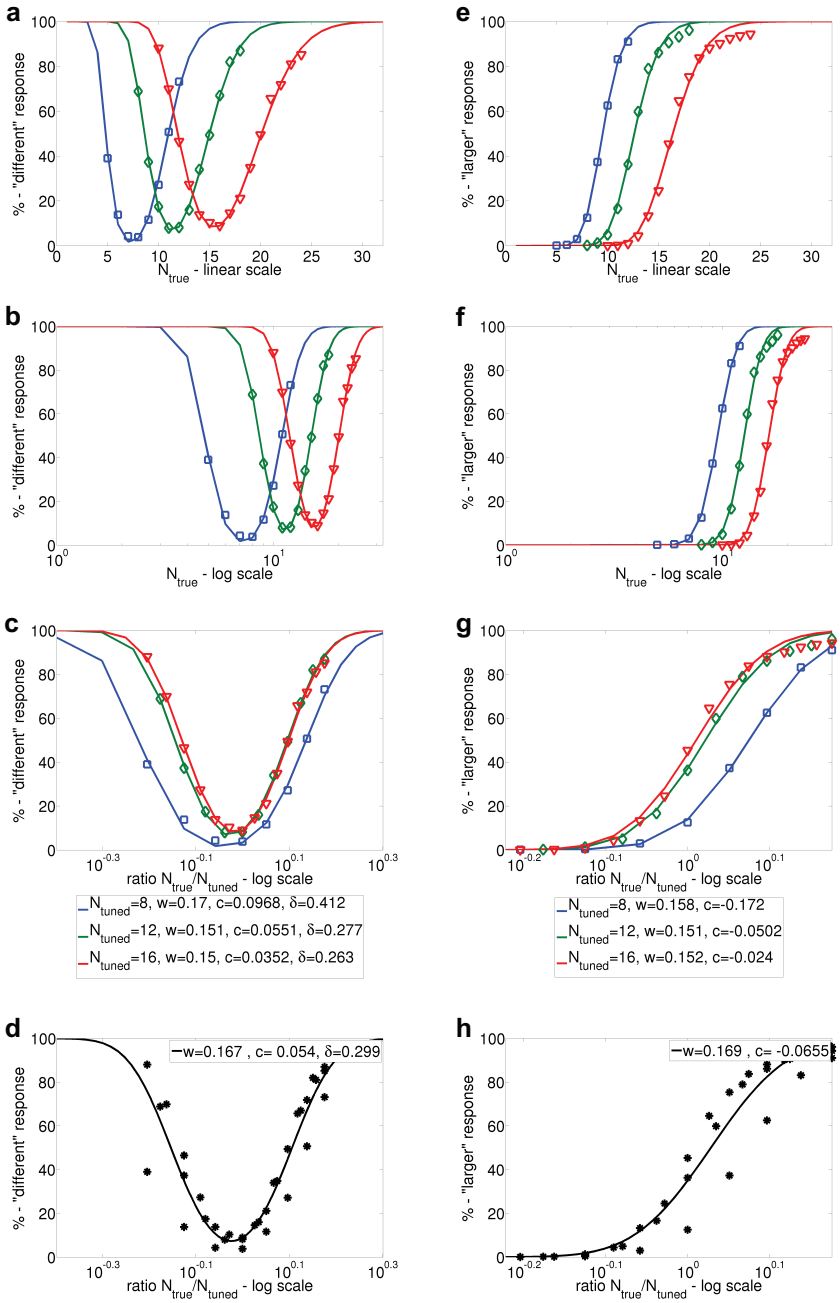


Figure 10: Computational results from same-different and larger-smaller tasks where the estimators were trained on the TR dataset and the evaluation was done on the R dataset. The noise parameters were $\sigma_{in} = 0.4$ and $\sigma_{lin} = 2$ and the corresponding estimator parameters can be found in Table 1. In the left column the graphs show the relative frequencies of the estimators response to “different” and in the right column the relative frequencies of the response “larger” are shown. In both cases the relative frequencies are plotted against functions of the true numerosities in the stimulus. The resulting data points from the evaluation process are shown as squares, rhombuses, and triangles. The continuous graphs are the fitted log-normal distributions as described in the Materials and Methods section. The graphs are skewed on a linear scale (a,e) and become symmetric on a logarithmic scale (b,f). Plotting the frequencies against the ratio with the reference numerosity N_{tuned} on a logarithmic scale shows that the graphs for all reference numerosities are approximately covering each other (c,g). Using data points of all reference numerosities in the logarithm of the ratio scale to fit the continuous functions yields the overall numerosity discrimination ability (d,h) regarding the previously illustrated reference numerosities.

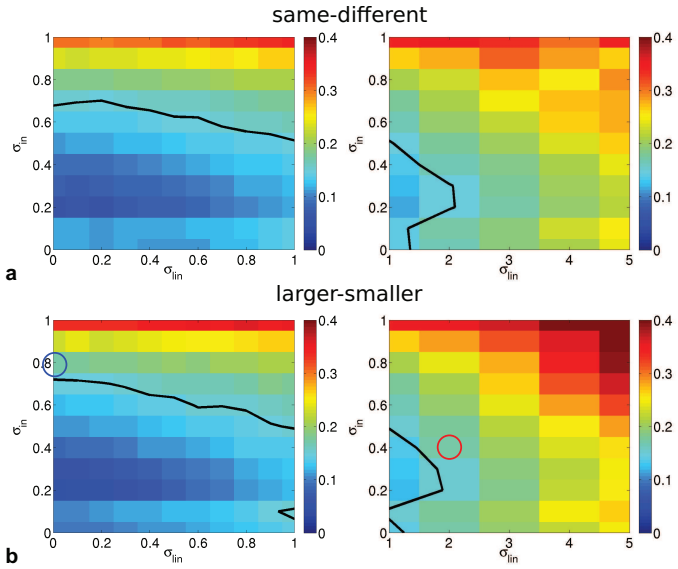


Figure 11: The distribution of Weber fractions over various noise combinations of input noise σ_{in} and linear filter output noise σ_{lin} is illustrated for (a) the same-different task and (b) the larger smaller task. The Weber fractions are a result of the evaluation of the dataset (R) with the parameters for the optimal estimators obtained from the dataset (TR). The black line shows a constant Weber fraction of $w = 0.15$ within the shown distributions. The red and the blue circle highlight the parameter instantiations used for Figure 12.

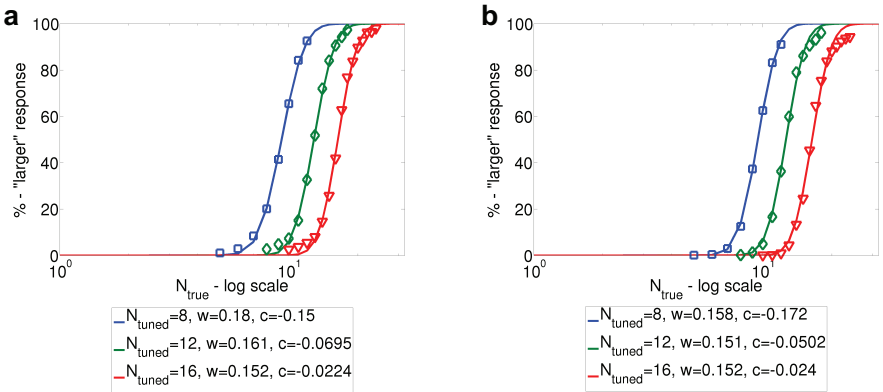


Figure 12: The response of the optimal estimator in the case of the larger-smaller task for different noise parameter combinations $(\sigma_{in}, \sigma_{lin})$ is shown on a log scale. The noise parameter combinations are (a) $(0.8, 0)$ and (b) $(0.4, 2)$ resulting in similar overall Weber fractions (a) $w = 0.172$ and (b) $w = 0.169$.

Discussion

In this paper, we proposed a general mathematical approach to the computation of numerosity. We suggested to formalize the problem and its solution by considering numerosity as a mathematical invariant. We then developed a model which takes the restrictions imposed by the structure of the visual system and the known neurophysiology into account. The proposed model was applied to visual stimuli and finally validated in computational experiments.

We first considered the problem on an abstract level with the goal to provide a formal mathematical solution without having to consider the specific constraints induced by the human visual system and the neural computations of the visual cortex. For this, we formalized it as the problem of an assignment of the property number to a configuration c of objects in three-dimensional space $\mathbb{R}^3$ (Figure 1). We then argued, that number specified in this sense has to be regarded as a mathematical *invariant* (Figure 2). This simple fact is in our view a central contribution of this paper: The understanding and modeling of numerosity should be considered as an investigation of the specific invariance properties that have to be provided by the perceptual processes and neural computations.

The special nature of this invariant - numerosity does not depend on the *shape* of the objects involved - suggested to consider the problem in the context of topology. Since the crucial property that constitutes an object is the fact that it occupies a *connected* region of space, the appropriate formal invariant turned out to be the zeroth Betti number, which counts the number of connected components of a topological space. It should be noted that this invariant represents the *ideal* solution to the problem of numerosity, and should hence be taken into account in future investigations, both in experimental and in model research. However, if we consider its computational properties, it turns out that the Betti number is not compatible with the established properties of numerosity perception: the biological numerosity system acts quite fast (Roitman et al., 2007; Park et al., 2015) and it is based on parallel computations, since there exists no hints on a dependence of processing time on the number of elements in the set (Mandler & Shebo, 1982). Connected component labeling algorithms from digital topology are able to estimate the number of components as a by-product of the labelling process but these algorithms are inherently sequentially and require multiple passes through the image, e.g. (He, Chao, & Suzuki, 2007; Suzuki, Horiba, & Sugie, 2003) .

This led us to the consideration of an alternative approach that is suited for a parallel implementation. It turned out, that such a solution can be achieved if we are willing to accept slightly more idealized conditions regarding the definition of what constitutes an object. If we consider objects as *simply connected* entities in topological terms, a solution which is fully parallel and has also a high degree of neurophysiological plausibility can be found, since the required invariant then becomes equivalent to the Euler characteristic $\chi(C)$. This in turn is computable via the famous Gauss-Bonnet theorem by measuring in a local, point-wise fashion the surface curvature of the objects in the set. By this we have specified numerosity as an invariant that can be computed in parallel by making use of local mathematical operations (derivatives).

Up to here, our analysis abstracted from the specific conditions of the visual system insofar as the objects and their surfaces are assumed to be defined in 3-D space. If we assume that the visual system is able to create some sort of 3-D representation of the environment (even a 2 1/2 D representation (Marr, 1982) may suffice), the problem of numerosity can indeed be solved by direct application of Gauss-Bonnet to this representation. However, we do not want to be logically dependent on this specific assumption, since there exist also theories

which assume that the handling of 3-D information in the early stages of the visual system (e.g. for 3-D object recognition) is based on a representation which consists of multiple views of which each is only a two-dimensional image (Wallis & Bülthoff, 1999).

We also considered alternatives of how Gauss-Bonnet can be applied directly to the processing of two-dimensional images (optical projections of the the 3-D configurations to 2-D retinal images). It turned out that this is basically possible by interpreting the 2-D luminance function as a surface (Figure 3). It can be shown that this leads to three further model variants that represent different strategies in how to deal with deviations from the ideal conditions of the theory (Kluth & Zetzsche, 2014). In this article we use the specific variant model which takes the deviations along the object borders explicitly into account by a second term (cf. equation 2). It can be assumed that this yields the most stable model version, because it combines the robust large-area integration of the surface curvature with a correction term.

Our analysis being largely based on formal mathematical considerations raises the immediate question of empirical support. Regarding the neurophysiological plausibility we have shown that the model can be easily realized by the available neural hardware. The local core operations are first- and second-order derivatives which correspond to oriented mechanisms with odd and even symmetry. It is well known that Gaussian derivatives are well suited for the description of neurons in the primary visual cortex, e.g. (Koenderink & Van Doorn, 1990; Lindeberg, 2013; Marr & Ullman, 1981; Martens, 1990; Young, 1987; Young & Lesperance, 2001). The computation of the local curvature terms requires an AND-like multiplication which can be provided by a number of neuropysiologically plausible mechanisms, cf. sect. “Neural Implementation” (Mel, 1993; Koch & Segev, 2000; Adelson & Bergen, 1985; Resnikoff & Wells Jr, 2015; Zetzsche & Barth, 1990a; Zetzsche & Nuding, 2005). Furthermore, it has been shown that certain forms of end-stopping, of the hypercomplex property, and of extra-classical receptive field properties show a close relation to the computation of curvature (Zetzsche & Barth, 1990a, 1990b; Zetzsche & Roehrbein, 2001; Zetzsche & Nuding, 2005). The computation of the curvature terms further requires a ratio operation. This is very similar to the mechanism of cortical gain control which is available at early stages of the visual cortex and is regarded as a ”‘canonical”’ neural computation of the cortex which exists in various versions (Carandini & Heeger, 2012).

Essential questions regarding the proposed model are whether it leads to clear testable predictions and how it compares to other existing models. In the following comparison we will only consider models which can be considered as ”‘image processing models”’. This criterion can range from models which process only binary images to ”‘full image processing models”’ which accept an arbitrary gray level image as input and compute the corresponding numerosity. As mentioned before, the first numerosity model of this type has to our knowledge been suggested in (Zetzsche & Barth, 1990b). This model can be regarded as a simplified version of the present one which represents shapes as polygons with +/-90 deg corners and straight segments aligned to the Cartesian grid. Aside from this it is based on the same invariance principle and will lead to similar predictions, such that we will not further considerate it as separate model in the following discussion. The classic model by (Dehaene & Changeux, 1993) can be considered as a restricted version of a full image processing model since it is only suited for one-dimensional images. Of greatest interest for the comparison are two more recent models which match the criteria. These are the ”‘contrast energy model”’ (Dakin et al., 2011; Morgan, Raphael, Tibber, & Dakin, 2014) and the neural network of (Stoianov & Zorzi, 2012). The latter one exists in two versions, as full neural network and as an abstracted mathematical model version (Stoianov & Zorzi, 2012; Cappelletti et al., 2014) which will be considered as ”‘network model”’ in the following comparison.

The model of (Dehaene & Changeux, 1993), henceforth designated as "'normalization model"', differs from the two other models and from our one not only in being only one-dimensional but also with respect to the invariance principle being used. In the normalization model the invariance is not achieved in an implicit fashion by the distributed computation of local features but with an explicit "'normalization stage"' in which input objects are mapped to a standard representation which does no longer vary in dependence on the shape properties of the input objects. This is achieved by regarding each object as "'blob"' which is represented by a dedicated "'blob detector"'. This principle can generate a certain size invariance but cannot cope with the different shape variations which become only fully apparent in the 2-D case, for example in form of elongated line-like elements. A single blob matching system will not be able to bring such different element types as lines, circles, and non-convex shapes which can additionally be arranged in quite different spatial patterns into one standardized form without interference between the invariance and the numbering properties. The existence of a blob matching stage as opposed to the spatially distributed computation of local features, as performed in the contrast energy model, the network model and our model, is a structural property which should also be testable on the neural level. While the model of (Dehaene & Changeux, 1993) predicts the existence of local units which represent a single object in an invariant fashion, the invariance in the other models is an emergent property which will only become apparent in a final spatial summation stage comprising several objects.

The comparison of the contrast energy model and the network model with our model is described on a very detailed level in appendix A4. Here we will thus only discuss the essential differences. The comparison of the models can be done with two methods, by consideration of the underlying invariance principle and by direct comparison of the detailed computations. We will start with the first approach.

On a closer look, the contrast energy model exist in two variants. In the first variant, the invariance is basically attributed to the high-frequency filtering stage and the low-frequency filters are only considered for handling a moderate bias from the size of the stimulus configuration (Dakin et al., 2011). In this interpretation, the contours of the elements are suggested as a crucial factor in the computation, since the aggregated contour length is directly proportional to the number of objects (Morgan et al., 2014). However, without an additional compensation mechanism this solution would predict a strong influence of element size since this would also directly contribute to the aggregated contour length. Deviating from the interpretation of (Dakin et al., 2011; Morgan et al., 2014) we will hence consider a second interpretation of the model, in which the low-frequency filters could contribute to size invariance by computing a cumulated area estimate. This is of special interest, since the invariance properties of the network model are also suggested to be based on a trade-off term which computes cumulated area (Stoianov & Zorzi, 2012). As shown in appendix A4 it is indeed possible to achieve size invariance by a trade-off computation between area and contour length. However, if invariance is achieved by this mechanism then the models will depend in a strong fashion on the shape of the elements. This effect is a well-known property of a famous invariant that is based on the area-length trade-off. This invariant is known as the "isoperimetric quotient" and it is a measure of the *compactness* of a shape (Kesavan, 2006). While providing perfect size invariance, it thus will show a strong dependency on the shape. The consequence of this is contour length and area are not sufficient for the computation of the full invariance properties required for the computation of numerosity.

This becomes also apparent if we use the second approach for the model comparison, the consideration of the local features which are extracted by the different models. A detailed analysis of this sort can also be found in appendix A4. It can provide information about where the two other models differ from our approach, and

this can also help in understanding how the different invariance properties are generated. The simplest case for such a comparison is the processing of binary images. In this case the interior region is completely flat. The Gauss-Bonnet theorem tells us that there should not come any contribution from these points, hence the only contributions should come from the object boundary, cf. appendix A5. In this sense it is a reasonable strategy to use only the contour region for the computation, as in the contrast energy model, or, in an implicit fashion, also in the network model (see appendix A4). However, not all contour points should contribute the same amount to the final spatially integrated numerosity variable. Rather, the contribution should depend on the *curvature* of the boundary. In particular, there should be a zero contribution from straight contour parts. That the contribution from straight contours or low-curvature contours is not appropriately reduced by the contrast energy model and the network model is, in our view, the essential reason for the systematic deviations from the ideal invariance properties that have to be expected for these models.

In conclusion, models which are based on some trade-off between area and contour will show a systematic dependency on shape, or on element size, or on both and the same is true for models which rely on contour features without providing an appropriate curvature-dependent weighting. It should be noted, however, that this does not necessarily imply that these models are not suited as models of human numerosity perception. In particular, if they do not realize the full size invariance but instead use some compromise which combines a medium size dependency with a medium shape dependency, it remains to be tested whether perceived numerosity does not exhibit the same form of deviation from perfect invariance. It is well known that the human numerosity system shows a systematic dependency on non-numerical parameters like the size of the elements (e.g., (Hurewitz, Gelman, & Schnitzer, 2006; Ross, 2003; DeWind, Adams, Platt, & Brannon, 2015)). However, it has been argued that many studies focus only on finding a statistical significant effect instead of quantifying the exact *quantitative* relation between the parameter and the numerosity bias, and that they rather interpret the influence of the parameter as an estimation error (DeWind et al., 2015). If the quantitative influence of non-numerical cues is instead explicitly modeled, the results suggest that the quantitative effect of the cues is relatively small for most individuals (DeWind et al., 2015). Nevertheless, the above considerations suggest that systematic quantitative measurements of the influence of non-numerical parameters on perceived numerosity are required in order to draw further conclusions on the different models. We would assume, that the explicit consideration of invariance mechanisms and invariance properties, as exemplified in the present analysis, could be one valuable strategy for such a systematic analysis. It is clear, however, that the prediction from our model on the outcome of such investigations would be that the quantitative deviations from invariance should always be relatively small.

Empirical tests of the invariance properties are in our view also essential with respect to the debate about a “true sense of number” as opposed to a texture-density based mechanism (Dakin et al., 2011). It is clear that we should expect to find strong interrelations between density, numerosity and cumulative area, since formally density and number are completely equivalent because they can be transformed into each other via the area. However, it would be justified in our view to regard numerosity perception as just a by-product of texture processing if it is derived from some texture related processing, like a bandpass filtering, and if the quantitative deviations from invariance are predicted by the properties of this mechanism. If on the other hand, the invariance properties should turn out to be more compatible with our sort of approach, then irrespective of whether the crucial variables in the system would covary *directly* or *inversely* with number, we would consider it adequate to speak of a sense of number.

In this context it is also of interest to consider a major argument for a "direct visual sense for number" which is given by the fact that number seems to be a primary visual property like color, orientation, or motion, to which the visual system can be adapted by prolonged viewing (Ross & Burr, 2010). If we consider the invariance mechanism of our model, the only stage where adaptation would selectively influence only the number property would be the stage of the final summation units. Adaptation at this stage would cause a systematic influence even if we later test with entirely different elements and spatial configurations as being used during adaptation. However, there should also arise systematic influences if we adapt the basic curvature features (cf. e.g. (Blakemore & Over, 1974; Bell, Gheorghiu, & Kingdom, 2009; Hancock & Peirce, 2008)). If we adapt to patterns with high local curvatures, for example, the contribution of the curvature mechanisms to the integral should be reduced, such that the resulting numerosity perceived in later tests should also be reduced.

The very nature of our approach leads to further testable predictions. The most prominent prediction, which applies to the basic model and all its variants, is based on its intimate relation to topology, and in particular to the topological concept of connectivity: On the one hand, many possible non-numeric changes of a configuration of objects are predicted to have only a small influence, even if they are quite dramatic with respect to basic signal-level properties. These are, for example, drastic changes of the sizes of the objects, or substantial alterations of their shapes (e.g. from a thin elongated to a compact round shape). On the other hand, changes of the topology, and in particular changes of connectivity, should have a strong influence, even if the corresponding signal-level changes are very small. An example for this would be the connection of two big blobs by a thin line, where the line width can be assumed to be represented on a substantially smaller spatial scale than the blob size (Figure 6). Our model would predict a clear decrease of perceived numerosity in spite of the small signal-level change. In a model using area as an essential variable, the influence should be relatively weak, whereas a mechanism relying on aggregated contour length would predict an increase of perceived numerosity. The predictions of our model are supported by findings which show that visual perception is sensitive to topological quantities (Chen, 2005), and in particular that a change of topological connectivity affects visual numerosity estimation (He, Zhang, et al., 2009; Franconeri et al., 2009).

As a last point for empirical tests it should be remembered that the restriction to simply connected objects has only been caused by plausibility arguments, since for the computation of the invariant zeroth Betti number no parallel algorithm is known. It would thus be of special interest to perform experiments with human subjects in which exactly this property is manipulated (e.g. by determining how the perceived numerosity is influenced by making "holes" into the objects). In this contex we again mention that there exists already evidence that the human perception in general is sensitive to topological quantities, and this includes a significant influence of "holes" on perception (Chen, 2005),

As our approach is derived from formal mathematical considerations it seems to generate an undesired prediction: there seems to be no obvious role for errors, and in particular not for Weber-like behavior. However, it thus can be regarded as important supporting evidence that only on the basis of quite natural assumptions about noise sources and without any explicit structural support (e.g., by logarithmic transfer functions or gain control mechanisms) the model exhibits such a Weber-like behavior. The noisy model combined with a decision making system was able to closely reproduce the number discrimination abilities of human subjects: In the literature Weber fractions of 0.174 (Piazza et al., 2004), 0.108 (Halberda & Feigenson, 2008), and 0.15 (Piazza et al., 2010) were reported for adults. The presented parametrization of the noise levels results in Weber fractions of 0.167 and 0.169 for the rectangular dataset and 0.176 and 0.197 for the circular dataset.

In conclusion, we have provided a formal analysis of the problem of numerosity. The essential and most basic result is that numerosity should be regarded as a mathematical invariant. Based on concepts from topology we have also derived a basic model structure and several specific variants of this model. Its key property is given by the Gauss-Bonnet formula which provides the desired invariance properties by the parallel integration of local curvature measures. These in turn are based on neurophysiologically highly plausible operations: directional derivatives (oriented receptive fields), nonlinear AND-like combinations related to extraclassical receptive field properties, and divisive operations similar to cortical gain control (normalization) mechanisms. The properties that turned up in our analysis are so basic from a mathematical point of view that it seems difficult to believe that there could be any mathematically reasonable model for the "sense of number" which is based on parallel computations but does do not somehow relate to the invariance principles described here. It may thus be hoped that the conceptual framework suggested here can serve as a fruitful basis for future research into the basic cognitive capacity of numerosity perception.

Appendix

A1. Optimal estimator

In order to relate the internal representation of numerosity to behavioral results we need models for decision making. Theses models must connect the analog quantity n resulting from the proposed noisy model with a decision regarding the specific task. For the same-different task and the larger task, both requiring a binary decision, we used the receiver operator characteristic (ROC) (Fawcett, 2006; Chang, 2010) from signal detection theory to obtain the optimal parameters for the estimator. Each parameter setup defines one point in the ROC space which is defined as the space spanned by the true positive rate and the false positive rate. The higher the true positive rate and simultaneously smaller the false positive rate, the better is the parametrization of the classifier. This is equivalent to the maximization of the area under the curve (AUC) defined by thy classifier in the ROC space.

In the same-different task, the classifier is defined as

$$D_{\text{different}}(n|lb,d) := \begin{cases} 1 & , \ n \in [lb, lb+d] \\ 0 & , \ \text{else,} \end{cases} \tag{8}$$

where lb defines the left bound and d the length of the detection interval.

In the smaller-larger task, the classifier is defined as

$$D_{\text{larger}}(n|t) := \begin{cases} 1 & , \ n \geq t \\ 0 & , \ \text{else,} \end{cases} \tag{9}$$

where t defines the threshold to distinguish between smaller or larger.

In both cases the optimal parameters of the estimator were determined in order to maximize the receiver operator characteristic on the dataset and one fixed reference number N_{tuned}. The parameters of the detectors can be found in Table 1. The optimal estimators on dataset TR were then used to obtain the statistics on dataset R of the behavioral tasks shown as squares, triangles and rhombuses in Figure 10.

A2. Fitting functions

We fitted the behavior of the estimators to a continuous function which is dependent on the internal weber fraction w as described in (Piazza et al., 2004). This fitting method was also used by Stoianov and Zorzi (Stoianov & Zorzi, 2012) to analyze their model such that the results are obviously comparable. However, we obtained the conditional relative frequencies of the response of the optimal estimators $D_{\text{different}}$ and D_{larger} given the true numerosities N_{true}. For the same-different task we fitted this data points to the function

$$\begin{aligned} h_{\text{different}}(N_{true}|N_{tuned},\delta,c,w) &= 1-\tfrac{1}{2}\left[erf\left(\frac{\delta+c+ln(\frac{N_{true}}{N_{tuned}})}{\sqrt{2}w}\right)\right. \\ &\quad \left. + erf\left(\frac{\delta-c-ln(\frac{N_{true}}{N_{tuned}})}{\sqrt{2}w}\right)\right], \end{aligned} \tag{10}$$

where erf is the standard error function, c controls the internal representation of the reference numerosity N_{tuned}, δ controls the variance of the corresponding probability density function and w is the internal weber fraction.

For the smaller-larger task we fitted the results of the optimal estimator to another function

$$h_{\text{larger}}(N_{true}|N_{tuned},c,w) = \frac{1}{2}\left(1+erf\left(\frac{c+ln(\frac{N_{true}}{N_{tuned}})}{\sqrt{2}w}\right)\right), \tag{11}$$

with the functions and parameters as described above. In order to obtain representative fits for the number discrimination ability (Piazza et al., 2004), we only used numerosities which are close to the reference numerosity for the fit, i.e. we used all N with the ratio $\frac{N}{N_{tuned}} \in [0.625, 1.5]$. In both cases the parameters were estimated by a standard least squares minimization of the residual. The fitted curves obtained from dataset R are illustrated as continuous lines in Figure 10. The respective parameters can be found in Table 2.

A3. SNR and PSNR analysis

In order to evaluate the noise behavior of the input unit and the linear filter units with respect to their neural plausibility the peak-signal-to-noise-ratio ($PSNR$) is derived from theory as described in the following. The input signal is a function $l : [-1,1]\times[-1,1] \to [0,1]$ with amplitude $A(l) = 1$ such that the $PSNR$ becomes

$$PSNR(l) = 10\log_{10}\left(\frac{1}{\sigma_{in}^2}\right) \text{ dB}. \tag{12}$$

The implementation of the derivatives strongly depends on the discretization and thus on the step sizes Δx and Δy. We thus can derive from the differential quotient

$$|l_x(x,y)| \approx \frac{l(x+\Delta x,y)-l(x,y)}{\Delta x} \leq \frac{1}{\Delta x}. \tag{13}$$

The upper bound for l_y can be derived analogously. The second derivative in x-direction is bounded by

$$|l_{xx}(x,y)| \approx \frac{l(x+\Delta x,y)-l(x,y)+l(x-\Delta x,y)-l(x,y)}{\Delta x^2} \leq \frac{2}{\Delta x^2}. \tag{14}$$

Table 3: **PSNR and SNR.** The $PSNR$ and SNR functions and values for the two stages *input* and *linear filters* are shown for the theoretical setup and the three datasets (TR), (R), and (C).

Signal	**PSNR** (in 10 dB) formula	**PSNR** (in 10 dB) $\Delta x = \Delta y = 1/50$	**SNR** (in 10 dB) TR	R	C
l / Mean	$\log_{10}\left(\frac{1}{\sigma_{in}^2}\right)$	$-\log_{10}(\sigma_{in}^2)$	$-0.7948 - \log_{10}(\sigma_{in}^2)$	$-0.7949 - \log_{10}(\sigma_{in}^2)$	$-0.7919 - \log_{10}(\sigma_{in}^2)$
l_x	$\log_{10}\left(\frac{4}{\Delta x^2 \sigma_{lin}^2}\right)$	$4.0000 - \log_{10}(\sigma_{lin}^2)$	$0.8431 - \log_{10}(\sigma_{lin}^2)$	$0.8431 - \log_{10}(\sigma_{lin}^2)$	$0.8372 - \log_{10}(\sigma_{lin}^2)$
l_y	$\log_{10}\left(\frac{4}{\Delta y^2 \sigma_{lin}^2}\right)$	$4.0000 - \log_{10}(\sigma_{lin}^2)$	$0.8429 - \log_{10}(\sigma_{lin}^2)$	$0.8432 - \log_{10}(\sigma_{lin}^2)$	$0.8375 - \log_{10}(\sigma_{lin}^2)$
l_{xx}	$\log_{10}\left(\frac{16}{\Delta x^4 \sigma_{lin}^2}\right)$	$8.0000 - \log_{10}(\sigma_{lin}^2)$	$3.7885 - \log_{10}(\sigma_{lin}^2)$	$3.7896 - \log_{10}(\sigma_{lin}^2)$	$3.7207 - \log_{10}(\sigma_{lin}^2)$
l_{yy}	$\log_{10}\left(\frac{16}{\Delta y^4 \sigma_{lin}^2}\right)$	$8.0000 - \log_{10}(\sigma_{lin}^2)$	$3.7861 - \log_{10}(\sigma_{lin}^2)$	$3.7879 - \log_{10}(\sigma_{lin}^2)$	$3.7233 - \log_{10}(\sigma_{lin}^2)$
l_{xy}	$\log_{10}\left(\frac{16}{\Delta x^2 \Delta y^2 \sigma_{lin}^2}\right)$	$8.0000 - \log_{10}(\sigma_{lin}^2)$	$2.8204 - \log_{10}(\sigma_{lin}^2)$	$2.8207 - \log_{10}(\sigma_{lin}^2)$	$2.8766 - \log_{10}(\sigma_{lin}^2)$
Mean	-	$6.4000 - \log_{10}(\sigma_{lin}^2)$	$2.4162 - \log_{10}(\sigma_{lin}^2)$	$2.4168 - \log_{10}(\sigma_{lin}^2)$	$2.3991 - \log_{10}(\sigma_{lin}^2)$

The upper bound for l_{yy} and l_{xy} can be derived analogously. The linear filter units are computed by a convolution operation with a Gaussian function g ($\|g\|_{L^1} = 1$) such that we need an upper bound for the filter output. By using Young's inequality we get for l_x

$$\|l * g_x\|_{L^\infty} = \|l_x * g\|_{L^\infty} \leq \|l_x\|_{L^\infty} \underbrace{\|g\|_{L^1}}_{=1} \leq \frac{1}{\Delta x}. \tag{15}$$

The peak-to-peak signal amplitude is then twice the maximum norm of the respective filter unit, i.e $A(l_x) = \frac{2}{\Delta x}$. The $PSNR$ then becomes

$$PSNR(l_x) = 10 \log_{10}\left(\frac{A(l_x)^2}{\sigma_{lin}^2}\right) \text{ dB} = 10 \log_{10}\left(\frac{4}{\Delta x^2 \sigma_{lin}^2}\right) \text{ dB}. \tag{16}$$

The $PSNR$ formulas for all signals are summarized in Table 3.

In contrast to the $PSNR$ which can be derived from theory easily, we also consider the SNR of the datasets used for the analysis with respect to the Weber fraction. The SNR is determined by the expected integral of the squared signal over each dataset. The SNR for the function l and all images in the dataset R is given by

$$SNR(l) = 10 \log_{10}\left(\frac{E_{l \in R}(\frac{1}{4} \int_{[-1,1]\times[-1,1]} l(x)^2 \, dx)}{\sigma_{in}^2}\right) \text{ dB}. \tag{17}$$

The values for the different signal types and the different datasets can also be found in Table 3.

In order to relate the given SNR values with information encoded by neurons we determine the rate of an uniform quantizer for the given SNR values. In good approximation we can use the "6-dB-per-bit-rule" (Gray & Neuhoff, 1998) such that the rate R is given approximately by

$$R(l) \approx \frac{SNR(l)}{6.02} \text{ bit} \tag{18}$$

for a signal l.

A4. Model comparison

On the general level, our proposed model differs from the other image-based models ((Dehaene & Changeux, 1993; Dakin et al., 2011; Morgan et al., 2014; Stoianov & Zorzi, 2012; Cappelletti et al., 2014)) by the moti-

Table 4: **Model comparison by operations.** The different models (left to right) are split into their basic operations (top to bottom). The operations are split into layers with a linear operation followed by a nonlinear one (NL).

Model	contrast energy (Dakin et al., 2011)	network (Stoianov & Zorzi, 2012)	proposed
1st layer	Isotropic derivatives of 2D-Gaussian filters	2D-Gaussian filter & spatial integration of cumulative area	Orientation-selective derivatives of 2D-Gaussian filters
NL	Absolute value	Sigmoid function & log-norm	Multiplicative feature combination
2nd layer	Spatial integration	2D-Gaussian filter minus constant (nonlinear function of cumulative area)	Additive combination of different features
NL	Ratio of different filter bands (density)	-	Ratio of different combinations (curvature quantities)
Top	Multiplication of density and cumulative area estimate from low frequency filter	Linear weighting	Spatial integration

vation behind its design. The central motivation for our model is the idea to find a sound mathematical basis for the provision of the invariance properties required by an ideal numerosity mechanism. As a second step, we then considered how these ideal principles could be implemented and approximated by neurobiologically plausible mechanisms. The other models have different goals or they are based on a different type of reasoning. The model of Dehaene and Changeux (Dehaene & Changeux, 1993) was an early model that tried to model the image processing aspect to a certain degree. However, it is only a one-dimensional model and, as such, cannot deal with two-dimensional shapes and the associated invariance properties. We therefore will not include it in the following detailed comparison. The contrast energy model (Dakin et al., 2011; Morgan et al., 2014) is motivated by empirical observations in psychophysical experiments which suggested the possibility of a close connection between numerosity and density (Durgin, 2008). The actual model has then be derived from considerations of how a density mechanism can be provided by use of established filter-based texture computations (Dakin et al., 2011; Morgan et al., 2014). The model of Stoianov and Zorzi (Stoianov & Zorzi, 2012) is based on a deep learning neural network. It is then abstracted by a computational analysis to a spatial filter model with point nonlinearities (Stoianov & Zorzi, 2012; Cappelletti et al., 2014). (This model is henceforth designated as network model.)

In order to compare the other models with our model there exist in principle two different strategies. One is to find out the invariance properties which underly the other models, and to compare them to the invariance properties of our model. The other strategy is to map the model to a common framework and to compare the model components within this framework. We will make use of both strategies in our analysis.

We start by considering in how far the other models make use of alternative invariance mechanisms. Two basic shape properties seem somehow to be involved in the computations of the two models, one is the boundary of objects and the other one is their area. For the contrast energy model, the contour length of the elements is explicitly mentioned as one crucial factor in its computations (Morgan et al., 2014). In the network model, the cumulated area of all elements is seen to play a crucial role as a covarying factor (Stoianov & Zorzi, 2012).

How can contour length and area be used to obtain invariant properties? It is a well known fact that the ratio

$$q = \frac{A}{L^2} \tag{19}$$

with A being the area of an object and L its contour length, is an invariant with respect to size changes. We can thus use this invariance property of the ratio to compute the number of objects based on the aggregated area $\sum A_i$ and the aggregated contour length $\sum L_i$ as

$$n = q\frac{(\sum L_i)^2}{\sum A_i} = q\frac{n^2L^2}{nA} = qn\frac{1}{q}, \tag{20}$$

where all objects have the same area $A_i = A$. However, although it is size-invariant, the ratio q depends in general in a systematic fashion on an important second factor, the shape of the elements. It has long been known that the normalized variant of q, the so called "isoperimetric quotient" ($Q = \frac{4\pi A}{L^2}$) is a measure of the *compactness* of a shape (Kesavan, 2006). This measure attains its maximum of 1 in case of circles but can easily get much smaller if the object becomes more "'ragged"'. For certain patterns, like the well-known Koch snowflake, Q can even approach zero. But also for commonplace patterns, U-like shapes for example, Q is around 1/3, and hence much smaller than 1. This would imply that three U-shaped elements should appear perceptually as numerous as 9 circles. It is not directly evident that the two models are aimed at this ratio invariance, since only one of its parts is emphasized in each description of the models. However, in the contrast energy model the low-frequency filtering could be seen to bear a certain resemblance to area-related computations. And in the network model, a contour-related feature can be seen to be implicitly computed, as will be explained in detail in the common framework analysis presented below. However, in so far as the models deviate form the ratio invariant, they will loose the size invariance, and if they manage to come close to the ratio invariant, they will substantially depend on the shape of the elements. If we want to allow for arbitrary object shapes, making use of the ratio invariant seems not to be a viable solution.

Now let us pursue the second strategy for the model comparison, the common framework approach. In Table 4 the three models are split into their operations at different stages/layers. From this representation it becomes apparent that all three models use similar basic computations. They have a similar first layer consisting of linear filter operations (derivatives and lowpass filtering), and of the spatial pooling of local luminance signals into an aggregated cumulative area. The contrast energy model and the network model then use a standard nonlinear transfer function (a sigmoid function and an absolute value function). In contrast, our proposed model requires a nonlinear AND-like (multiplicative) interaction. (As described in the main text, this could be realized in neural hardware by a variety of mechanisms.)

In the second layer, the contrast energy model and the network model perform a spatial integration (the first a global and the second a local integration). Our model also performs such a spatial integration but only as the last processing step on its top layer. On the second layer, our model performs a linear combination of different local features (different terms of the curvature computation). This combination of different local features is one distinction to the other models which both use only one type of local spatial feature (a Laplacian bandpass feature or a Gaussian lowpass feature).

The contrast energy model has a ratio operation between the two aggregated contrast energy values as its last operation. However, this operation is considered only for the incorporation of a moderate bias term for the

influence of patch size (Dakin et al., 2011), whereas the basic numerosity variable is argued to be provided by the high-frequency filter output (Dakin et al., 2011; Morgan et al., 2014). The network model ends with a linear weighting over the corrected Gaussian filter outputs of the second layer by a linear classifier.

In the following, we will try to compare the models with respect to their general behavior. This will require to make some simplifying approximations but we think that the basic trend will be similar to the behavior of the complete models. For the contrast energy model, the authors suggest that the core computation is the spatial pooling of the local energy from a Laplacian high-frequency filter (Dakin et al., 2011; Morgan et al., 2014). The main argument is to use this to measure the amount of contour, since adding more objects to an image amounts to adding more contour (Morgan et al., 2014). For a given type of element, e.g. for squares of some fixed size, the estimate works perfect: Changing the numbers of elements will proportionally change the pooled filter output. But how does this operation behave with respect to the invariance properties, i.e. if we change the shape of the elements? For some changes the influence will be relatively small, in particular if we only use concave elements, as often done in numerosity experiments. For example, if we replace the squares by disks of the same area the difference in contour length will be only about 10%. If we use also concave elements, differences can become larger. For the L-shaped element shown in Figure 14a6 the difference will become as large as 113% in terms of number units (Figure 13). The differences will become even larger if we also consider elements with different size. For example, if we compare small squares of side length d/4 with large squares of side length d, a set of N large squares will appear to have the same numerosity as a set of 4N small squares. Thus the contrast energy model makes clear predictions about systematic deviations from the ideal invariance properties. In how far these are also present in human bias terms remains to be determined (see discussion).

The analysis of the network model is somewhat more complicated. We make the following simplifying assumptions, which we think are valid as far as systematic deviations from invariance are considered: First, the 2nd layer operation and the final weighting by the classifier are both linear. The weighting of the 2nd layer units at the different positions by the classifier can be expected to be similar, since these units have all the basically same status with respect to the computation of numerosity (any systematic differences between units at different positions would induce position-dependent biases). We thus assume that these two steps can be combined into one global spatial integration $\sum_{(m,n)\in X} O^{mn}$. Here the O^{ij} are the local Gaussian filter outputs after the sigmoid point nonlinearity in the first layer and X is the set of indices of the discretized domain of the input image. We further assume that the aggregated luminance can also be rewritten into such a spatial integration as $c = \sum_{(m,n)\in X} kI^{mn}$. For this, we linearize the logarithm (the argument can assume values only between 1 and 2) as $c = log(1 + \frac{\sum_X I}{c_{max}}) \approx \sum_{(m,n)\in X} kI^{mn}$. The final sum can then be written as $\sum_{(m,n)\in X}(O^{mn} - kI^{mn})$, i.e., it is a sum over a difference image between the original image and a nonlinear low-pass filtered version of it. This analysis suggests that the constant k should be an absolute constant, which does not depend on further parameters, and that it should be chosen in a way to avoid contributions from the interior object area. This is necessary because otherwise there would remain a systematic dependency of the numerosity estimate on the object area. Formally, k depends on the normalization constant c_{max}. This is an absolute constant within an experiment, but it is not quite clear how a subject can have knowledge of it or what this implies for the relation between different types of experiments. In the following, we consider k as an absolute constant which is chosen to produce a zero difference throughout all areas of constant luminance of the input image. We can then analyze the contributions from the remaining non-zero areas of the difference image. For a class of simple examples, patterns with straight edges and right angled corners only (cf. Figure 14a5-10), we know the correct solution

from the Gauss-Bonnet theorem and from the Euler formula: The desired invariant can then be computed by simply summing up the number of (signed) corners as $n = 1/4 \sum v_i$ where each corner i contributes +1 or -1. For this special case the difference image should thus ideally only generate contributions from the corners and no contributions from the straight borders. This is also evident by considering the case of enlarging such a figure, since then the number of corners remains constant but the length of the straight contours is increased. Since the lowpass filtering smooths the corners, the network model will indeed provide the desired contribution from the corners. However, there seems also to be a non-vanishing contribution from the straight borders, such that there will be a dependency on the size as well as on the shape of the elements. While the shape dependency is moderate for convex shapes it becomes larger for concave shapes, since for those the total contour length is significantly larger (see Figures 14a6 and 13).

In conclusion, both the contrast energy model and the network model can be seen to make use of similar basic local features which result from some version of nonlinear bandpass filtering. This operation produces the basic local features (explicitly represented by the rectified bandpass features in case of the contrast energy model and implicitly represented by the subtraction of the cumulative area from the lowpass filtered features in the network model). These local features are then spatially pooled (globally in one step in the contrast energy model and in two steps, first the second Gaussian lowpass and then the linear classifier, in the network model). Mapped to this type of architecture, the two models can be directly compared to our model, which can also be seen as consisting of the computation of nonlinear local features and a subsequent spatial pooling. From the mathematical basis of our model we know which local parts of the input image have to play which role, if we want to achieve the desired invariance. We know, for example, that all constant image areas should not generate a systematic contribution. This enforces the mutual cancellation of the lowpass responses and the per-sample contribution of the cumulative area measurement in the interior object areas in the network model. If this is violated, a systematic influence of the cumulative area is unavoidable. We also know, that the contributions from the contour regions (in case of binary images) should systematically vary with the contour curvature. In particular, straight contours should not generate any contribution, since otherwise there will result a systematic dependency on the total contour length, and on the size of the elements. This is a problem for both the contrast energy model and the network model, since both produce nonvanishing contributions from straight contours. It should further be noted that it is generally not possible to trade off the false contributions from one class (say area) against the false contributions of the other class (contour), since there exists no shape-independent relation between the two (see the arguments above regarding the isoperimetric quotient).

A5. Invariance properties of binary objects

Assuming the special case of binary images/objects allows the derivation of problem-specific computational principles. Here we describe the general rules for the invariant computation of numerosity for this particular signal class. The binary setup can be interpreted from different point of views. In the following we consider two possible interpretations.

a) We can see it as an inherently one-dimensional problem. The objects are bounded two-dimensional subsets of the x-y-plane such that they can be represented by their one-dimensional boundary curve, cf. Figure 15. In this case the one-dimensional counterpart of the Gauss-Bonnet theorem is the following standard corollary in differential geometry.

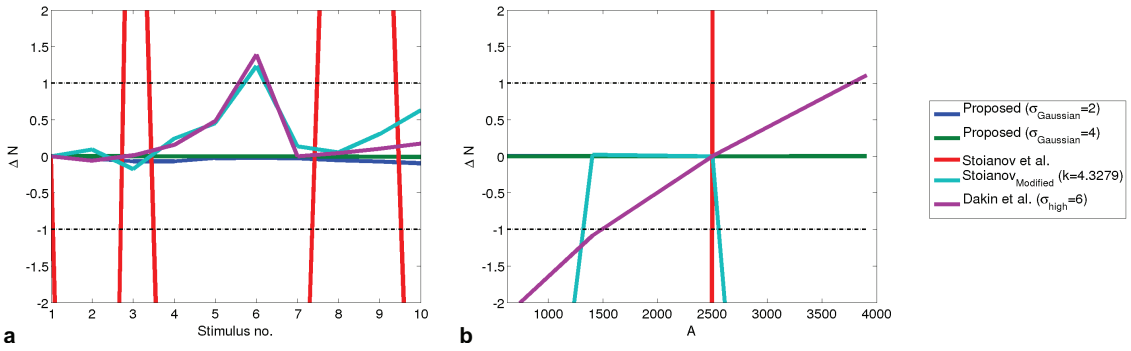


Figure 13: Responses of different models to stimuli illustrated in Figure 14 are shown. (a) shows the relative responses (difference to reference stimulus no. 1) to the convex and concave stimuli from Figure 14a. Analogously (b) shows the relative responses (difference to reference stimulus A=2500) to stimuli with different cumulative area A, cf. Figure 14b. The dataset in Figure 14c was used to determine the mean difference in model response per numerosity. This value was used to determine ΔN. The dotted black lines mark the absolute difference 1 in numerosity with respect to the reference stimulus.

Corollary 2 *Let Γ be a closed, regular, plane curve. Then the quantity*

$$\int_{\Gamma} \kappa \, ds = 2\pi n, \tag{21}$$

where κ is the curvature of the plane curve and n is an integer called the rotation index of the curve.

As the rotation index is always one for simply connected objects and the integral operator is linear, the integral over the disjunction of multiple boundary curves results in their number. The simple one-dimensional variant of the Gauss-Bonnet theorem thus tells us that we have to integrate the curvature of the boundary curve(s). In contrast to the isoperimetric quotient the integral over the curvature is independent of the shape. In Figure 15a this quotient decreases from left to right whereas the integral over the curvature remains constant. This is due to the fact that the increase of positive curvature from left to right in Figure 15b is compensated by an additional contribution of negative curvature.

If we assume a piecewise constant boundary, the integral over the curvature becomes the discrete sum over the arcs. In particular this setup includes the special case of objects having right-angled corners and otherwise vertical or horizontal straight boundary segments, see Figure 14a5-10 for examples. The corners can be divided into two classes. We have external corners where the object area covers one quarter within a circular neighborhood. These corner have an arc of $\pi/2$. The corners where the object area covers three quarter build the second class, the internal corners with a contribution of $-\pi/2$. If we divide both sides of the corollary by 2π, we obtain a standard solution to determine the number of simply connected binary objects in computer vision (Umbaugh, 2005). The number of objects then becomes the number of external rectangular corners minus the number of internal rectangular corners divided by four.

b) We can consider the objects as two-dimensional surfaces such that the approximate invariance requires (i) no contribution from the interior area of the object, (ii) no contribution from straight edges, and (iii) a contribution from the remaining regions. The following approach which matches all criteria in the binary image case is motivated by the detection of the contour line of the region with the aid of the Laplace operator. If the contribution is zero on constant luminance regions and the computed quantity at the boundary allows a mapping

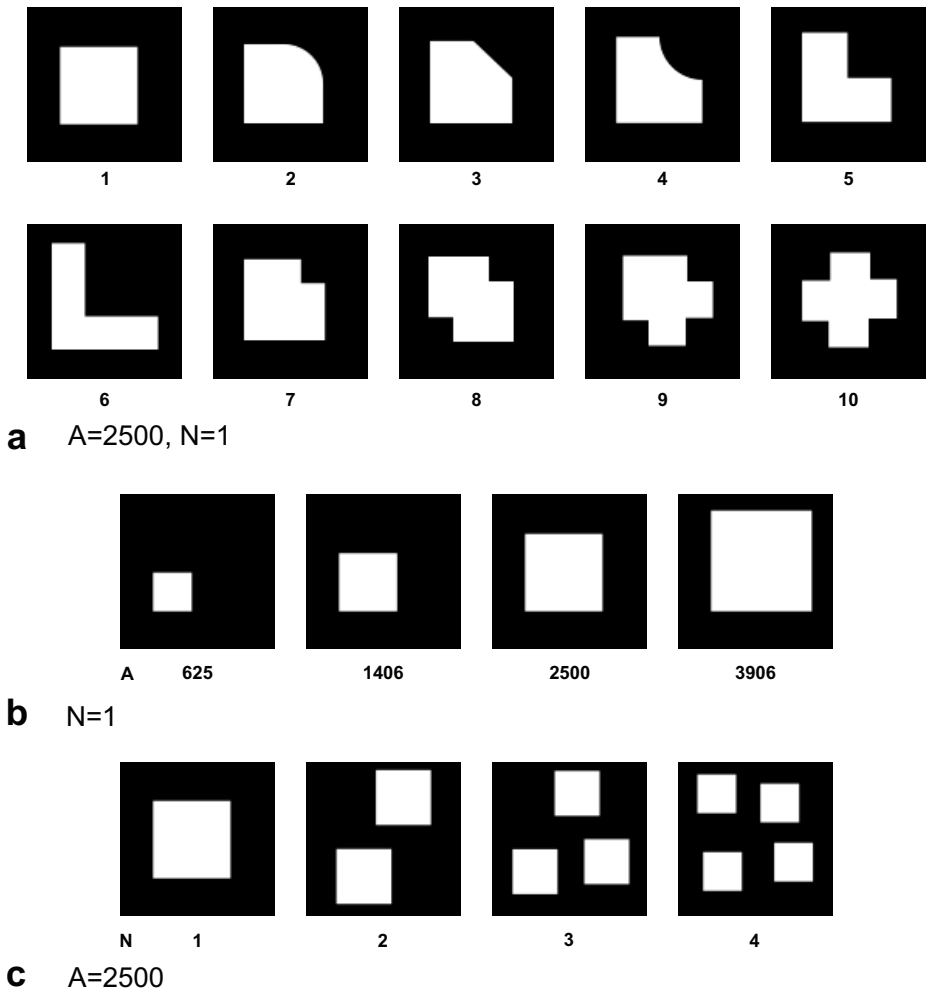


Figure 14: Three datasets of various stimuli are illustrated. (a) shows convex and concave objects with constant cumulative area A and numerosity N. (b) shows one rectangular object with increasing cumulative area A. (c) shows multiple rectangular objects with constant cumulative area A. All stimuli are binary images of size 100×100 pixels. Cumulative area A in pixels.

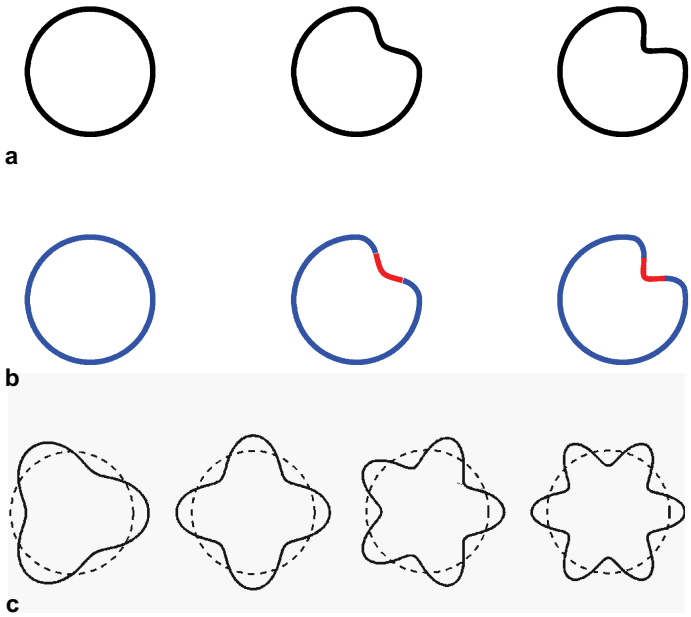


Figure 15: Three objects are illustrated by their one dimensional boundary curves. The ratio of squared perimeter and area increases from left to right in (a). The positive curvature (blue) and the negative curvature (red) of the boundary curve are illustrated in (b). In (c) the area is constant for all objects and the frequency of the sinusoidal grating along the boundary of a circle increases from left to right. The higher the frequency the larger is the contour length resulting in an decreasing isoperimetric quotient from left to right.

to the curvature of the boundary curve, the integration over the whole domain should result in an estimate for numerosity. An easy way to detect the contour is to determine the zero crossing line of the Laplacian filtered image. In order to guarantee the differentiability for the Laplace operator the binary image must be filtered by an appropriate filter function. One standard filter function is the Gaussian function. This function is not optimal for further considerations as its support is infinite. We assume a filter function g with compact circular support Ω around the origin. Furthermore the Laplace of the function g is positive in an inner disk $\Omega_{in} \subset \Omega$ and negative in $\Omega_{out} = \Omega \setminus \Omega_{in}$. The integral of Δg over Ω is assumed to be zero. Thus, the Laplace of the convolution with the filter function yields zero for constant luminance regions. Furthermore, if we consider a straight edge with a length greater or equal the diameter of Ω, the line integral in the orthogonal (to the edge) direction of the filter output becomes zero for all boundary points where the neighborhood Ω contains the straight edge only. Thus, the integral over the whole domain of the Laplacian filter output has already two of the three desired properties. It causes no contributions at constant luminance regions (i) and no contributions at straight edges (ii). That the spatial integration of the filter output does not result in the desired estimate for numerosity can be seen easily by the following equation

$$\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} l(y)\Delta g(x-y)\,dy\,dx = \int_{\mathbb{R}^2} l(y)\underbrace{\int_{\mathbb{R}^2} \Delta g(x-y)\,dx\,dy}_{=0,\ \forall y\in\mathbb{R}^2} = 0. \tag{22}$$

Note that the following solution does not depend on the Laplace operator. We can replace Δg by an arbitrary filter function h which has the same characteristic behavior on $\Omega = \Omega_{in} \cup \Omega_{out}$. In order to obtain contributions

from the curved regions a nonlinear function $F : \mathbb{R} \to \mathbb{R}$ is introduced. This function must be odd-symmetric such that the zero contribution on constant regions and on straight edges is still preserved. The problem of the estimation of numerosity n then becomes the problem of finding an appropriate combination of a nonlinear function F and a filter function h such that

$$\int_{\mathbb{R}^2} F\left(\int_{\mathbb{R}^2} l(y)h(x-y)\,dy\right)\,dx \sim n. \tag{23}$$

Again, the success of this approach is due to the accuracy in approximating the curvature of the boundary curve. We assume that the boundary curve is a piecewise constant line with a minimum distance of ϵ between vertices and that the diameter of the region Ω is chosen smaller than ϵ (i.e. the radius $\delta < \epsilon/2$). In this particular case we know that integrating the linear filter output over the neighborhood with radius δ around a vertex results in zero. At an edge the integration domains of positive and negative contributions, which sum up to zero, equal in size. At a vertex the size of these integration domains does not equal anymore but the overall integral remains zero. Thus a monotonic nonlinear function F fulfilling the previously formulated constraints is sufficient to guarantee a contribution from the vertices. If this output is proportional to the arc, the integration over the whole domain results in an estimate for numerosity. Figure 16 illustrates one choice of h and F which is sufficient to extract the desired information. We also applied the filter function from Figure 16a and the modified sigmoid function to the stimuli in Figure 14c producing an output of 133.73, 261.58, 399.45, and 533.46 from left to right. Relative to the output of the first stimulus with one object, the responses become 1.00, 1.95, 2.98, and 3.98. This simple choice of h and F thus results in the desired proportionality to numerosity.

In conclusion for binary images, approximating the curvature of the boundary curve and integrating this quantity results in an estimate for numerosity. Furthermore, the fundamental principle which allows the estimation of numerosity from binary images is the invariance property provided by the Gauss-Bonnet theorem.

References

Abbott, L., Varela, J., Sen, K., & Nelson, S. (1997). Synaptic depression and cortical gain control. *Science*, *275*(5297), 221–224.

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *JOSA A*, *2*(2), 284–299.

Allik, J., Tuulmets, T., & Vos, P. G. (1991). Size invariance in visual number discrimination. *Psychological Research*, *53*(4), 290–295. , doi:10.1007/BF00920482

Anobile, G., Cicchini, G. M., & Burr, D. C. (2014). Separate mechanisms for perception of numerosity and density. *Psychological Science*, *25*(1), 265-270. , doi:10.1177/0956797613501520

Barth, E., Ferraro, M., & Zetzsche, C. (2001). Global topological properties of images derived from local curvature features. In *Visual form 2001* (pp. 285–294). Springer.

Bell, J., Gheorghiu, E., & Kingdom, F. (2009). Orientation tuning of curvature adaptation reveals both curvature-polarity-selective and non-selective mechanisms. *Journal of Vision*, *9*(12), 3.

Blakemore, C., & Over, R. (1974). Curvature detectors in human vision. *Perception*, *3*(1), 3–7.

Bonn, C. D., & Cantlon, J. F. (2012). The origins and structure of quantitative concepts. *Cognitive Neuropsychology*, *29*(1-2), 149–173. , doi:10.1080/02643294.2012.707122

Brannon, E. M. (2006). The representation of numerical magnitude. *Current Opinion in Neurobiology*, *16*(2),

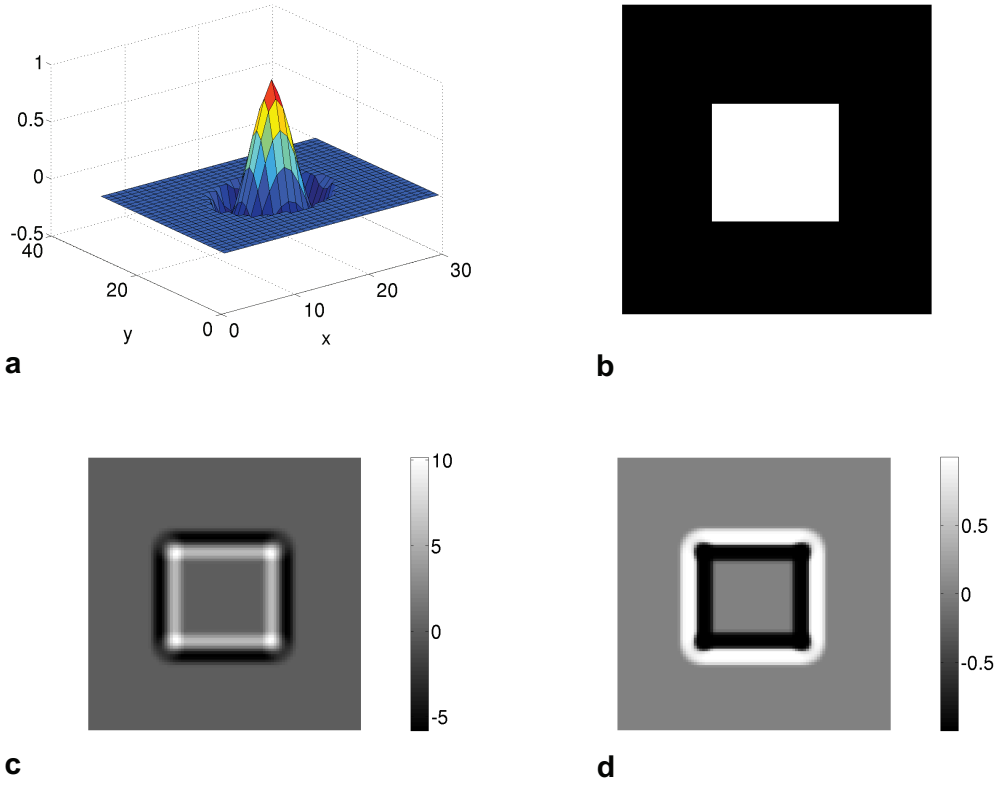


Figure 16: In (a) the used filter function h is illustrated. The radius of the inner disk Ω_{in} is 4 and the radius of the whole domain Ω is 7.24. The filter function was determined by a cubic spline interpolation such that its integral over the whole domain is zero. In (b) an example binary stimulus is illustrated. (c) shows the linear filter output $l * h$. The integral over the whole domain in (c) is approximately -0.015. In (d) a modified sigmoid function $F(x) = -2(\frac{1}{1-e^{-x}} - \frac{1}{2})$ is applied to the filter output. The area effect of the positive and negative contributions at the corners becomes apparent. This results in an integral over the whole domain in (d) of approximately 131.273.

222–229. , doi:10.1016/j.conb.2006.03.002

Buzsáki, G., & Mizuseki, K. (2014). The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews Neuroscience*, *15*(4), 264–278. , doi:10.1038/nrn3687

Cappelletti, M., Didino, D., Stoianov, I., & Zorzi, M. (2014). Number skills are maintained in healthy ageing. *Cognitive psychology*, *69*, 25–45.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51-62. , doi:10.1038/nrn3136

Chang, C.-I. (2010). Multiparameter receiver operating characteristic analysis for signal detection and classification. *Sensors Journal, IEEE*, *10*(3), 423–442. , doi:10.1109/JSEN.2009.2038120

Chen, L. (2005). The topological approach to perceptual organization. *Visual Cognition*, *12*(4), 553–637. , doi:10.1080/13506280444000256

Dakin, S. C., Tibber, M. S., Greenwood, J. A., Kingdom, F. A. A., & Morgan, M. J. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences*, *108*(49), 19552-19557. , doi:10.1073/pnas.1113195108

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems* (Vol. 806). Cambridge, MA: MIT Press.

Dehaene, S., & Changeux, J. P. (1993). Development of elementary numerical abilities: a neuronal model. *Journal of Cognitive Neuroscience*, *5*(4), 390–407. , doi:10.1162/jocn.1993.5.4.390

DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (2015). Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition*, *142*, 247–265. , doi:10.1016/j.cognition.2015.05.016

Durgin, F. H. (2008). Texture density adaptation and visual number revisited. *Current Biology*, *180*(18), R855 - R856. , doi:10.1016/j.cub.2008.07.053

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874. , doi:10.1016/j.patrec.2005.10.010

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. , doi:10.1016/j.tics.2004.05.002

Franconeri, S., Bemis, D., & Alvarez, G. (2009). Number estimation relies on a set of segmented objects. *Cognition*, *113*(1), 1–13.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1–2), 43–74. , doi:10.1016/0010-0277(92)90050-R

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences*, *4*(2), 59–65. , doi:10.1016/S1364-6613(99)01424-2

Geisler, W. S., & Albrecht, D. G. (1992). Cortical neurons: isolation of contrast gain control. *Vision research*, *32*(8), 1409–1410.

Gray, R. M., & Neuhoff, D. L. (1998). Quantization. *Information Theory, IEEE Transactions on*, *44*(6), 2325–2383.

Gross, H. J., Pahl, M., Si, A., Zhu, H., Tautz, J., & Zhang, S. (2009). Number-based visual generalisation in the honeybee. *PloS one*, *4*(1), e4263. , doi:10.1371/journal.pone.0004263

Grossberg, S. (1988). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural networks*, *1*(1), 17–61. , doi:10.1016/0893-6080(88)90021-4

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457–1465. , doi:10.1037/a0012682

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668. , doi:10.1038/nature07246

Hancock, S., & Peirce, J. W. (2008). Selective mechanisms for simple contours revealed by compound adaptation. *Journal of Vision*, *8*(7), 11.

He, L., Chao, Y., & Suzuki, K. (2007). A linear-time two-scan labeling algorithm. In *Image processing, 2007. icip 2007. ieee international conference on* (Vol. 5, pp. V-241–V-244).

He, L., Chao, Y., Suzuki, K., & Wu, K. (2009). Fast connected-component labeling. *Pattern Recognition*, *42*(9), 1977 - 1987. , doi:10.1016/j.patcog.2008.10.013

He, L., Zhang, J., Zhou, T., & Chen, L. (2009). Connectedness affects dot numerosity judgment: Implications for configural processing. *Psychonomic Bulletin & Review*, *16*(3), 509–517. , doi:10.3758/PBR.16.3.509

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual neuroscience*, *9*(02), 181–197. , doi:10.1017/S0952523800009640

Hegde, J., & Felleman, D. (2007). Reappraising the Functional Implications of the Primate Visual Anatomical Hierarchy. *The Neuroscientist*, *13*(5), 416–421. , doi:10.1177/1073858407305201

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366. , doi:10.1016/0893-6080(89)90020-8

Hurewitz, F., Gelman, R., & Schnitzer, B. (2006). Sometimes area counts more than number. *Proceedings of the National Academy of Sciences*, *103*(51), 19599–19604. , doi:10.1073/pnas.0609485103

Kaski, S., & Kohonen, T. (1994). Winner-take-all networks for physiological models of competitive learning. *Neural Networks*, *7*(6-7), 973–984. , doi:10.1016/S0893-6080(05)80154-6

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*(4), 498–525.

Kesavan, S. (2006). *Symmetrization & applications* (Vol. 3). Singapore: World Scientific.

Kluth, T., & Zetzsche, C. (2014). Spatial numerosity: A computational model based on a topological invariant. In C. Freksa, B. Nebel, M. Hegarty, & T. Barkowsky (Eds.), *Spatial cognition ix* (Vol. 8684, p. 237-252). Heidelberg: Springer International Publishing.

Kluth, T., & Zetzsche, C. (2015). *A topological view on numerosity.* (In preparation.)

Koch, C., & Segev, I. (2000). The role of single neurons in information processing. *Nature Neuroscience*, *3*, 1171–1177.

Koenderink, J. J., & Doorn, A. J. van. (2003). Shape and shading. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 1090–1105). Cambridge: MIT Press.

Koenderink, J. J., & Van Doorn, A. (1990). Receptive field families. *Biological cybernetics*, *63*(4), 291–297. , doi:10.1007/BF00203452

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, *16*(1), 37–68.

Lampl, I., Ferster, D., Poggio, T., & Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. *Journal of neurophysiology*, *92*(5), 2704–2713.

Lindeberg, T. (2013). A computational theory of visual receptive fields. *Biological cybernetics*, *107*(6), 589–635. , doi:10.1007/s00422-013-0569-z

Mandler, G., & Shebo, B. J. (1982). Subitizing: an analysis of its component processes. *Journal of Experimental Psychology: General*, *111*(1), 1–22. , doi:10.1037/0096-3445.111.1.1

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman and Company.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London B: Biological Sciences*, *207*(1167), 187–217.

Marr, D., & Ullman, S. (1981). Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London B: Biological Sciences*, *211*(1183), 151–180.

Martens, J.-B. (1990). The hermite transform-theory. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *38*(9), 1595–1606. , doi:10.1109/29.60086

Mead, C., & Ismail, M. (2012). *Analog vlsi implementation of neural systems* (Vol. 80). Springer Science & Business Media.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(3), 320-334. , doi:10.1037/0097-7403.9.3.320

Mel, B. W. (1993). Synaptic integration in an excitable dendritic tree. *Journal of Neurophysiology*, *70*(3), 1086–1101.

Morgan, M. J., Raphael, S., Tibber, M. S., & Dakin, S. C. (2014). A texture-processing model of the visual sense of number. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1790), 20141137. , doi:10.1098/rspb.2014.1137

Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, *297*(5587), 1708–1711. , doi:10.1126/science.1072493

Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review*, *1*(1), 29–55. , doi:10.3758/BF03200760

Park, J., DeWind, N. K., Woldorff, M. G., & Brannon, E. M. (2015). Rapid and direct encoding of numerosity in the visual stream. *Cerebral Cortex*, bhv017. , doi:10.1093/cercor/bhv017

Phong, B. T. (1975). Illumination for computer generated pictures. *Communications of the ACM*, *18*(6), 311–317. , doi:10.1145/360825.360839

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., et al. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition*, *116*(1), 33–41. , doi:10.1016/j.cognition.2010.03.012

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 547–555. , doi:10.1016/j.neuron.2004.10.014

Piazza, M., Mechelli, A., Butterworth, B., & Price, C. J. (2002). Are subitizing and counting implemented as separate or functionally overlapping processes? *NeuroImage*, *15*(2), 435–446. , doi:10.1006/nimg.2001.0980

Raphael, S., Dillenburger, B., & Morgan, M. (2013). Computation of relative numerosity of circular dot textures. *Journal of Vision*, *13*(2). , doi:10.1167/13.2.17

Reich, D. S., Mechler, F., Purpura, K. P., & Victor, J. D. (2000). Interspike intervals, receptive fields, and information encoding in primary visual cortex. *The Journal of neuroscience*, *20*(5), 1964–1974.

Resnikoff, H. L., & Wells Jr, R. O. (2015). *Mathematics in civilization*. Courier Dover Publications.

Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annu. Rev. Neurosci.*, *27*, 611–647. , doi:10.1146/annurev.neuro.26.041002.131039

Roitman, J. D., Brannon, E. M., & Platt, M. L. (2007). Monotonic coding of numerosity in macaque lateral intraparietal area. *PLoS biology*, *5*(8), e208. , doi:10.1371/journal.pbio.0050208

Ross, J. (2003). Visual discrimination of number without counting. *Perception*, *32*(7), 867–870. , doi:10.1068/p5029

Ross, J., & Burr, D. C. (2010). Vision senses number directly. *Journal of Vision*, *10*(2), 10.1–8. , doi:10.1167/10.2.10

Salinas, E., & Thier, P. (2000). Gain modulation: a major computational principle of the central nervous system. *Neuron*, *27*(1), 15–21. , doi:10.1016/S0896-6273(00)00004-0

Schlick, C. (1993). A customizable reflectance model for everyday rendering. In M. Cohen, C. Puech, & F. Sillion (Eds.), *Fourth eurographics workshop on rendering* (pp. 73–83). Paris.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature neuroscience*, *4*(8), 819–825.

Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *Computer vision and pattern recognition, 2005. cvpr 2005. ieee computer society conference on* (Vol. 2, pp. 994–1000).

Starkey, P., Spelke, E. S., & Gelman, R. (1990). Numerical abstraction by human infants. *Cognition*, *36*(2), 97–127. , doi:10.1016/0010-0277(90)90001-Z

Stoianov, I., & Zorzi, M. (2012). Emergence of a 'visual number sense' in hierarchical generative models. *Nature Neuroscience*, *15*(2), 194–196. , doi:10.1038/nn.2996

Strauss, M. S., & Curtis, L. E. (1981). Infant perception of numerosity. *Child Development*, *52*(4), 1146–1152. , doi:10.2307/1129500

Suzuki, K., Horiba, I., & Sugie, N. (2003). Linear-time connected-component labeling based on sequential local operations. *Computer Vision and Image Understanding*, *89*(1), 1–23. , doi:10.1016/S1077-3142(02)00030-9

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. *Psychological review*, *101*(1), 80. , doi:10.1037/0033-295X.101.1.80

Umbaugh, S. E. (2005). *Computer imaging: digital image analysis and processing*. CRC press.

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: a neural model. *Journal of Cognitive Neuroscience*, *16*(9), 1493–1504. , doi:10.1162/0898929042568497

Wallis, G., & Bülthoff, H. (1999). Learning to recognize objects. *Trends in cognitive sciences*, *3*(1), 22–31. , doi:10.1016/S1364-6613(98)01261-3

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488. , doi:10.1016/j.tics.2003.09.002

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*(2), 130–137. , doi:10.1111/1467-9280.00120

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1–B11. , doi:10.1016/S0010-0277(99)00066-9

Young, R. A. (1987). The gaussian derivative model for spatial vision: I. retinal mechanisms. *Spatial vision*, *2*(4), 273–293. , doi:10.1163/156856887X00222

Young, R. A., & Lesperance, R. M. (2001). The gaussian derivative model for spatial-temporal vision: Ii. cortical data. *Spatial vision*, *14*(3), 321–389. , doi:10.1163/156856801753253591

Zetzsche, C., & Barth, E. (1990a). Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision Research*, *30*(7), 1111–1117.

Zetzsche, C., & Barth, E. (1990b). Image surface predicates and the neural encoding of two-dimensional signal variations. In *Human vision and electronic imaging: Models, methods, and applications* (Vol. 1249, pp. 160–177).

Zetzsche, C., Gadzicki, K., & Kluth, T. (2013, April). Statistical invariants of spatial form: From local and to numerosity. In O. Kutz, M. Bhatt, S. Borgo, & P. Santos (Eds.), *Proc. of the 2nd interdisciplinary workshop the shape of things* (Vol. 1007, pp. 163–172). Aachen: CEUR-WS.org.

Zetzsche, C., & Nuding, U. (2005). Nonlinear and higher-order approaches to the encoding of natural scenes. *Network: Computation in Neural Systems*, *16*(2-3), 191–221.

Zetzsche, C., & Roehrbein, F. (2001). Nonlinear and extra-classical receptive field properties and the statistics of natural scenes. *Network: Computation in Neural Systems*, *12*(3), 331–350. , doi:10.1080/net.12.3.331.350

4 Action selection for object recognition and the influence of isotropic $i2D$-features

Where to look to identify an object? This is the major question of this section. As has been discussed in Section 1.1, the receptors on the retina are not equally distributed. The fovea has a high resolution and in the periphery the resolution decreases with an increase in eccentricity. To gather new important information, the eye has to perform saccades. Then the new fixated region lies in the high resolution area of the fovea again. How the human visual system decides where to look next is still ongoing research. Assuming that the human visual system is optimized by evolution and that this also holds true for the control of eye movements, the behavioral findings and developed models are good candidates to improve active computer vision models for object recognition. Although the biological example can move through the environment and move its eyes, the majority of standard object recognition approaches in computer vision have a static nature. If a system is able to perform movements or if it can interact with the environment, not using this ability results in less information available to the system. An unsystematic approach of performing actions cannot guarantee an increase in the perceived information within a limited time. In order to circumvent this fact, a systematic action selection strategy is required. In particular, information theoretical concepts are applied to an active vision system, i.e. a steerable camera, within this chapter. The concept of information gain is used to determine the most informative next movement to perform object recognition tasks. The active vision model is developed in the context of a sensorimotor system and it is evaluated in Section 4.3. Further analysis and an interpretation in the context of Gibson's affordances are presented in Section 4.4. And finally in Section 4.5, various action selection strategies are implemented and compared to each other. This also includes an $i2D$-feature approach which is based on the concept of a clipped eigenvalue operator. The used operator is derived in Section 4.2 and it is also proved that this operator defines an $i2D$-system.

4.1 Related Work

In this section a brief overview about related findings in the literature is given. In order to avoid repetition, this section is restricted to a minimum and the reader is referred to the introductory parts of the articles in Sections 4.3, 4.4, and 4.5.

The classical view on visual perception is that eye movements only have the function to fixate objects or parts of them. In this case vision is interpreted as a sequence of snapshots of the environment [23]. Based on this classical view, vision models were developed which conceptually separate the perceptual sensory process and the movement process. This is also the case for common object recognition approaches, e.g. the feed-forward HMAX model [71].

Opposed to the classical view, Gibson proposed his *ecological approach to visual perception*

[25] which states that an agent has a relation to its environment. This relation has to be taken into account to explain the behavior of the agent. Research in the direction of affordances, i.e. interaction possibilities offered by the environment dependent on the capabilities of a particular agent, provides evidence that affordances are key ingredients of the perceptual process, see [27] for a review. Studies regarding object recognition show that visual information of a manipulable object cause an activation of representations of actions which can typically be executed on the object [28]. Action and perception thus seem to have an intertwined relation and also depend on each other to some degree.

A stronger relation between action and perception is proposed by the *sensorimotor* approach in [58, 59]. The sensory and the motor part are not treated independently anymore. Rather, the information is combined in a sensorimotor representation where the classic notion of separate cognitive processing stages for sensory and motor information does not hold, i.e. they are integrated into one sensorimotor coding.

Active vision approaches based on the concept of an sensorimotor system allow to control the movements performed by the system. Equipped with an appropriate cost function the movements can be determined to obtain new information after performing the chosen movement. Information theoretical approaches to the cost function were used for object recognition [66, 9] and in particular an information gain strategy was used for scene recognition [67, 96].

The relation between image features extracted by $i2D$-selective operators and saccadic eye movements was investigated in [45, 68]. It was shown that saccadic eye movements can be predicted by image features extracted by $i2D$-operators. The fixation points of humans were analyzed in [95] regarding their higher-order statistics. It turned out that the subjects have a bias for image regions with multiple spatial frequencies, e.g. corners. This result supports the use of $i2D$-operators as a detector for informative image regions.

4.2 Mathematical preliminaries

This section briefly reviews the *generalized curvature* and the *clipped eigenvalue* which were introduced by Zetzsche and Barth [87, 6] in the context of intrinsic two-dimensional operators. The clipped eigenvalue is used as a bottom-up region-of-interest detector for object recognition in Section 4.5. It is also shown that the operators defined by the generalized curvature and the clipped eigenvalue are able to satisfy the conditions of an $i2D$-operator. The generalized curvature is similar to the curvature operator T_1 defined in Theorem 3.24 as can be seen in its following definition.

Definition 4.1 (Generalized curvature)**.** The operator $T : C^2(\Omega) \to C(\Omega)$ with compact $\Omega \subset \mathbb{R}^2$ is defined for $n \in \mathbb{N}$ by

$$T(u)(x) := \frac{1}{4}\left((\Delta u)^2 - \epsilon_n(u)^2\right) \tag{4.1}$$

with eccentricity

$$\epsilon_n(u)^2 := (c_n * u)^2 + (s_n * u)^2. \tag{4.2}$$

The convolution kernels c_n and s_n are defined by their Fourier transform in polar coordinates ($x_1 = r\cos(\phi)$, $x_2 = r\sin(\phi)$) by

$$\mathcal{F}(c_n)(r,\phi) = (i)^n f(r)\cos(n\phi), \tag{4.3}$$
$$\mathcal{F}(s_n)(r,\phi) = (i)^n f(r)\sin(n\phi). \tag{4.4}$$

f is a continuous function of the radius r. The operator T is then called the *generalized curvature* operator.

Within this definition it is not clear what kind of functions f are a good choice and where the term curvature has its origin. The following lemma clarifies these issues.

Lemma 4.2. *Let $f(r) = 2\pi r^2$ and $n = 2$. Then the generalized curvature operator T becomes*

$$T(u)(x) = \frac{\partial^2}{\partial x_1^2}u\frac{\partial^2}{\partial x_2^2}u - (\frac{\partial^2}{\partial x_1 \partial x_2}u)^2. \tag{4.5}$$

Note that this expression is exactly the nominator of the Gaussian curvature as can be seen in Theorem 3.24. Thus, the operator with these parameters automatically fulfills the requirements of an i2D-operator.

Proof. For convenience the index of the function u determines the derivative in the corresponding variable, .i.e. $u_i := \frac{\partial}{\partial x_i}u$. We start from the nominator of the Gaussian curvature

$$\begin{aligned} & u_{11}u_{22} - u_{12}^2 \\ =& \frac{1}{4}(u_{11}+u_{22})^2 - \frac{1}{4}\underbrace{((u_{11}-u_{22})^2 + 4u_{12}^2)}_{=:\epsilon^2} \\ =& \frac{1}{4}\left((\Delta u)^2 - \epsilon^2\right) \end{aligned} \tag{4.6}$$

where ϵ is the eccentricity. The eccentricity can be rewritten by

$$\epsilon = ((u_{11}-u_{22})^2 + (2u_{12})^2). \tag{4.7}$$

If we are now able to rewrite the base of each summand by a convolution with the right filter kernel, we are done. For this purpose we use the following property of the Fourier transform.

$$\mathcal{F}(D^\alpha f)(z) = i^{|\alpha|} z^\alpha \mathcal{F}(z) \tag{4.8}$$

where $\alpha \in \mathbb{N}^n$ is a multi-index ($D^\alpha = D_1^{\alpha_1} \dots D_n^{\alpha_n}$ and $z^\alpha = z_1^{\alpha_1} \cdot \dots \cdot z_n^{\alpha_n}$). Applying this to $u_{11} - u_{22}$ with $z_1 = r\cos(\phi)$ and $z_2 = r\sin(\phi)$ yields

$$\begin{aligned}
\mathcal{F}(u_{11} - u_{22})(z) &= i^2 z_1^2 \mathcal{F}(u)(z) - i^2 z_2^2 \mathcal{F}(u)(z) \\
&= (z_2^2 - z_1^2)\mathcal{F}(u)(z) \\
&= r^2(\sin(\phi)^2 - \cos(\phi)^2)\mathcal{F}(u)(z) \\
&= -r^2\cos(2\phi)\mathcal{F}(u)(z) = \mathcal{F}(c_2)(z)\mathcal{F}(u)(z) = \frac{1}{2\pi}\mathcal{F}(c_2 * u)
\end{aligned} \tag{4.9}$$

with $f(r) := 2\pi r^2$. Applying the derivative relation of the Fourier transform to $2u_{12}$ yields

$$\begin{aligned}
\mathcal{F}(2u_{12})(z) &= 2i^2 z_1 z_2 \mathcal{F}(u)(z) \\
&= -2r^2\cos(\phi)\sin(\phi)\mathcal{F}(u)(z) \\
&= -r^2\sin(2\phi)\mathcal{F}(u)(z) = \mathcal{F}(s_2)(z)\mathcal{F}(u)(z) = \frac{1}{2\pi}\mathcal{F}(s_2 * u)
\end{aligned} \tag{4.10}$$

with $f(r) = 2\pi r^2$. This proves the assumption. □

This lemma also states that the generalized curvature is an $i2D$-operator for $n = 2$. The following theorem generalizes this statement.

Theorem 4.3. *Let $n = 2, 4, 6, \dots$ and $f(r) = 2\pi r^2$. Then the generalized curvature operator defined in Definition 4.1 is an $i2D$-operator.*

Proof. The main goal is to show that $(\Delta u)^2 - \epsilon_n^2$ does not respond to $i0D$- and $i1D$-points in a signal. We rewrite the operator by an equivalent second-order Volterra system and use Theorem 2.10 to show that this system is an $i2D$-system. First the Fourier transform of the Laplace operator is derived. By using Equation (4.8), $z_1 = r\cos(\phi)$, and $z_2 = r\sin(\phi)$ we get

$$\begin{aligned}
\mathcal{F}(u_{11} + u_{22})(z) &= -(z_1^2 + z_2^2)\mathcal{F}(u)(z) \\
&= -r^2\mathcal{F}(u)(z) =: \frac{1}{2\pi}\mathcal{F}(l * u)(z)
\end{aligned} \tag{4.11}$$

such that $\mathcal{F}(l)(r, \phi) = -2\pi r^2$. We thus get

$$(\Delta u)^2 - \epsilon_n^2 = (l * u)^2 - (c_n * u)^2 - (s_n * u)^2. \tag{4.12}$$

Let $h \in L^1(\mathbb{R}^2$ and $u \in L^2(\mathbb{R}^2)$ be arbitrary functions. Then the following holds

$$((h * u)(x))^2 = \int_{\mathbb{R}^4} h(x_1)h(x_2)u(x - x_1)u(x - x_2)\ dx_1\ dx_2. \tag{4.13}$$

This expression can be described by a second-order Volterra system with the second-order Volterra kernel $\tilde{h}(x_1, x_2) = h(x_1)h(x_2)$. With $z = (z_1, z_2)^T \in \mathbb{R}^4$ and $z_1, z_2 \in \mathbb{R}^2$ the Fourier transform can be determined by

$$\mathcal{F}(\tilde{h})(z) = \mathcal{F}(h)(z_1)\mathcal{F}(h)(z_2). \tag{4.14}$$

ow we can apply this relation to the resulting operator $\tilde{l} - \tilde{c}_n - \tilde{s}_n$ of the right-hand side of Equation (4.12) such that the Fourier transform in polar coordinates becomes

$$\begin{aligned}
&\mathcal{F}(\tilde{l} - \tilde{c}_n - \tilde{s}_n)(r_1, \phi_1, r_2, \phi_2) \\
&= (2\pi)^2 r_1^2 r_2^2 - (i)^{2n}(2\pi)^2 r_1^2 r_2^2 \cos(n\phi_1)\cos(n\phi_2) - (i)^{2n}(2\pi)^2 r_1^2 r_2^2 \sin(n\phi_1)\sin(n\phi_2) \\
&= (2\pi)^2 r_1^2 r_2^2 \left(1 - (-1)^n(\cos(n\phi_1)\cos(n\phi_2) + \sin(n\phi_1)\sin(n\phi_2))\right).
\end{aligned} \tag{4.15}$$

In order to apply Theorem 2.10, we have to guarantee that the function values of the Fourier transform vanish for all arguments $z \in \mathbb{R}^4$ given by

$$z = \begin{pmatrix} r_1\cos(\phi) \\ r_1\sin(\phi) \\ r_2\cos(\phi) \\ r_2\sin(\phi) \end{pmatrix}, \forall r_1, r_2 \geq 0,\ \phi \in [0, 2\pi]. \tag{4.16}$$

This parametrization matches the polar coordinate system in Equation (4.15) with $\phi_1 = \phi_2 = \phi$. Then it follows

$$\mathcal{F}(\tilde{l} - \tilde{c}_n - \tilde{s}_n)(r_1, \phi, r_2, \phi) = (2\pi)^2 r_1^2 r_2^2 \left(1 - (-1)^n \underbrace{(\cos(n\phi)^2 + \sin(n\phi)^2)}_{=1}\right). \tag{4.17}$$

Consequently it is equal to zero for even n. With Theorem 2.10 follows the assumption. □

Remark 4.4. A Gaussian blurring is sometimes incorporated in $f(r)$ given by

$$f(r) = 2\pi r^2 e^{\frac{1}{2}\frac{r^2}{\sigma_r^2}}. \tag{4.18}$$

Note that the resulting operator is not an $i2D$-operator in a strict sense anymore. An increase of σ_r causes an increase in $i2D$-selectivity.

It was shown that the generalized curvature becomes the nominator of the Gaussian curva-

ture. One important property of the Gaussian curvature is its sign classifying different kinds of curvature, i.e. elliptic, hyperbolic, and parabolic curvature. If one wants to distinguish the two different types of elliptic curvature which can be seen as a valley or a hill for illustrative purposes, another operator is required. The clipped eigenvalue operator with respect to the generalized curvature in the following definition solves this problem.

Definition 4.5 (Clipped eigenvalue)**.** The operator $N : C^2(\Omega) \to C(\Omega)$ with compact $\Omega \subset \mathbb{R}^2$ is defined for $n \in \mathbb{N}$ by

$$N(u)(x) = |\min(0, \frac{1}{2}(\Delta u + |\epsilon_n(u)|))| - |\max(0, \frac{1}{2}(\Delta u - |\epsilon_n(u)|))| \tag{4.19}$$

where ϵ_n is as defined in Definition 4.1. The operator N is called the *clipped eigenvalue* operator. $\lambda_1(u) = \frac{1}{2}(\Delta u + |\epsilon_n(u)|)$ and $\lambda_2(u) = \frac{1}{2}(\Delta u - |\epsilon_n(u)|)$ are the *generalized eigenvalues.*

Remark 4.6. Note that the product of the generalized eigenvalues is exactly the generalized curvature. The term eigenvalues is motivated by the case $n = 2$ and $f(r) = 2\pi r^2$ where the generalized curvature becomes the denominator of the Gaussian curvature. Then the generalized eigenvalues become the eigenvalues of the Hessian matrix of u.

The following theorem concludes this section with a statement which gives the warranty to use the clipped eigenvalue operator as an $i2D$-operator in the subsequent application.

Theorem 4.7. *Let $n = 2, 4, 6, \ldots$ and $f(r) = 2\pi r^2$. Then the clipped eigenvalue operator defined in Definition 4.5 is an $i2D$-operator.*

Proof. Let $u \in C(\Omega)$ be an arbitrary signal and let $x \in I_0(u) \cup I_1(u)$ be an arbitrary point. The relation between the generalized eigenvalues and the generalized curvature is

$$T(u) = \lambda_1(u)\lambda_2(u). \tag{4.20}$$

Note that $\lambda_1(u) \geq \lambda_2(u)$ by definition. We thus can distinguish three cases such that the clipped eigenvalue operator becomes

$$N(u)(x) = \begin{cases} -\lambda_2(u)(x) & , \ 0 \leq \lambda_2(u)(x) \leq \lambda_1(u)(x), \\ -\lambda_1(u)(x) & , \ \lambda_2(u)(x) \leq \lambda_1(u)(x) \leq 0, \\ 0 & , \ \lambda_2(u)(x) \leq 0 \leq \lambda_1(u)(x). \end{cases} \tag{4.21}$$

With $x \in I_0(u) \cup I_1(u)$ follows $T(u)(x) = 0$. Consequently, the product of the generalized eigenvalues is also zero. As a result at least one generalized eigenvalue has to be zero. In the first case, where both generalized eigenvalues are positive, $\lambda_2(u)(x) = 0$ or both are zero. Thus $N(u)(x) = 0$. In the second case where both generalized eigenvalues are negative, $\lambda_1(u)(x) = 0$ or both are zero. Thus $N(u)(x) = 0$. The third case is always zero. This holds for arbitrary $x \in I_0(u) \cup I_1(u)$ such that the operator N is an $i2D$-operator. □

4.3 Article: Active sensorimotor object recognition in three-dimensional space

Reference

D.N., T.K., T.R. and C.Z. designed research; D.N., T.K., and T.R. performed research; D.N and T.K. developed and implemented the system; D.N. analyzed the data; D.N., T.K., T.R., C.Z. and K.S. wrote the paper.

The paper was published in *Spatial Cognition IX* under the following reference [54]:

D. Nakath, T. Kluth, T. Reineking, C. Zetzsche, and K. Schill. Active sensorimotor object recognition in three-dimensional space. In C. Freksa, B. Nebel, M. Hegarty, and T. Barkowsky, editors, *Spatial Cognition IX*, volume 8684 of *Lecture Notes in Computer Science*, pages 312–324. Springer International Publishing, 2014.

Active Sensorimotor Object Recognition in Three-Dimensional Space

David Nakath, Tobias Kluth, Thomas Reineking,
Christoph Zetzsche, and Kerstin Schill

Cognitive Neuroinformatics, University of Bremen,
Enrique-Schmidt-Straße 5, 28359 Bremen, Germany
dnakath@informatik.uni-bremen.de
http://www.informatik.uni-bremen.de/cog_neuroinf/en

Abstract. Spatial interaction of biological agents with their environment is based on the cognitive processing of sensory as well as motor information. There are many models for sole sensory processing but only a few for integrating sensory and motor information into a unifying sensorimotor approach. Additionally, neither the relations shaping the integration are yet clear nor how the integrated information can be used in an underlying representation. Therefore, we propose a probabilistic model for integrated processing of sensory and motor information by combining bottom-up feature extraction and top-down action selection embedded in a Bayesian inference approach. The integration of sensory perceptions and motor information brings about two main advantages: (i) Their statistical dependencies can be exploited by representing the spatial relationships of the sensor information in the underlying joint probability distribution and (ii) a top-down process can compute the next most informative region according to an information gain strategy. We evaluated our system in two different object recognition tasks. We found that the integration of sensory and motor information significantly improves active object recognition, in particular when these movements have been chosen by an information gain strategy.

Keywords: sensorimotor, object recognition, Bayesian inference, information gain.

1 Introduction

The capabilities of artificial systems are easily exceeded by humans when it comes to perception-based interaction with the environment. With this in mind, it seems reasonable to take a closer look at the main principles of the human visual perception process. Especially the reciprocal advantageous interplay of motion and sensory information, which was early recognized by Gibson [1] and Neisser [2], should be considered here. Based on analog arguments, an "active perception" approach was proposed by [3,4,5]. The strong interrelation between movements in space and corresponding sensory perceptions can foster the even stronger concept of a *sensorimotor representation* [6,7,8,9]. In this concept, the

C. Freksa et al. (Eds.): Spatial Cognition 2014, LNAI 8684, pp. 312–324, 2014.

classic notion of separate cognitive processing stages for sensory and motor information does not hold. In fact, they are integrated into one sensorimotor coding. This is a precondition for a sensorimotor representation which is established from the specific pattern of alternating sensory perceptions and spatial motor actions [6,10]. The constant checking and confirmation of sensory and motor information against an internal cognitive model then constitutes a scanpath, and thus the perception of a particular object [11,12].

To be able to check such an internal sensorimotor model of an object, the next motor action has to be chosen accordingly by an object recognition system. Generally, the problem of action selection can be solved in numerous ways, but as information gathering should be the purpose of motor actions it seems reasonable to choose an information-theoretic criterion. Prior research in this area often found that the principle of *information gain* is well suited to select an appropriate next action. This has been shown by [13] in the context of decision trees, where information gain was used to decide which attributes are the most relevant ones. Robotics also proofed to be a suitable domain, as information gain can be used there to actively reduce the uncertainty of the robot regarding its position and spatial environment [14,15]. Additionally, information gain was not only used to explain human selection behavior [16,17] but also to mimic it: Both in the form of human-like expert systems [18,19] and with a modeled sensorimotor loop in a saccadic eye movement control system [20,21].

Based on the preceding considerations the basic sketch of a sensomotoric object recognition system becomes apparent. Building upon the research of Schill, Zetzsche, and coworkers [20,21], we propose a sensomotoric probabilistic reasoning system for active object recognition integrating sensory perceptions and motor actions. Our system is inspired by the human perception process and therefore should model a sequential pattern of actions controlled by a top-down and a bottom-up process. Evidence suggests that sensory perception and motor actions partly share the same cognitive processing stage which makes it reasonable to integrate them into one single sensorimotor feature (SMF) [22,23,24]. Through this integration two improvements come into effect: (i) The accuracy of the recognition process is improved through the additional motor information which encodes spatial relations and (ii) the next motor action can be chosen according to the maximum information gain principle, thus supplying the sensors with an optimized input in the next recognition step.

The basic architecture of the system we propose is outlined in Sect. 2. In Sect. 3, we describe the implementation of the system. Subsequently, Sect. 4 shows the results of the evaluation in two different scenarios: Optimized control of 3D movements of a camera mounted on a robotic arm and simulated sensor movements on images from the Caltech 256 dataset. The paper is concluded with a discussion of the specific advantages offered by the proposed sensorimotor architecture.

2 Sensomotoric Object Recognition System

The sensorimotor system described in the following is a generic architecture (see Fig. 1). In the case of visual object recognition, the senorimotor loop starts out with a particular pose of the active sensor which passes its raw sensor data to the sensory processing module. After processing, the sensory data becomes part of a new sensorimotor feature, which is then fed into the probabilistic reasoning module. The Bayesian inference module calculates the new posterior distribution based on a previously-learned sensorimotor representation. The posterior distribution constitutes the current belief of the system. This belief is used by the information gain strategy to choose an optimal next movement from the set of possible motor actions. The selected movement then also becomes part of the sensorimotor feature and is subsequently executed by the active sensor. The whole process results in a new sensor pose, which in turn delivers new raw sensory data to enter the next cycle of the sensorimotor loop.

More formally spoken, the system depends on an *active sensor* (AS), which can be controlled such that it delivers information about a specific aspect of the

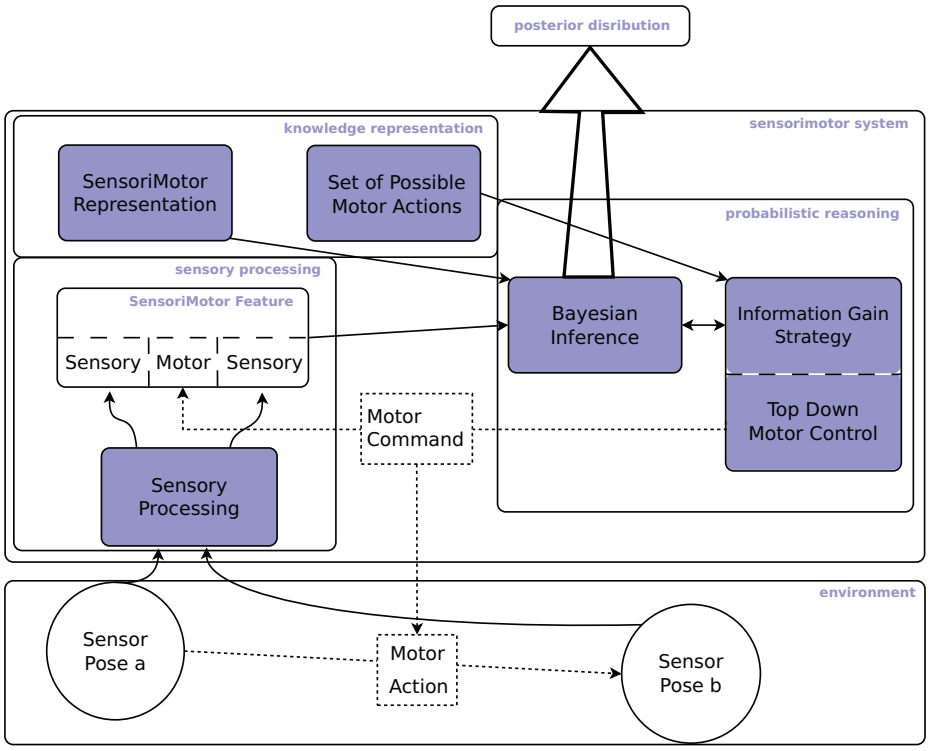


Fig. 1. Sensorimotor Information Processing System

world. In Fig. 1, the two arrows pointing from the sensor poses to the sensory processing module correspond to the mapping $A : U \times X \to R$, where U is the space of all motor actions which are currently possible, X is the state space of the active sensor, and R is the raw sensor data space.

The system has no knowledge about the actual state of the AS, instead it is only informed about the currently possible motor actions U. These are of course dependent on the state in the sense of $U : X \to \mathcal{P}(\Omega_U)$, where Ω_U is the set of all possible motor controls and $\mathcal{P}$ denotes the power set. Assuming that the output of the function U is given, we write U instead of $U(x),\ x \in X$, for convenience. Considering the state-agnostic behavior, the formal representation of the AS can be redefined to $A_x : U \to R$ where the index x recalls the dependency on the state

$$A_x(u) := A(x, u) = r,\ x \in X,\ u \in U(x),\ r \in R. \tag{1}$$

The only time-dependent variables are the sensor state x and the relative motor control u. In contrast, the world is assumed to be static which implies no dynamic changes in the raw sensor data $r \in R$.

This data is fed into the *sensory processing* (SP) which mainly extracts the relevant features belonging to a feature space F, i.e., $SP : R \to F$. Subsequently, the quantization operation $Q_S : F \to S$ maps the features to a specific feature class in the finite and countable space S. The possible motor actions are mapped with $Q_M : \Omega_U \to M$ to the finite countable set of actions M to yield a manageable product space of sensory and motor information. The results of these quantizations then become part of a sensorimotor feature (SMF). The single quantizations are represented in Fig. 1 by the arrows from the sensory processing and the motor command to the first-order sensorimotor feature which is defined as the triple

$$SMF_i := \{s_{i-1}, m_{i-1}, s_i\}, \tag{2}$$

where $m_{i-1} := Q_M(u_{i-1})$ is the intermediate motor action between the sensor information s_{i-1} and s_i at time step t_{i-1} and t_i (see Fig. 2). The whole chain of operations to obtain the sensor information at a time step t_i can be described by

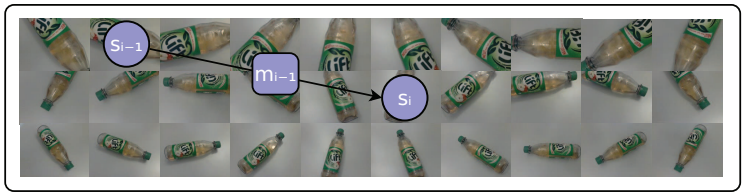
$$s_i := (Q_S \circ SP \circ A_x)(u_{i-1}). \tag{3}$$

The *knowledge representation* is comprised of the currently available motor actions U and the learned sensorimotor representation (SMR), which is a full joint probability distribution of $SMFs$ and the classes represented by the discrete random variable Y. Every possible SMF is generated on a set of known objects in a training phase. This means that, from every possible state x, every possible motor action u is performed, resulting in

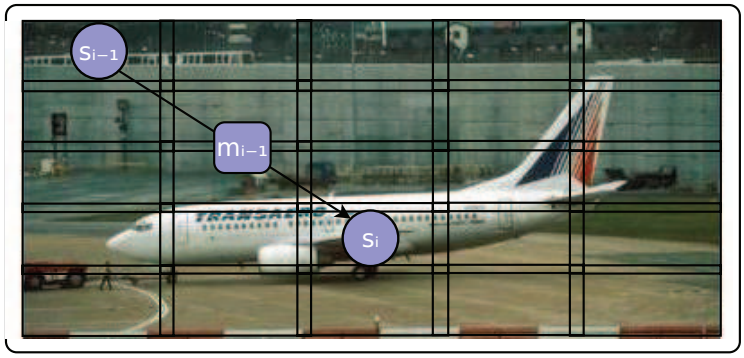
$$SMR := P(SMF, Y) = P(S_{i-1}, M_{i-1}, S_i, Y). \tag{4}$$

The *probabilistic reasoning* module consists of a Bayesian inference approach accompanied by an information gain strategy. They rely on bottom-up sensory data and top-down information from the knowledge representation. This design enables the Bayesian inference system to take into account motor actions, thus

316 D. Nakath et al.



(a) Robotic arm in 3D space



(b) Simulated active sensor in 2D space

Fig. 2. Exemplary sensorimotor features (SMF) drawn on the discretized views on an object, both for the 3D case shown in (a) and 2D case shown in (b). Here, s_{i-1} denotes the preceding sensory input, m_{i-1} denotes the preceding movement, and s_i denotes the current sensory input.

improving the posterior distribution over the object classes Y. Furthermore, the information gain strategy can choose an optimal next motor action for the active sensor, thus improving the input of the following Bayesian inference step.

3 Model Implementation

Based on the schematic outline presented above, we applied our system in the field of active object recognition. We consider both the case of an active sensor moving in 3D space and a simulated active sensor moving in 2D space (see Fig. 2).

3.1 Active Sensor Implementation

For the 3D case, we used a discrete set M of movements of a camera mounted on a robotic arm (see Fig. 2a), which resembles an observer actually moving around an object. For the 2D case, we used simulated active sensor movements

on images of a reduced version of the Caltech 256 [25] dataset.[1] Here, M consists of all possible relative movements between the individual cells of a 5×5 grid (see Fig. 2b). The latter setup mimics eye movements of a stationary observer. Hence, in both cases holds $\Omega_U = M$ and the quantization Q_M is an identity operation.

Although the implemented sensors are of a fundamentally different nature, the following basic learning and recognition principles can be applied to both of them: In the learning phase, features are extracted from the training data (i.e., images from every reachable state of the active sensor), which corresponds to the mapping SP introduced above. As the robotic arm relies on views showing the entire object, GIST-features [26,27] are used, while the more local image patches of the 2D case are described by the SURF-feature [28] with the highest score on that patch. The quantization Q_S is then learned by performing a k-means clustering on the extracted features ($k = 15$).[2] In order to build the individual $SMFs$, features are extracted (see SP) and the results are assigned to clusters with the aid of the previously defined mapping Q_S. These labels are combined with the corresponding intermediate movement resulting in a set of $SMFs$. Finally, all generated $SMFs$ are stored in a Laplace-smoothed SMR.

3.2 Probabilistic Reasoning

The probabilistic reasoning is comprised of a Bayesian inference module in the form of a dynamic Bayesian network (BN) and a corresponding information gain strategy. Four of these probabilistic reasoning modules were implemented to examine the difference between *sensor networks*, which only take into account sensory information (which also implies that no information gain strategy is used), and *sensorimotor networks*, which take into account integrated sensory and motor information. The object recognition in the sense of machine vision then takes place by classification which is performed by choosing the class with the maximum posterior probability.

The first representative of the *sensor networks* is Bayesian network 1 (BN1) (see Fig. 3a), which resembles a naive Bayes model only taking into account the current sensory input s_i. Thus, the inference can be performed by

$$P(y|s_{1:n}) = \alpha P(y) \prod_{i=1}^{n} P(s_i|y), \tag{5}$$

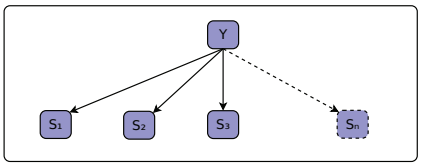
where α is a normalizing constant guaranteeing that the probability function properties are satisfied and $s_{1:n}$ is a short notation for the n-tuple $(s_1, \ldots, s_n)$.

The second representative of the *sensor networks* is Bayesian network 2 (BN2) (see Fig. 3b), which assumes additional statistical dependencies between the preceding and the current sensor information, s_{i-1} and s_i, resulting in

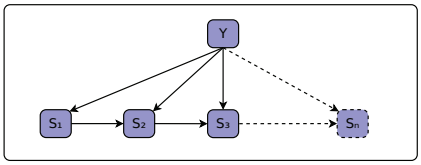
[1] The reduced dataset consists in each case of 100 randomly-selected images from the object classes: *airplanes*, *cowboy-hat*, *faces*, *motorbikes*, *swan*, *breadmaker*, *diamond-ring*, *ketch*, *self-propelled-lawn-mower*, and *teapot*.

[2] We use only a small number of clusters in order to limit the number of model parameters and to prevent overfitting.

318 D. Nakath et al.



(a) BN1



(b) BN2

Fig. 3. The dynamic Bayesian sensor networks process only sensory information. BN1, which is shown in (a), represents a naive Bayes approach where the current sensory input s_i depends only on the object hypothesis Y. BN2, which is shown in (b), assumes statistical dependencies between the object hypothesis Y and the preceding sensory input s_{i-1} for every sensory input s_i.

$$P(y|s_{1:n}) = \alpha P(y)P(s_1|y)\prod_{i=2}^{n} P(s_i|s_{i-1}, y). \tag{6}$$

Bayesian network 3 (BN3) (see Fig. 4a) uses the full information of the SMF and therefore belongs to the *sensorimotor networks*. The assumption that the current sensory input s_i depends on the preceding sensory input s_{i-1} and the intermediary motor action m_{i-1} integrates motor and sensor information in the recognition process and permits the application of the information gain strategy to choose the next optimal movement. The inference can then be conducted by

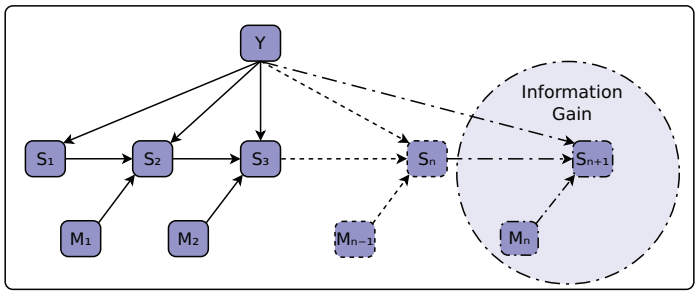
$$P(y|s_{1:n}, m_{1:n-1}) = \alpha P(y)P(s_1|y)\prod_{i=2}^{n} P(s_i|s_{i-1}, m_{i-1}, y). \tag{7}$$

Bayesian network 4 (BN4) (see Fig. 4b) mainly resembles BN3, but additionally allows statistical dependencies between the preceding sensory input s_{i-1} and the motor action m_{i-1}. The inference can thus be conducted by

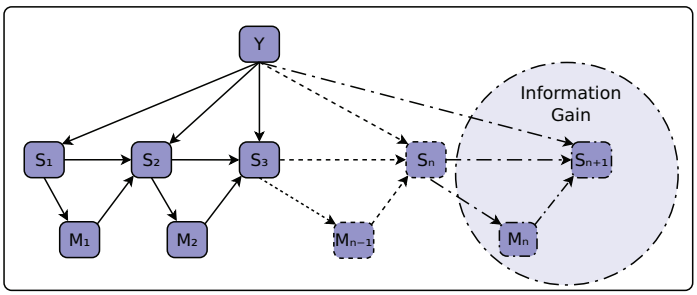
$$P(y|s_{1:n}, m_{1:n-1}) = \alpha P(y)P(s_1|y)\prod_{i=2}^{n} P(s_i|s_{i-1}, m_{i-1}, y)P(m_{i-1}|s_{i-1}). \tag{8}$$

3.3 Information Gain

The strategy for action selection should satisfy two main properties: (i) The strategy should adapt itself to the current belief state of the system and (ii) the



(a) BN3



(b) BN4

Fig. 4. The dynamic Bayesian sensorimotor networks process integrated sensorimotor information. As motor information is taken into account, the next movement m_n can be chosen by the information gain strategy. BN3, which is shown in (a), assumes statistical dependencies between the current visual input s_i, the preceding movement m_{i-1}, the visual input s_{i-1}, and the hypothesis y. BN4, which is shown in (b), allows for the same dependencies as BN3 and, in addition, allows for a dependency of the preceding movement m_{i-1} on the preceding visual input s_{i-1}.

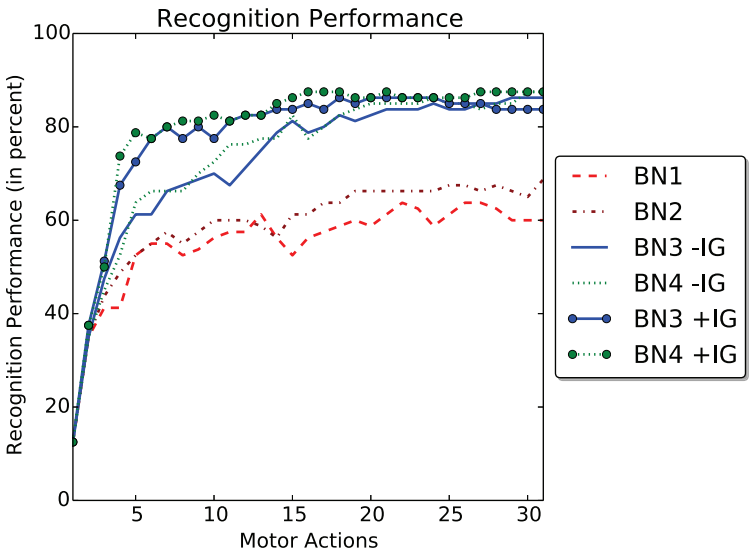
strategy should not make decisions in an heuristic fashion but tightly integrated into the axiomatic framework used for reasoning. The information gain strategy presented in this paper complies with both of these properties.

The information gain IG of a possible next movement m_n is defined as the difference between the current entropy $H(Y)$ and the conditional entropy $H(Y|S_{n+1}, m_n)$, i.e.,

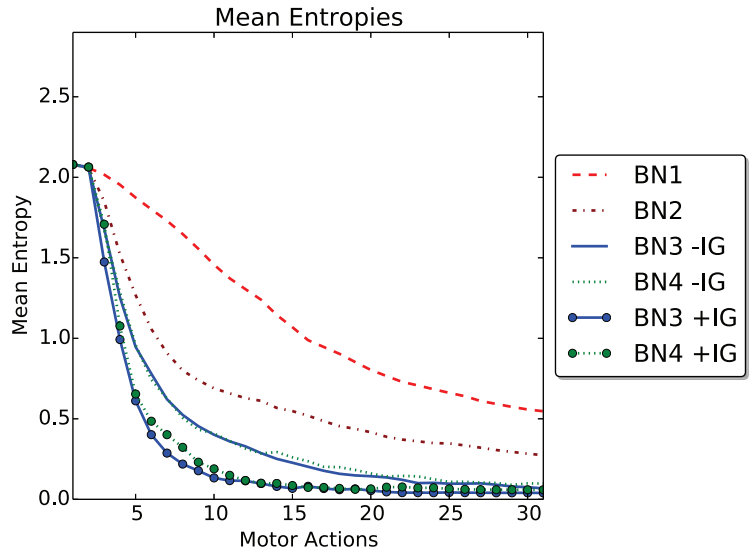
$$IG(m_n) := H(Y) - H(Y|S_{n+1}, m_n), \tag{9}$$

where all probabilities are conditioned by $s_{1:n}, m_{1:n-1}$. This is equivalent to the mutual information of Y and (S_{n+1}, m_n) for an arbitrary m_n. As the current entropy $H(Y)$ is independent of the next movement m_n the most promising motor action m^* can be calculated by minimizing the expected entropy with respect to S_{n+1}, i.e.,

320 D. Nakath et al.

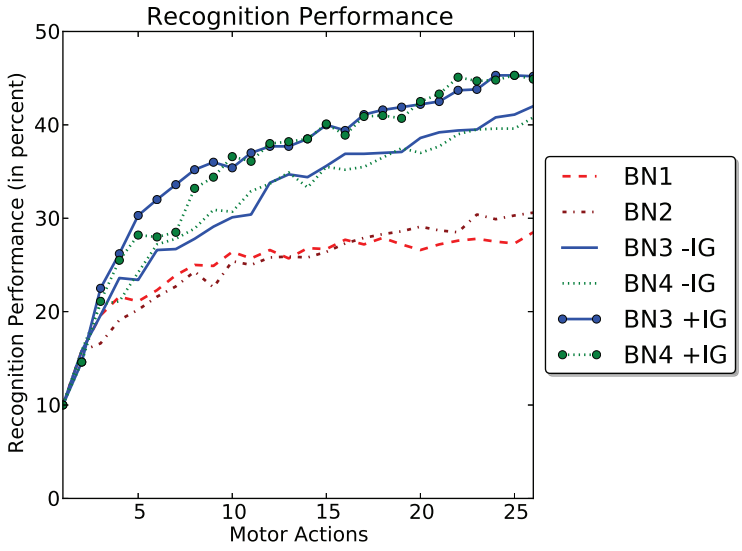


(a) Recognition performance

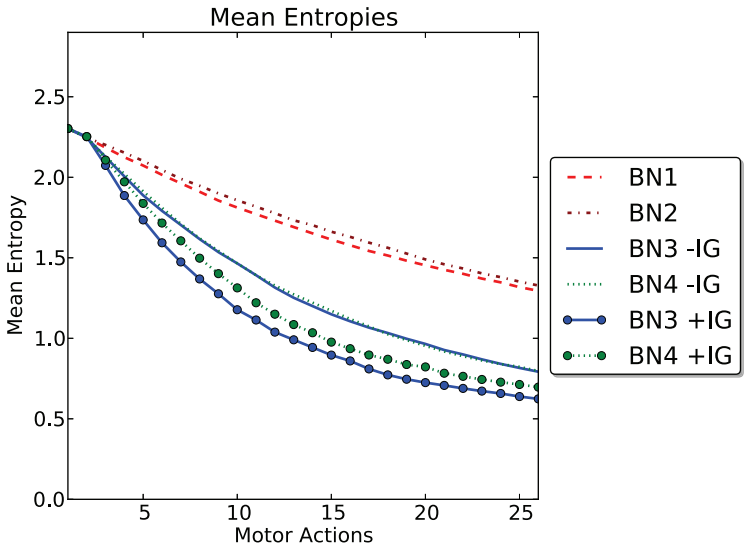


(b) Mean entropies

Fig. 5. Results of the robotic arm evaluation. BN 1, 2, 3 -IG, and 4 -IG executed random movements while BN 3 +IG and BN4 +IG executed information-gain-guided movements (GIST features, 15 clusters, 94 possible relative movements, inhibition of return). Recognition performance shown in (a) and mean entropy of the posterior distribution shown in (b) are both plotted against the number of performed motor actions.



(a) Recognition performance



(b) Mean entropies

Fig. 6. Results of the Caltech 256 subset evaluation. BN 1, 2, 3 -IG, and 4 -IG executed random movements while BN 3 +IG and BN4 +IG executed information-gain-guided movements (SURF features, 15 clusters, 80 possible relative movements, inhibition of return). Recognition performance shown in (a) and mean entropy of the posterior distribution shown in (b) are both plotted against the number of performed motor actions.

$$m^* = \arg\min_{m_n}(\underset{S_{n+1}}{E}[H(Y|s_{1:n}, S_{n+1}, m_{1:n})]). \tag{10}$$

Because the sensory input s_{n+1} is not known prior to executing m_n, the expected value over all possible sensory inputs S_{n+1} is taken into account. Subsequently, the so chosen motor action $m^* \in M$ can be integrated into the sensorimotor feature. The inverse mapping of Q_M can then be used to obtain a top-down motor command $u \in U$, which is then executed by the active sensor.

4 Evaluation

Both active sensor implementations were evaluated on two datasets based on a k-fold cross validation scheme ($k = 10$ for the 3D case and $k = 5$ for the 2D case). The case of 3D movements can be seen as a realistic test for robustness with noisy movements and sensor data. This realistic setting only allows a small dataset, consisting of 8 object classes, each containing 10 objects from 30 different points of view (see Fig. 2a). The case of simulated 2D movements allows for a larger dataset, as movements are simulated on a 5×5-grid on images. Therefore, the Caltech 256 dataset was chosen to serve as a scalability test. Our aim here is not to compete with state of the art recognition approaches but rather to investigate the effects of taking motor information into account while relying on a larger data basis consisting of 10 object classes, each with 100 samples.

Figure 5 depicts the results of the 3D case. The integration of information-gain-guided motor actions in the sensorimotor networks (BN3 +IG, BN4 +IG) proves to be beneficial in terms of recognition performance (see Fig. 5a). The sensor networks (BN1, BN2) perform worse, which holds true for the recognition performance as well as for the mean entropy reduction (see Fig. 5b). To illustrate the effect of the information gain strategy, the sensorimotor networks performing information-gain-guided movements (BN3 +IG, BN4 +IG) and random movements (BN3 -IG, BN4 -IG) were compared to each other. The sensorimotor networks with information-gain-guided movements perform better within the first 15 movements (see Fig. 5a), which is reflected by a steeper reduction in entropy (see Fig. 5b).

In the Caltech 256 evaluation (see Fig. 6) the advantage of using sensorimotor networks with information-gain-guided movements (BN3 +IG, BN4 +IG) for recognition persists over time (see Fig. 6a). This holds true compared to the sensorimotor networks with random movements (BN3 -IG, BN4 -IG) as well as compared to the sensor networks (BN1, BN2). This persisting advantage is also shown by the corresponding evolution of the mean entropies plotted in Fig. 6b.

5 Discussion

We have examined a sensorimotor object recognition system which chooses the next perspective on an object according to the principle of maximum information gain. The underlying sensorimotor representation improved the recognition

performance and enabled the system to optimize its selective serial information intake. It could be shown that the proposed information gain strategy is well suited to control such a selection process.

In this paper, we restricted our focus to the recognition rate and the information gain strategy. However, the criteria for the optimal next step in a selective information intake process may vary in other contexts, e.g., the amount of time or energy required to perform individual actions. The system could be adapted to different contexts on the basis of multicriteria optimization approaches [29].

In principle, our system is able to cope with situations where it only has partial access to information about its environment at a given moment. To overcome this shortcoming, it can act in a sequential fashion to establish the full picture. This is often seen as a contradiction but we could show here that, by integrating sensory and motor information, the underlying sensorimotor contingencies become usable, thus improving the process of sequential information intake controlled by reasonable intermediate actions.

Acknowledgements. This work was supported by DFG, SFB/TR8 Spatial Cognition, Project A5-[ActionSpace].

References

1. Gibson, J.: The ecological approach to visual perception. Houghton Mifflin, Boston (1992)
2. Neisser, U.: Cognition and reality: Principles and implications of cognitive psychology. WH Freeman/Times Books/Henry Holt & Co. (1976)
3. Bajcsy, R.: Active perception. Proceedings of the IEEE 76(8), 966–1005 (1988)
4. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active vision. International Journal of Computer Vision 1(4), 333–356 (1988)
5. Ballard, D.H.: Animate vision. Artificial intelligence 48(1), 57–86 (1991)
6. O'Regan, J.K., Noë, A.: A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences 24(5), 939–972 (2001)
7. Noë, A.: Action in Perception. MIT Press (2004)
8. Prinz, W.: A common coding approach to perception and action. Springer (1990)
9. Hommel, B., Müsseler, J., Aschersleben, G., Prinz, W.: The theory of event coding (TEC): A framework for perception and action planning. Behavioral and Brain Sciences 24(05), 849–878 (2001)
10. O'Regan, J.K.: What it is like to see: A sensorimotor theory of perceptual experience. Synthese 129(1), 79–103 (2001)
11. Noton, D., Stark, L.: Scanpaths in saccadic eye movements while viewing and recognizing patterns. Vision Research 11(9), 929–IN8 (1971)
12. Stark, L.W., Choi, Y.S.: Experimental metaphysics: The scanpath as an epistemological mechanism. In: Zangemeister, W.H., Stiehl, H.S., Freksa, C. (eds.) Visual Attention and Cognition. Advances in Psychology, vol. 116, pp. 3–69. North-Holland (1996)
13. Quinlan, J.R.: Induction of decision trees. Machine Learning 1(1), 81–106 (1986)
14. Cassandra, A.R., Kaelbling, L.P., Kurien, J.A.: Acting under uncertainty: Discrete Bayesian models for mobile-robot navigation. In: Proceedings of the 1996 IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 1996, vol. 2, pp. 963–972. IEEE (1996)

15. Stachniss, C., Grisetti, G., Burgard, W.: Information Gain-based Exploration Using Rao-Blackwellized Particle Filters. In: Robotics: Science and Systems, vol. 2, pp. 65–72 (2005)
16. Oaksford, M., Chater, N.: Information gain explains relevance which explains the selection task. Cognition 57(1), 97–108 (1995)
17. Friston, K., Kilner, J., Harrison, L.: A free energy principle for the brain. Journal of Physiology-Paris 100(1-3), 70–87 (2006); heoretical and Computational Neuroscience: Understanding Brain Functions
18. Schill, K., Pöppel, E., Zetzsche, C.: Completing knowledge by competing hierarchies. In: Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence, pp. 348–352. Morgan Kaufmann Publishers Inc. (1991)
19. Schill, K.: Decision support systems with adaptive reasoning strategies. In: Freksa, C., Jantzen, M., Valk, R. (eds.) Foundations of Computer Science. LNCS, vol. 1337, pp. 417–427. Springer, Heidelberg (1997)
20. Zetzsche, C., Schill, K., Deubel, H., Krieger, G., Umkehrer, E., Beinlich, S.: Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach. In: Proceedings of the 5th International Conference of Simulation of Adaptive Behaviour, vol. 5, pp. 120–126 (1998)
21. Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., Zetzsche, C.: Scene analysis with saccadic eye movements: top-down and bottom-up modeling. Journal of Electronic Imaging 10(1), 152–160 (2001)
22. Zetzsche, C., Wolter, J., Schill, K.: Sensorimotor representation and knowledge-based reasoning for spatial exploration and localisation. Cognitive Processing 9(4), 283–297 (2008)
23. Reineking, T., Wolter, J., Gadzicki, K., Zetzsche, C.: Bio-inspired Architecture for Active Sensorimotor Localization. In: Hölscher, C., Shipley, T.F., Olivetti Belardinelli, M., Bateman, J.A., Newcombe, N.S. (eds.) Spatial Cognition VII. LNCS, vol. 6222, pp. 163–178. Springer, Heidelberg (2010)
24. Schill, K., Zetzsche, C., Hois, J.: A belief-based architecture for scene analysis: From sensorimotor features to knowledge and ontology. Fuzzy Sets and Systems 160(10), 1507–1516 (2009)
25. Griffin, G., Holub, A., Perona, P.: Caltech-256 Object Category Dataset. Technical report, California Institute of Technology (2007)
26. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Prorgess in Brain Research 155, 23–36 (2006)
27. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
28. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer Vision and Image Understanding 110(3), 346–359 (2008)
29. Roy, N., Burgard, W., Fox, D., Thrun, S.: Coastal navigation-mobile robot navigation with uncertainty in dynamic environments. In: Proceedings of the 1999 IEEE International Conference on Robotics and Automation, vol. 1, pp. 35–40. IEEE (1999)

4.4 Article: Affordance-based object recognition using interactions obtained from a utility maximization principle

Reference

D.N., T.K., T.R. and C.Z. designed research; D.N., T.K., and T.R. performed research; D.N and T.K. developed and implemented the system; D.N. analyzed the data; D.N., T.K., T.R., C.Z. and K.S. wrote the paper.

The paper was published in *Computer Vision - ECCV 2014 Workshops* under the following reference [42]:

T. Kluth, D. Nakath, T. Reineking, C. Zetzsche, and K. Schill. Affordance-based object recognition using interactions obtained from a utility maximization principle. In L. Agapito, M. M. Bronstein, and C. Rother, editors, *Computer Vision - ECCV 2014 Workshops*, volume 8926 of *Lecture Notes in Computer Science*, pages 406–412. Springer International Publishing, 2015.

Affordance-Based Object Recognition Using Interactions Obtained from a Utility Maximization Principle

Tobias Kluth( ), David Nakath, Thomas Reineking, Christoph Zetzsche, and Kerstin Schill

Cognitive Neuroinformatics, University of Bremen, Enrique-Schmidt-Straße 5, 28359 Bremen, Germany
tkluth@math.uni-bremen.de

Abstract. The interaction of biological agents within the real world is based on their abilities and the affordances of the environment. By contrast, the classical view of perception considers only sensory features, as do most object recognition models. Only a few models make use of the information provided by the integration of sensory information as well as possible or executed actions. Neither the relations shaping such an integration nor the methods for using this integrated information in appropriate representations are yet entirely clear. We propose a probabilistic model integrating the two information sources in one system. The recognition process is equipped with an utility maximization principle to obtain optimal interactions with the environment

Keywords: Affordance · Sensorimotor object recognition · Information gain

1 Introduction

The ability of humans to reliably recognize objects in the environment is still a largely unsolved problem for artificial systems. Usually, object recognition is understood as a classification problem where a fixed mapping from feature vectors to object classes is learned. This is in line with the classical view of perception as the input from world to mind and of action as the output from mind to world [6], which implies a dissociation between the capacities for perception and action. However, there is strong evidence that object recognition cannot be understood independently of the interaction of an agent with its environment [8]. "Active perception" approaches [1,2] take this partially into account, but actions are not merely means for acquiring new information, they also provide evidence themselves for the recognition [5]. What an agent perceives is thus also determined by what it does or what it is able to do [8].

Research in the direction of affordances by Gibson [3] also provides evidence that affordances are key ingredients of the perceptual process. A variety of studies regarding object recognition show that the visual information of a manipulable

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 406–412, 2015.
DOI: 10.1007/978-3-319-16181-5_29

object causes an activation of representations of actions which can typically be executed on the object [4]. The advantageous interplay between sensory and action information, which was also recognized by Neisser [7], should be considered in the recognition process.

The strong interrelation between motor actions and sensory perceptions is basis for the concept of a sensorimotor representation [8,10]. Similarly to the affordance point of view the processing stages for sensory and motor information are not separated. The approach including the actions in the representation gives the opportunity to choose the next action such that a specific objective is pursued. Generally, the problem of action selection can be solved in numerous ways, but as information gathering should be one major purpose of motor actions it is appropriate to consider an information-theoretic utility function. Prior research in this area often found that the principle of *information gain* is well suited to select an appropriate next action.

In this paper, we propose a system for object recognition which incorporates both the information gain principle from sensorimotor systems and the theoretical concept of affordances. Building upon the investigations in [11], we developed a sensomotoric probabilistic reasoning system for affordance-based object recognition. The design of our architecture is motivated by two main goals: i) to provide a clear relation to Bayesian inference approaches, and ii) to enable a comparison between the classic sensory approach and the sensorimotor, affordance-oriented approach within one common probabilistic framework.

2 Object Recognition System

The system described in the following is a generic architecture (see Fig. 1). The recognition loop starts out with a particular pose of an object which is perceived by a sensor. The sensor passes its raw data to the sensory processing module. After processing, the sensory data becomes part of a new sensorimotor feature, which is then fed into the probabilistic reasoning module. The processed sensory data are also used to obtain a set of possible interactions, i.e., the affordances offered by the sensory data related to the abilities of the agent. The Bayesian inference module calculates the new posterior distribution based on a previously-learned sensorimotor representation. This representation contains the learned perceptual consequences of an interaction in a given state for every object class. The posterior distribution constitutes the current belief of the system. This belief is used by the information gain strategy to choose an optimal next action from the set of possible interactions. The selected interaction then also becomes part of the sensorimotor feature and is subsequently executed by the agent. The whole process results in a new state, which in turn delivers new raw sensory data to enter the next cycle of the recognition loop.

More formally speaking, the system depends on an *agent*, which can be controlled such that it perceives information about a specific aspect of the world. In Fig. 1, the two arrows pointing from the states to the sensory processing module correspond to the mapping $A : U \times X \rightarrow R$, where U is the space of all

408 T. Kluth et al.

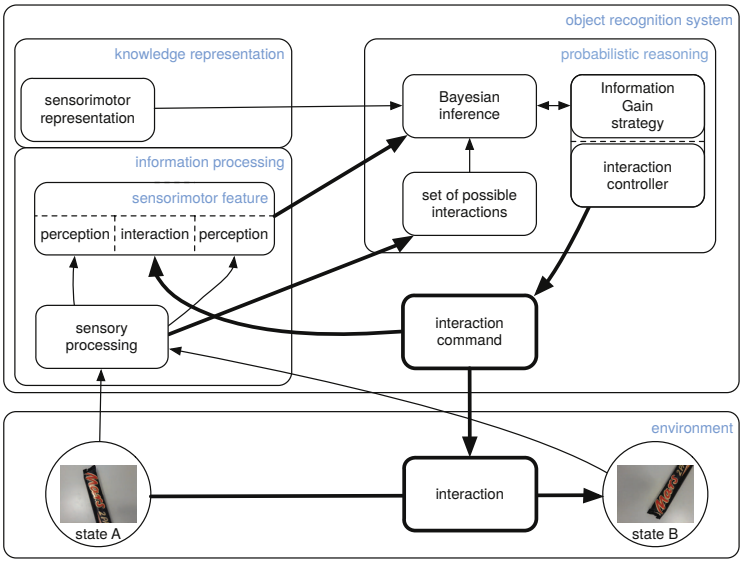


Fig. 1. Architecture of the object recognition system

interactions which are currently possible, X is the state space, and R is the raw sensor data space.

The system has no explicit knowledge about the actual state, and the currently possible interactions U. The possible interactions are of course dependent on the state but nevertheless both information must be obtained from the sensor data. The sensoric dependency on the state is formalized by the mapping $U : X \to \mathcal{P}(\Omega_U)$, where Ω_U is the set of all possible interactions and $\mathcal{P}$ denotes the power set. Note that U comprises the link from the state to the sensory processing module and the following link to the set of possible interactions in Fig. 1, i.e., the perceived affordances. Assuming that the output of the function U is given, we write U instead of $U(x),\ x \in X$, for convenience. Considering the state-agnostic behavior, the influence of the agent can be formally redefined to $A_x : U \to R$ with $A_x(u) := A(x, u) = r,\ x \in X,\ u \in U(x),\ r \in R$. The only time-dependent variables are the state x and the interaction u.

The raw sensor data $r \in R$ is fed into the *sensory processing* (SP) which mainly extracts the relevant features belonging to a feature space F, i.e., $SP : R \to F$. Subsequently, the quantization operation $Q_S : F \to S$ maps the features to a specific feature class in the finite space S. The possible interactions are mapped with $Q_M : \Omega_U \to M$ to the finite set of interactions M to yield a manageable product space of sensory information and actions. The results of these quantizations then become part of a sensorimotor feature (SMF). The single quantizations are represented in Fig. 1 by the arrows from the sensory processing module and the interaction command to the sensorimotor feature which is defined as the triple

$$SMF_i := (s_{i-1}, m_{i-1}, s_i), \tag{1}$$

where $m_{i-1} := Q_M(u_{i-1})$ is the interaction between the sensor information s_{i-1} and s_i at time step t_{i-1} and t_i. The whole chain of operations to obtain the sensor information at a time step t_i can be described by $s_i := (Q_S \circ SP \circ A_x)(u_{i-1})$.

The *knowledge representation* is comprised of the learned sensorimotor representation (SMR), which is a full joint probability distribution of SMFs and the classes represented by the discrete random variable Y. Every possible SMF is generated on a set of known objects in a training phase. This means that, from every possible state x, the sensory consequence of every possible action u is perceived, resulting in

$$SMR := P(SMF_i, Y) = P(S_{i-1}, M_{i-1}, S_i, Y). \tag{2}$$

The *probabilistic reasoning* module consists of a Bayesian inference approach accompanied by an information gain strategy. They rely on bottom-up sensory data and top-down information from the knowledge representation. The information gain strategy can choose an optimal next interaction for the agent, thus improving the input of the following Bayesian inference step.

3 Model Implementation and Outlook

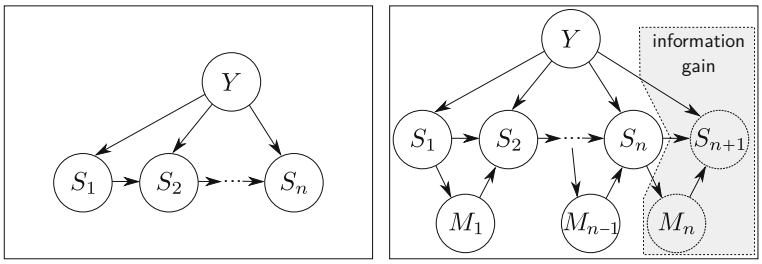
Based on the schematic outline presented above, we applied our system to object recognition using a robotic arm interacting with objects in 3D space. We used a discrete set of interactions M of a robotic arm with an object which comprise the relative position/pose of the visual sensor to the object ($\Omega_U = M$, $Q_M = Id$).

In the learning phase, features are extracted from the training data (images from every reachable state). GIST-features [9] are used to describe the sensory input, i.e., defining SP. The quantization Q_S is then learned by performing a k-means clustering on the extracted features. In order to build the individual SMFs, features are extracted and the results are assigned to clusters with the aid of the previously defined mapping Q_S. These labels are combined with the corresponding interactions in a set of SMFs. Finally, all generated SMFs are stored in a Laplace-smoothed SMR.

The probabilistic reasoning is comprised of a Bayesian inference module in the form of a dynamic Bayesian network (BN) and a corresponding information gain strategy. Two of these probabilistic reasoning modules were implemented to examine the difference between *sensor networks*, which only take into account sensory information (which also implies that no information gain strategy is used), and *affordance-based networks*, which integrate sensory perceptions and interactions. The object recognition in the sense of computer vision then takes place by classification which is performed by choosing the class with the maximum posterior probability.

The representative of the *sensor networks* is Bayesian network 1 (BN1) (see Fig. 2a), which resembles an extended naive Bayes model that additionally allows

410 T. Kluth et al.



(a) BN1 **(b)** BN2

Fig. 2. In Bayesian network BN1 (a) sensory information S_n is processed only to obtain the object class Y. Bayesian network BN2 (b) is equipped with the information gain startegy which takes also the action M_n into account.

for statistical dependencies between the preceding and the current sensor information, s_{i-1} and s_i, resulting in

$$P(y|s_{1:n}) \propto P(y)P(s_1|y)\prod_{i=2}^{n} P(s_i|s_{i-1}, y), \tag{3}$$

where $s_{1:n}$ is a short notation for the n-tuple $(s_1, \ldots, s_n)$.

Bayesian network 2 (BN2) (see Fig. 2b) uses the full information of the SMF and therefore belongs to the *affordance-based networks*. The assumption that the current sensory input s_i depends on the action m_{i-1} integrates sensory perceptions and actions in the recognition process and permits the application of the information gain strategy to choose the next optimal interaction. Additionally, it is assumed that the action m_{i-1} statistically depends on the preceding sensory input s_{i-1}. The inference can then be conducted by

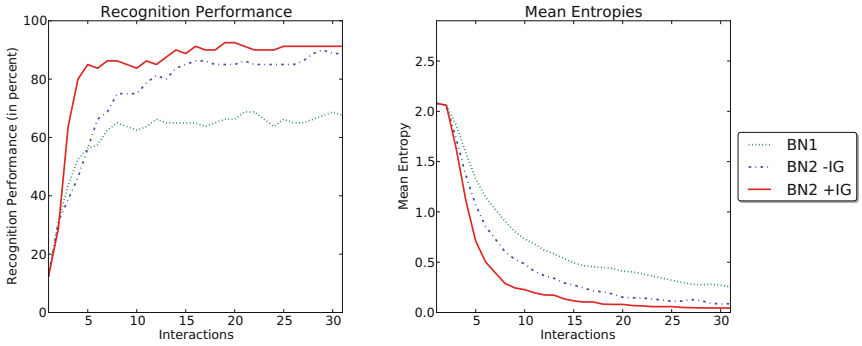
$$P(y|s_{1:n}, m_{1:n-1}) \propto P(y)P(s_1|y)\prod_{i=2}^{n} P(s_i|s_{i-1}, m_{i-1}, y)P(m_{i-1}|s_{i-1}). \tag{4}$$

The strategy for action selection should satisfy two main properties: (i) The strategy should adapt itself to the current belief state of the system and (ii) the strategy should not make decisions in an heuristic fashion but tightly integrated into the axiomatic framework used for reasoning. The information gain strategy presented in this paper complies with both of these properties.

The information gain IG of a possible next action m_n is defined as the difference between the current entropy and the conditional entropy,

$$IG(m_n) := H(Y|s_{1:n}, m_{1:n-1}) - H(Y|S_{n+1}, m_n, s_{1:n}, m_{1:n-1}). \tag{5}$$

This is equivalent to the mutual information of Y and (S_{n+1}, m_n) for an arbitrary m_n. As the current entropy $H(Y|s_{1:n}, m_{1:n-1})$ is independent of the next action



(a) Recognition performance **(b)** Mean entropies

Fig. 3. Results of the robotic arm evaluation (8 object classes, 10 objects per class, 30 discrete viewpoints). BN 1 and 2 -IG executed random movements while BN2 +IG executed information-gain-guided movements.

m_n the most promising action m^* can be calculated by minimizing the expected entropy with respect to S_{n+1},

$$m_n^* = \arg\min_{m_n}(\underset{S_{n+1}}{E}[H(Y|s_{1:n}, S_{n+1}, m_{1:n})]). \tag{6}$$

Because the sensory input s_{n+1} is not known prior to executing m_n, the expected value over all possible sensory inputs s_{n+1} is taken into account. The selected action $m^* \in M$ is integrated into the next sensorimotor feature. The inverse mapping of Q_M can then be used to obtain a top-down interaction command $u \in U$, which is then executed by the agent.

Preliminary results are shown in Fig. 3. In the future, we plan to conduct a more extensive evaluation of our approach (using different sensory features) by comparing it to established object recognition approaches on a larger data set. Furthermore we want to extend our approach by a saliency feature detector.

Acknowledgments. This work was supported by DFG, SFB/TR8 Spatial Cognition, project A5-[ActionSpace], and DLR projects "EnEx" and "KaNaRiA".

References

1. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active vision. International Journal of Computer Vision **1**(4), 333–356 (1988)
2. Bajcsy, R.: Active perception. Proceedings of the IEEE **76**(8), 966–1005 (1988)
3. Gibson, J.: The ecological approach to visual perception. Houghton Mifflin, Boston (1992)
4. Grèzes, J., Decety, J.: Does visual perception of object afford action? Evidence from a Neuroimaging study. Neuropsychologia **40**(2), 212–222 (2002)

5. Helbig, H.B., Graf, M., Kiefer, M.: The role of action representations in visual object recognition. Experimental Brain Research **174**(2), 221–228 (2006)
6. Hurley, S.L.: Consciousness in action. Harvard University Press (2002)
7. Neisser, U.: Cognition and reality: Principles and implications of cognitive psychology. WH Freeman/Times Books/Henry Holt & Co. (1976)
8. Noë, A.: Action in Perception. MIT Press (2004)
9. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Progress in Brain Research **155**, 23–36 (2006)
10. O'Regan, J.K., Noë, A.: A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences **24**(5), 939–972 (2001)
11. Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., Zetzsche, C.: Scene analysis with saccadic eye movements: top-down and bottom-up modeling. Journal of Electronic Imaging **10**(1), 152–160 (2001)

4.5 Article: Adaptive information selection in images: Efficient naive bayes nearest neighbor classification

Reference

T.R., T.K., and D.N. designed research; T.R., T.K. and D.N. performed research; T.R. and T.K. implemented the work; T.R. analyzed the data; T.R., T.K., and D.N. wrote the paper.

The paper is accepted at the *16th International Conference on Computer Analysis of Images and Patterns* and has the following title:

T. Reineking, T. Kluth, and D. Nakath. Adaptive information selection in images: Efficient naive bayes nearest neighbor classification.

Adaptive Information Selection in Images: Efficient Naive Bayes Nearest Neighbor Classification

Thomas Reineking, Tobias Kluth and David Nakath

Cognitive Neuroinformatics, University of Bremen,
Enrique-Schmidt-Str. 5, 28359 Bremen, Germany
{trking,tkluth,dnakath}@cs.uni-bremen.de

Abstract. We propose different methods for adaptively selecting information in images during object recognition. In contrast to standard feature selection, we consider this problem in a Bayesian framework where features are sequentially selected based on the current belief distribution over object classes. We define three different selection criteria and provide efficient Monte Carlo algorithms for the selection. In particular, we extend the successful Naive Bayes Nearest Neighbor (NBNN) classification approach, which is very costly to compute in its original form. We show that the proposed information selection methods result in a significant speed-up because only a small number of features needs to be extracted for accurate classification. In addition to adaptive methods based on the current belief distribution, we also consider image-based selection methods and we evaluate the performance of the different methods on a standard object recognition data set.

Keywords: object recognition, classification, information selection, Bayesian inference, information gain

1 Introduction

Selecting relevant information from a high-dimensional input is a fundamental problem pertaining many different areas ranging from computer vision to robotics. An effective selection strategy uses only a small subset of the available information without negatively impacting the task performance. An example of a successful selection strategy is the processing of visual information in humans where eye movements are performed in order to extract the relevant information from a scene in a very efficient manner [11]. A key feature of this selection is its adaptivity because the selection is strongly influenced by the current belief about the scene [17].

In this paper, we follow the idea of an adaptive belief-based information selection and we investigate it in the context of object recognition. While object recognition is usually viewed as a static pattern recognition problem, we model the recognition as an information gathering process unfolding in time, which is more akin to visual processing in humans. In this case, recognition becomes a

problem of Bayesian information fusion where the selection of relevant information is done adaptively with regard to the current belief distribution (in contrast to classical feature selection methods, e.g. [4,7]). We propose different criteria for optimal information selection and provide efficient algorithms for their application. In addition to belief-based selection methods, we also consider an image-based method that uses a saliency operator to identify relevant locations in an image.

We combine the information selection methods with the successful NBNN object recognition approach presented in [1]. We use NBNN because it is a probabilistic approach where local image features are sequentially processed in order to update a belief distribution over possible object classes.[1] For each extracted feature, multiple expensive nearest neighbor searches have to be performed, which is why selecting a small subset of relevant features greatly reduces the computational costs of NBNN classification (for making the nearest neighbor search itself more efficient, see [9]). Note that while we focus on object recognition in this paper, the proposed belief-based information selection methods are very versatile and could therefore also be applied in other contexts.

The paper is structured as follows. In the next section, the basics of the NBNN approach are introduced. In Sect. 3, the information selection methods are described in detail. In Sect. 4, the different selection methods are combined with the NBNN approach and compared empirically on a standard object recognition data set. The paper concludes with a short discussion of the proposed methods and possible extensions.

2 Naive Bayes Nearest Neighbor

For NBNN, a set of local image descriptors is extracted from the query image (e.g. SIFT descriptors [8]) which is then used to compute the posterior probability distribution over object classes. Let $\mathcal{C}$ denote the set of object classes, and let $d_{1:N}$ denote all descriptors extracted from the query image[2] where N is the total number of descriptors found in the image. By applying Bayes' rule and by making a naive Bayes assumption regarding the conditional independence of descriptors, the posterior is given by

$$P(c|d_{1:N}) \propto P(c) \prod_{i=1}^{N} p(d_i|c) \text{ with } c \in \mathcal{C}. \tag{1}$$

The likelihood $p(d_i|c)$ for the i-th descriptor is approximated using kernel density estimation (KDE). This avoids the severe errors caused by quantizing descriptors like in bag-of-words models [2]. To reduce computational complexity and in contrast to typical KDE, only the nearest neighbor (NN) of d_i in the training set is considered because the density contributions of descriptors that

[1] Other state-of-the-art classification approaches like deep networks [6] are not suited here because they do not allow for an incremental processing of features.

[2] We use the shorthand notation $d_{1:N} = d_1, \ldots, d_N$.

are farther away tend to be negligible. Using a Gaussian kernel, the likelihood is approximated by

$$p(d_i|c) = \frac{1}{|\mathcal{D}_c|} \sum_{d^{(j)} \in \mathcal{D}_c} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{||d_i - d^{(j)}||^2}{2\sigma^2}) \tag{2}$$

$$\approx \frac{1}{\sqrt{2\pi}\sigma|\mathcal{D}_c|} \exp(-\frac{||d_i - NN_c(d_i)||^2}{2\sigma^2}) \tag{3}$$

with

$$NN_c(d_i) = \arg\min_{d^{(j)} \in \mathcal{D}_c} ||d_i - d^{(j)}|| \tag{4}$$

where σ denotes the (class-independent) KDE bandwidth, $\mathcal{D}_c$ denotes the set of descriptors in the training set belonging to class c, and $NN_c(d_i)$ denotes the NN of d_i in $\mathcal{D}_c$. The posterior is thus given by

$$P(c|d_{1:N}) \propto P(c) \prod_{i=1}^{N} p(d_i|c) \propto P(c) \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} ||d_i - NN_c(d_i)||^2\right). \tag{5}$$

Note that we ignore the descriptor count $|\mathcal{D}_c|$ for the posterior because its influence is very limited and it simplifies the derivations below. Assuming a uniform class prior, the most probable class c^* can be found using the simple decision rule

$$c^* = \arg\max_{c \in \mathcal{C}} \log P(c|d_{1:N}) = \arg\min_{c \in \mathcal{C}} \sum_{i=1}^{N} ||d_i - NN_c(d_i)||^2. \tag{6}$$

Though the decision rule in Eq. (6) is independent of σ (it is therefore ignored in the original NBNN approach), the bandwidth turns out to be relevant for the selection of optimal descriptors in the next section. We determine the optimal bandwidth σ^* by maximizing the log-likelihood of all training set descriptors $\mathcal{D} = \cup_{c \in \mathcal{C}} \mathcal{D}_c$ according to

$$\sigma^* = \arg\max_{\sigma} \log p(\mathcal{D}|\sigma) = \sqrt{\frac{\sum_{c \in \mathcal{C}} \sum_{d^{(i)} \in \mathcal{D}_c} ||d^{(i)} - NN_c(d^{(i)})||^2}{|\mathcal{D}|}}. \tag{7}$$

3 Information Selection

For selecting the most relevant descriptors, we distinguish between belief-based selection methods and image-based ones. For belief-based selection, the probabilistic model introduced in the previous section is used to predict the effect of extracting a descriptor at a particular location in the image on the current belief distribution. In contrast, for image-based selection, the image information itself is used to determine which regions in the image are most relevant without considering the training data.

We model the information selection problem as one of finding the most promising *absolute* location in an image where the object is assumed to be depicted at the center of the image. This simplification allows us to ignore the problem of object detection, which would be necessary in case of more complex scenes with variable object locations. Let l_t denote the location of a descriptor d_{l_t} in an image at the t-th extraction step after already having extracted the first $t-1$ descriptors $d_{l_1:l_{t-1}}$. To select the next optimal location, we compute a score $S(l_t)$ for each location and choose the maximum

$$l_t^* = \arg\max_{l_t \in \mathcal{L}_t} S(l_t). \tag{8}$$

To limit the number of locations, we put a grid over each image where a location represents a grid cell. Because of the naive Bayes assumption, the likelihoods of the descriptors within a cell can simply be combined by multiplying them, i.e., each likelihood $p(d_{l_t}|c)$ represents a product of the likelihoods of individual descriptors located within the same grid cell.

In the remainder of this section, we first present two belief-based information selection methods and then an image-based one.

3.1 Maximum Expected Probability

For classification it is useful to select the descriptor that maximizes the expected posterior probability (MEP) of the true class. Because the value of the next descriptor is unknown prior to extracting it, it has to be modeled as a random variable D_{l_t}. The same applies to the value of the true object class of the query image, which is modeled as a random variable $C_{\text{true}} \in \mathcal{C}$. The score S_{MEP} is the conditional expectation of the true class posterior probability

$$\begin{aligned} S_{\text{MEP}}(l_t) &= E[P(C_{\text{true}}|d_{l_1:l_{t-1}}, D_{l_t})|d_{l_1:l_{t-1}}] && (9)\\ &= \int \sum_{c_{\text{true}} \in \mathcal{C}} p(c_{\text{true}}, d_{l_t}|d_{l_1:l_{t-1}})\, P(c_{\text{true}}|d_{l_1:l_t})\, dd_{l_t} && (10)\\ &= \int \sum_{c_{\text{true}} \in \mathcal{C}} p(c_{\text{true}}, d_{l_t})\, \frac{P(c_{\text{true}}|d_{l_1:l_{t-1}})}{P(c_{\text{true}})}\, P(c_{\text{true}}|d_{l_1:l_t})\, dd_{l_t} && (11)\\ &\approx \frac{1}{M} \sum_{i=1}^{M} \frac{P(c^{(i)}|d_{l_1:l_{t-1}})}{P(c^{(i)})}\, P(c^{(i)}|d_{l_1:l_{t-1}}, d_{l_t}^{(i)}) && (12) \end{aligned}$$

with respect to C_{true} and D_{l_t} given the previous descriptors $d_{l_1:l_{t-1}}$. Because the training samples are assumed to represent i.i.d. samples from the joint distribution $p(c_{\text{true}}, d_{l_t})$, the score can be approximated by a Monte Carlo estimate computed over the training set in Eq. (12) where $c^{(i)}$ denotes the class of the i-th image in the training set, $d_{l_t}^{(i)}$ denotes the descriptor in the i-th training image at location l_t, and M denotes the total number of images in the training set. All the posterior probabilities can be obtained using Eq. (5).

Computing the Monte Carlo estimate can be time-consuming because all descriptors in the training set have to be considered. However, the NN distances required for the likelihoods can be computed in advance so that the overall score computation is still significantly faster than having to process all descriptors from the query image. In addition, it would be possible to only use a subset of the training samples where each sample would be drawn with a probability given by the current belief distribution.

For the special case where no descriptors have been extracted ($t = 1$) or where one chooses to ignore previously extracted descriptors, we can compute a score that ignores the current belief distribution and only maximizes the normalized expected likelihood (MEL). Plugging in $P(c^{(i)})$ for the current belief distribution in Eq. (12) results in

$$S_{\mathrm{MEL}}(l_t) = E[P(C_{\mathrm{true}}|D_{l_t})] \tag{13}$$

$$\approx \frac{1}{M}\sum_{i=1}^{M}\frac{P(c^{(i)})}{P(c^{(i)})}\,P(c^{(i)}|d_{l_t}^{(i)}) \tag{14}$$

$$= \frac{1}{M}\sum_{i=1}^{M}\frac{P(c^{(i)})\,p(d_{l_t}^{(i)}|c^{(i)})}{\sum_{c\in\mathcal{C}}P(c)\,p(d_{l_t}^{(i)}|c)}. \tag{15}$$

Because this score is independent of previous descriptors, it can be computed offline and is thus extremely fast.

3.2 Maximum Expected Information Gain

A popular method for feature selection is the maximum expected information gain (MIG) [18]. Here we consider a "dynamic" information gain version that takes previous descriptors into account during the recognition process [12,15]. It is given by the expected uncertainty/entropy reduction resulting from observing a new descriptor d_{l_t}. The information gain score S_{MIG} is the conditional expectation of this reduction with respect to D_{l_t} given the previous descriptors $d_{l_1:l_{t-1}}$:

$$S_{\mathrm{MIG}}(l_t) = H(C|d_{l_1:l_{t-1}}) - E[H(C|d_{l_1:l_{t-1}}, D_{l_t})|d_{l_1:l_{t-1}}] \tag{16}$$

$$= H(C|d_{l_1:l_{t-1}}) - \int\sum_{c_{\mathrm{true}}\in\mathcal{C}} p(c_{\mathrm{true}}, d_{l_t}|d_{l_1:l_{t-1}})H(C|d_{l_1:l_t})\,dd_{l_t} \tag{17}$$

$$\approx H(C|d_{l_1:l_{t-1}}) - \frac{1}{M}\sum_{i=1}^{M}\frac{P(c^{(i)}|d_{l_1:l_{t-1}})}{P(c^{(i)})}\,H(C|d_{l_1:l_{t-1}}, d_{l_t}^{(i)}) \tag{18}$$

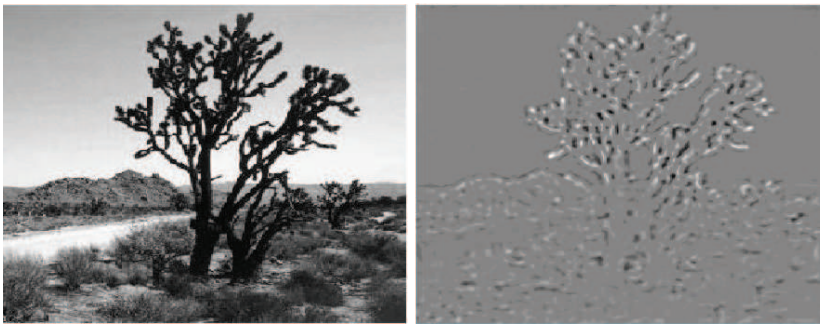
with entropy

$$H(X) = -\sum_{x\in\mathcal{X}} P(x)\log P(x). \tag{19}$$

Like for S_{MEP}, the expected value is approximated by a Monte Carlo estimate using samples from the training set in Eq. (18). Note that the information gain is

independent of the true class, meaning that a high MIG score only requires the resulting posterior distribution to be "non-uniform", thus completely ignoring how probable the true class is.

3.3 Intrinsically Two-Dimensional Signals



(a) Original (b) $I2D$-saliency

Fig. 1: Extracted $I2D$-saliency (b) of the image shown in (a). The extracted $I2D$-score is the clipped eigenvalue computed with the following parameters: $n = 6$, $\sigma_r = 0.2$. Positive elliptically curved regions are light and negative elliptically curved regions are dark.

The following image-based selection method uses a saliency operator which detects intrinsically two-dimensional ($I2D$) signals [19]. The intrinsic dimensionality of a signal $u(x, y)$ is defined as $I0D$ for all signals that are constant and as $I1D$ for all signals that can be written as a function of one variable in an appropriately rotated coordinate system (e.g. an image of an oriented straight edge). In contrast, $I2D$-signals make full use of the two degrees of freedom (e.g. an image of a corner or crossing lines). The $I2D$-saliency also appears to play an important role in the control of saccadic eye movements [5,16] which motivates its use as a score function within the context of this work. In order to identify the interesting $I2D$-points, we make use of the generalized curvature operator introduced in [19]: The generalized curvature operator $T_n : C^2(\Omega) \to C(\Omega)$ with compact $\Omega \subset \mathbb{R}^2$ is defined for $n \in \mathbb{N}$ by

$$T_n(u)(x) = \frac{1}{4}\left((\Delta u)^2 - \epsilon_n(u)^2\right) = \frac{1}{4}\underbrace{(\Delta u + |\epsilon_n(u)|)}_{=\lambda_1(u)}\underbrace{(\Delta u - |\epsilon_n(u)|)}_{=\lambda_2(u)} \tag{20}$$

with eccentricity $\epsilon_n(u)^2 = (c_n * u)^2 + (s_n * u)^2$. The convolution kernels c_n and s_n are defined by their Fourier transform in polar coordinates ($x_1 = r\cos(\phi)$,

$x_2 = r\sin(\phi)$) by

$$\mathcal{F}(c_n)(r,\phi) = (i)^n f(r)\cos(n\phi)$$
$$\text{and } \mathcal{F}(s_n)(r,\phi) = (i)^n f(r)\sin(n\phi).$$

f is a continuous function of the radius r given by $f(r) = 2\pi r^2 e^{\frac{1}{2}\frac{r^2}{\sigma_r^2}}$. λ_1 and λ_2 are the eigenvalues of the Hessian matrix of u in the case of $n = 2$ where the generalized curvature becomes the Gaussian curvature. The Gaussian curvature allows a distinction between elliptic, hyperbolic, and parabolic regions on the curved surface $\{(x, y, u(x, y))^T | (x, y)^T \in \mathbb{R}^2\}$. Using the eigenvalues, the clipped eigenvalue is defined by

$$CE(u) = |\min(0, \lambda_1(u))| - |\max(0, \lambda_2(u))|. \tag{21}$$

In contrast to directly using generalized curvature as a score function, the advantage of the clipped eigenvalue is that it can distinguish between positive elliptic and negative elliptic points, i.e., both eigenvalues are positive or negative. Furthermore, the clipped eigenvalue does not respond to hyperbolic regions. The latter is useful because hyperbolic regions are often found right next to elliptic ones, in which case the hyperbolic regions would only provide redundant information. The score function is then defined with respect to the luminance function u of the grid cell $\Omega(l_t)$ at location l_t by

$$S_{I2D}(l_t) = \frac{1}{|\Omega(l_t)|}\int_{\Omega(l_t)} |CE(u)(x)|\, dx. \tag{22}$$

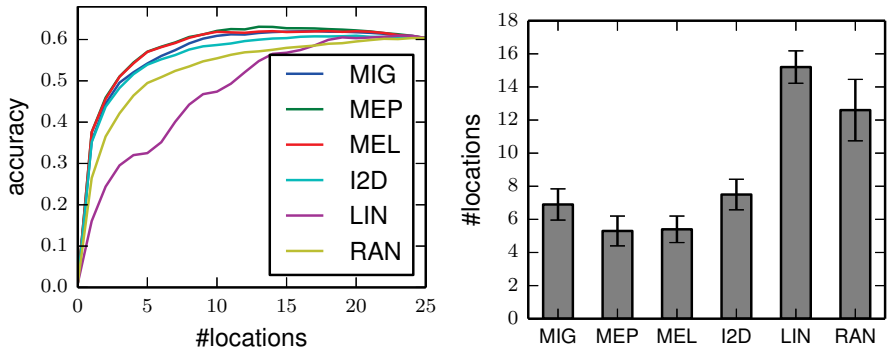
In contrast to belief-based score functions, the $I2D$-saliency is a purely image-based method. Consequently, it does not require any training data. The $I2D$-score function of an example image is illustrated in Fig. 1.

4 Evaluation

We evaluate the proposed information selection methods on the Caltech 101 data set [3]. We use 15 randomly selected images from each of the 101 object classes for training and 10 for testing. All images are scaled such that they have a maximum width or height of 300 pixels. Afterwards, densely-sampled SIFT descriptors are extracted (several thousands for each image depending on the size) and the NN distances are computed.[3]

Fig. 2 shows the mean accuracy over time for the different selection methods using a 5×5 grid and 10-fold cross validation. The MEP and MEL methods result in the quickest increase in accuracy and only require extracting descriptors from less than 6 grid cells on average for reliable classification (even though the MEL method ignores the current belief distribution). The MIG and I2D methods perform only slightly worse and all of the considered methods significantly

[3] We use the code provided at `https://github.com/sanchom/sjm` for SIFT descriptor extraction and the FLANN library [10] for fast NN matches.



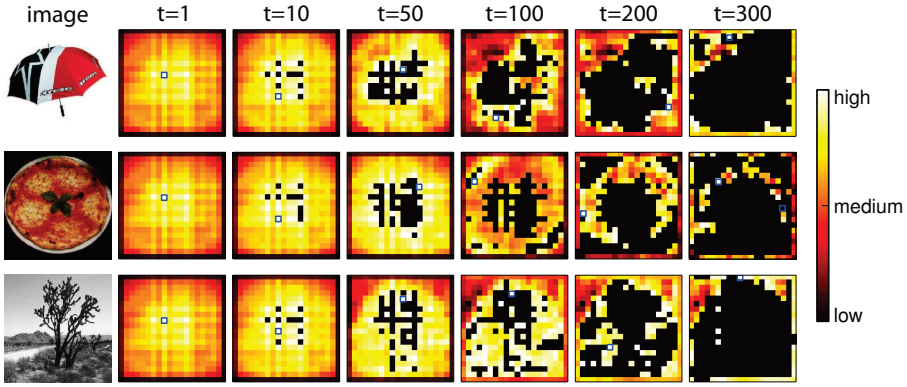
(a) accuracy (b) location count

Fig. 2: (a) Mean accuracy on the entire Caltech data set plotted for different time steps/location counts using different selection methods. (b) Mean number of time steps/location counts required for reaching at least 90% of the final accuracy where all descriptors have been extracted. The indicated standard deviation is computed with respect to the different folds.

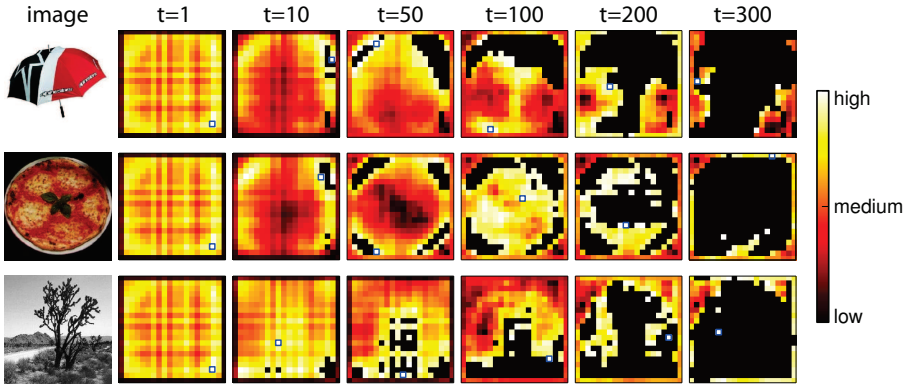
outperform the baseline methods where descriptors are either selected randomly ("RAN") or line by line starting at the top of the image ("LIN"). The final accuracy after having extracted *all* descriptors is identical for each method because the extraction order is irrelevant for the classification model. Interestingly, the accuracy is highest after having extracted about half of all descriptors (except for the baseline methods), showing that the remaining descriptors tend to only decrease the recognition performance.

To illustrate the process of sequentially selecting descriptors, Fig. 3 shows score distributions over time using a 20×20 grid for three example images. For the belief-based MEP and MIG selection methods shown in (a) and (b), the score distributions change significantly over time and adapt themselves to the query image based on the current belief distribution. The I2D score distribution remains constant over time aside from setting the score of previously selected locations to 0 (the apparent change in other locations is due to scaling in the visualization). At $t = 1$, both the MEP and the MIG scores are independent of the query image and only the I2D method uses the image information. Over time, the MEP and MIG scores adapt themselves to the current belief distribution over object classes, whereas the I2D score remains unchanged. The visible "grid pattern" (especially for $t \leq 10$) is an artifact resulting from some grid cells containing more descriptors than others (this could be avoided if all cells contained roughly the same number of descriptors).

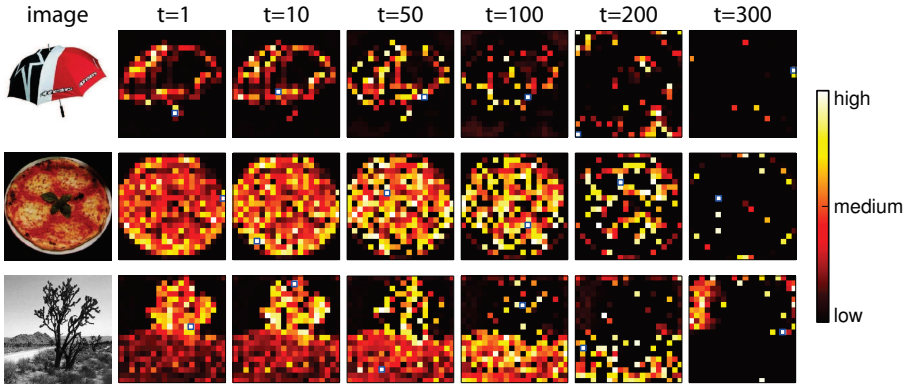
Perhaps surprisingly, the MEP score is highest at the center while the MIG score is initially highest in the periphery. One possible explanation for this effect is that the MEP method can be interpreted as a "confirmation strategy" whereas the MIG method can be interpreted as a "discriminative strategy". For MEP,



(a) MEP



(b) MIG



(c) I2D

Fig. 3: Examples of score distributions over time using a 20×20 grid for different selection methods and query images. The small blue square indicates the cell with the highest score from which the next descriptor(s) are extracted. Cells that have already been selected have a score value of 0 (black).

extracting descriptors from the center of an object usually increases the probability of the true class without necessarily resulting in a unique classification (i.e. the overall belief distribution can still be very uniform). In contrast, the MIG method is agnostic with respect to the true class and only seeks to reduce uncertainty (e.g. by ruling out large numbers of classes). This could be accomplished by analyzing the "context" of objects, which is why the MIG method might first focus on the background.

5 Conclusion

We have proposed different methods for adaptive information selection from images where the current belief distribution directly determines which image locations should be considered next. In addition, we have also considered an image-based selection method that does not require any training data. Using these methods, we have extended the NBNN approach and we have shown that the selection methods make it possible to only consider a small subset of the available information while maintaining the original recognition performance. In particular for NBNN, where computing the NN distances for each descriptor is very time-consuming, the result is a significantly reduced computation time.

One of the problems not addressed in this paper is the fact that features in close proximity to each other are highly correlated. While the naive Bayes assumption can be justified for inference by the greatly reduced computational complexity, for the information selection it would be possible to use a more sophisticated model where correlations are explicitly considered. As a result, there would be a penalty for extracting features located very closely to each other, thus avoiding processing of redundant information.

In this paper, we have considered belief-based selection strategies (MEP, MIG) and image-based strategies (I2D) separately. A more promising approach could be a combination of both strategies [16] because the belief-based strategy completely ignores what is readily available in the image while a purely image-based strategy has difficulties selecting the relevant information because it ignores the training data. Due to the complementary nature of these strategies, a hybrid strategy could further improve the selection process.

We believe that the proposed selection methods can also be useful for problems beyond recognizing single objects. Especially for complex scenes containing many objects, an adaptive information selection strategy could predict the likely locations of objects and thereby facilitate understanding of the entire scene.

Finally, the general nature of the proposed information selection approaches allows for the application to systems which must perform actions to obtain new information from their environments (e.g. an autonomous spacecraft [14] or a melting probe [13]). These actions can cause high costs in terms of, for example, energy consumption or execution time. In these situations, it is thus highly desirable to avoid non-informative actions by using adaptive selection strategies.

Acknowledgments. This work was supported by the German Federal Ministry for Economic Affairs and Energy (DLR project "KaNaRiA", funding no. 50 NA 1318, and DLR project "CAUSE", funding no. 50 NA 1505).

References

1. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8. IEEE (2008)
2. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV. vol. 1, pp. 1–2 (2004)
3. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding 106(1), 59–70 (2007), special issue on Generative Model Based Vision
4. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (Mar 2003)
5. Krieger, G., Rentschler, I., Hauske, G., Schill, K., Zetzsche, C.: Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. Spatial vision 13(2-3), 201–214 (2000)
6. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
7. Liu, H., Sun, J., Liu, L., Zhang, H.: Feature selection with dynamic mutual information. Pattern Recognition 42(7), 1330–1339 (2009)
8. Lowe, D.G.: Object recognition from local scale-invariant features. In: Computer vision, 1999. The proceedings of the seventh IEEE international conference on. vol. 2, pp. 1150–1157. Ieee (1999)
9. McCann, S., Lowe, D.G.: Local naive Bayes nearest neighbor for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 3650–3656. IEEE (2012)
10. Muja, M., Lowe, D.G.: Scalable nearest neighbor algorithms for high dimensional data. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36 (2014)
11. Najemnik, J., Geisler, W.S.: Optimal eye movement strategies in visual search. Nature 434(7031), 387–391 (2005)
12. Nakath, D., Kluth, T., Reineking, T., Zetzsche, C., Schill, K.: Active sensorimotor object recognition in three-dimensional space. In: Spatial Cognition IX, pp. 312–324. Springer (2014)
13. Niedermeier, H., Clemens, J., Kowalski, J., Macht, S., Heinen, D., Hoffmann, R., Linder, P.: Navigation system for a research ice probe for antarctic glaciers. In: IEEE/ION PLANS 2014. pp. 959–975. IEEE (2014)
14. Pavone, M., Acikmese, B., Nesnas, I.A., Starek, J.: Spacecraft autonomy challenges for next generation space missions (2013), `http://goo.gl/nU8xG0`, online, to appear in Lecture Notes in Control and Information Systems
15. Reineking, T., Schill, K.: Evidential object recognition based on information gain maximization. In: Cuzzolin, F. (ed.) Belief Functions: Theory and Applications, Lecture Notes in Computer Science, vol. 8764, pp. 227–236. Springer International Publishing (Sep 2014)

16. Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., Zetzsche, C.: Scene analysis with saccadic eye movements: Top-down and bottom-up modeling. Journal of Electronic Imaging 10(1), 152–160 (2001)
17. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological review 113(4), 766 (2006)
18. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML. vol. 97, pp. 412–420 (1997)
19. Zetzsche, C., Barth, E.: Image surface predicates and the neural encoding of two-dimensional signal variations. In: SC-DL tentative. pp. 160–177. International Society for Optics and Photonics (1990)

5 Summary and outlook

In this thesis the concept of intrinsic dimensionality was investigated in the context of second-order Volterra systems in application to low level functionalities reported in the early visual cortex. Furthermore, this concept was applied to numerical cognition and to action selection for active object recognition. The thesis aimed to answer the following research questions:

- How can neurons of the visual cortex be modeled so that they show a significantly nonlinear behavior in line with the concept of intrinsic dimensionality?
- How can numerical cognition be modeled from operations determined by the concept of intrinsic dimensionality so that human behavior can be explained?
- How can the action selection for active object recognition be influenced by information theoretical quantities and operations determined by the concept of intrinsic dimensionality?

The first question was addressed in Section 2. The insight that the reported behavior of neurons in early visual cortex cannot be explained solely by linear systems anymore makes the problem more complex. The number of possible models increases dramatically. To overcome this issue, the second-order Taylor series of nonlinear systems, i.e. second-order Volterra systems, is used to design nonlinear systems which are able to explain the reported behavior. Based on this formalism a nonlinear generalized Gabor filter was developed and parametrized to obtain the selectivity to oriented $i2D$-signals. In particular, systems being selective to crossing lines, end-stopped lines, and corners were developed. It was shown that the proposed parametrization can qualitatively describe phenomena reported in the literature. The results give first insights in the four-dimensional domain of the filter functions and the abilities to extract relevant features from images. In comparison to the linear approach, the nonlinear approach is far away from being well understood. This and the investigation of higher-order Volterra systems thus remain future research. The qualitative results regarding neural behavior directly lead to the question whether the developed systems are able to explain the empirical data quantitatively. In order to do the evaluation a data fitting algorithm for the generalized Gabor approach has to be developed in the future. Another open question is its relation to the proposed non-classical receptive field models in the literature.

The second question regarding the development of a computational model for numerical cognition was addressed in Section 3. In this section, a computational model for numerosity estimation was developed from scratch. The immense abstraction ability of the human system was investigated from a mathematical point of view. Under certain assumptions the topological invariant Euler characteristic can be used to develop a model for the number of objects. The relation between topology and differential geometry provided by the Gauss-Bonnet

theorem was then used to derive a computational model to obtain the number from a specific class of visual stimuli. This model is based on the geodesic curvature and the Gaussian curvature. It was also shown that both operators fulfill the requirements of an $i2D$-system. Finally, by the introduction of noise to the system, behavioral results of humans in standard numerosity estimation tasks were reproduced. These results raise some questions for future research. The proposed model suggests that a first attempt to numerosity can be computed by operations already provided in early stages of the visual system. Whether numerosity is represented in higher cortical areas or whether there exists a representation in early stages, is still an open question in numerosity research. As the proposed model is sensitive to the kind of connectedness of the objects by definition of the Euler characteristic, this sensitivity should be tested in behavioral experiments. This is directly related to the assumption that objects are simply connected. The generalization to arbitrarily connected objects requires the study of other invariants like the Betti numbers. The investigation of these invariants regarding their usability to derive a computational model remains future research.

The third research question which was addressed in Section 4 is split into two parts. In the first part a sensorimotor system based on a visual sensor device was developed which provides the ability to choose the next appropriate action for the information gathering process. The reasoning system was designed on top of Bayesian networks which were used to infer an object class from the features extracted from the visual input. The probabilistic knowledge base was used to formulate an information theoretical score function to obtain the next action to be performed by the sensorimotor system. The information gain strategy was able to decrease the number of performed actions which were necessary to reach a certain level of performance. The second part considers a similar sensorimotor system with small improvements: The quantization of feature vectors into a finite number of classes in the previously considered reasoning system annihilates important information for object recognition. It thus was replaced by a continuous approximation of the probability distribution which provides the knowledge base. In this framework various probabilistic and information theoretical score functions were investigated. In contrast to the score functions, relying on previously gained experience, an $i2D$-operator, the clipped eigenvalue operator, was used to provide a score function which does not rely on knowledge. The comparison of numbers of performed actions required to reach a certain level of performance showed that the knowledge-based score functions perform best followed by the $i2D$-operator based score function. This raises the question whether an improvement of the extracted $i2D$-features or hybrid approaches which combine knowledge- with image-based approaches can yield better score functions for object recognition. Whether humans use one of these strategies is an open question and should be investigated in future behavioral experiments.

In summary, it is an impressive result that the concept of intrinsic dimensionality seems to play an important role in low level vision and in an increasing number of higher cognitive abil-

ities. Further exciting applications of intrinsic dimensionality in human brain functionalities remain future research.

List of Figures

1.1 Mexican hat function 5
1.2 Classical Gabor filter kernels 7
1.3 Illustration of typical $i0D$-, $i1D$-, and $i2D$-signals 9
2.1 Classical Gabor filter kernels in Fourier space 19
2.2 Illustration of analyzed $i1D$-signals 22
2.3 Three-dimensional illustration of the forbidden region M 24
2.4 Three-dimensional illustration of the generalized Gabor filter H_2 35
2.5 Test stimuli of the kind "crossing" and "corner" 36
2.6 Three-dimensional illustration of the results of the generalized Gabor filters H_1, H_2, and H_3 37
2.7 Color-encoded illustration of the results of the generalized Gabor filters H_1, H_2, and H_3 38

List of Tables

1 Parameter constraints for the generalized Gabor approach 33
2 Parametrization of the generalized Gabor filters H_1, H_2, and H_3 33

References

[1] D. L. Adams and J. C. Horton. A precise retinotopic map of primate striate cortex generated from the representation of angioscotomas. *The Journal of neuroscience*, 23(9):3771–3789, 2003.

[2] F. Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183–193, 1954.

[3] H. Aubert and C. Foerster. Beiträge zur kenntniss des indirecten sehens. (i). untersuchungen über den raumsinn der retina. *Archiv fuer Ophthalmologie*, 3:1–37, 1857.

[4] R. Bamler. *Mehrdimensionale lineare Systeme: Fourier-Transformation und Delta-Funktionen.* Nachrichtentechnik ; 20. Springer, Berlin, 1989.

[5] T. Banchoff and S. T. Lovett. *Differential geometry of curves and surfaces.* Peters, Natick, Mass., 2010.

[6] E. Barth, T. Caelli, and C. Zetzsche. Image encoding, labeling, and reconstruction from differential geometry. *CVGIP: Graphical models and image processing*, 55(6):428–446, 1993.

[7] I. Biederman. Human image understanding: Recent research and a theory. *Computer vision, graphics, and image processing*, 32(1):29–73, 1985.

[8] C. D. Bonn and J. F. Cantlon. The origins and structure of quantitative concepts. *Cognitive Neuropsychology*, 29(1-2):149–173, 2012.

[9] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. Appearance-based active object recognition. *Image and Vision Computing*, 18(9):715–727, 2000.

[10] M. B. Brilliant. Theory of the analysis of nonlinear systems. Technical Report 304, MIT Research Lab of Electronics, 1958.

[11] F. W. Campbell and J. Robson. Application of fourier analysis to the visibility of gratings. *The Journal of Physiology*, 197(3):551–566, 1968.

[12] M. Carandini. What simple and complex cells compute. *The Journal of Physiology*, 577(2):463–466, 2006.

[13] M. Carandini, D. J. Heeger, and J. A. Movshon. Linearity and normalization in simple cells of the macaque primary visual cortex. *The Journal of Neuroscience*, 17(21):8621–8644, 1997.

[14] J. R. Cavanaugh, W. Bair, and J. A. Movshon. Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *Journal of Neurophysiology*, 88(5):2530–2546, 2002.

[15] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A*, 2(7):1160–1169, 1985.

[16] R. L. de Valois and K. K. de Valois. *Spatial vision*. Oxford psychology series; 14. Oxford University Press, New York, 1988.

[17] S. Dehaene. *The number sense: How the mind creates mathematics*. Oxford University Press, New York, 2011.

[18] S. Dehaene, G. Dehaene-Lambertz, and L. Cohen. Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences*, 21(8):355–361, 1998.

[19] J. E. Dowling. *The retina: an approachable part of the brain*. Harvard University Press, 1987.

[20] E. Famiglietti and H. Kolb. Structural basis for on-and off-center responses in retinal ganglion cells. *Science*, 194(4261):193–195, 1976.

[21] G. T. Fechner. *Elemente der Psychophysik*. Breitkopf und Härtel, Leipzig, 1860.

[22] L. Feigenson, S. Dehaene, and E. Spelke. Core systems of number. *Trends in Cognitive Sciences*, 8(7):307–314, 2004.

[23] D. K. Freedheim and I. B. Weiner. *Handbook of Psychology: History of Psychology*. Wiley, Hoboken, NJ, 2003.

[24] C. R. Gallistel and R. Gelman. Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences*, 4(2):59–65, 2000.

[25] J. J. Gibson. *The ecological approach to visual perception*. Houghton Mifflin, Boston, 1979.

[26] A. Gray. *Modern differential geometry of curves and surfaces*. Studies in advanced mathematics. CRC Press, Boca Raton, 1993.

[27] J. G. Greeno. Gibson's affordances. *Psychological Review*, 101(2):336–342, 1994.

[28] J. Grèzes and J. Decety. Does visual perception of object afford action? evidence from a neuroimaging study. *Neuropsychologia*, 40(2):212–222, 2002.

[29] J. Halberda, M. M. M. Mazzocco, and L. Feigenson. Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455(7213):665–668, 2008.

[30] M. Hashemi-Nezhad and D. C. Lyon. Orientation tuning of the suppressive extraclassical surround depends on intrinsic organization of v1. *Cerebral Cortex*, 22(2):308–326, 2012.

[31] S. Hattar, H.-W. Liao, M. Takao, D. M. Berson, and K.-W. Yau. Melanopsin-containing retinal ganglion cells: architecture, projections, and intrinsic photosensitivity. *Science*, 295(5557):1065–1070, 2002.

[32] M. D. Hauser, S. Carey, and L. B. Hauser. Spontaneous number representation in semi–free–ranging rhesus monkeys. *Proceedings of the Royal Society of London B: Biological Sciences*, 267(1445):829–833, 2000.

[33] H. Heuser. *Lehrbuch der Analysis Teil 2.* Vieweg + Teubner, Wiesbaden, 2008.

[34] D. H. Hubel. *Eye, brain, and vision.* Scientific American library; 22. Scientific American Books, New York, NY, 1988.

[35] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, 28(2):229–289, 1965.

[36] D. H. Hubel and T. N. Wiesel. Early exploration of the visual cortex. *Neuron*, 20(3):401–412, 1998.

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

[38] W. S. Jevons. The power of numerical discrimination. *Nature*, 3:281–282, 1871.

[39] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987.

[40] E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann. The discrimination of visual number. *The American Journal of Psychology*, 62(4):498–525, 1949.

[41] T. W. Kjaer, T. J. Gawne, J. A. Hertz, and B. J. Richmond. Insensitivity of v1 complex cell responses to small shifts in the retinal image of complex patterns. *Journal of Neurophysiology*, 78(6):3187–3197, 1997.

[42] T. Kluth, D. Nakath, T. Reineking, C. Zetzsche, and K. Schill. Affordance-based object recognition using interactions obtained from a utility maximization principle. In

L. Agapito, M. M. Bronstein, and C. Rother, editors, *Computer Vision - ECCV 2014 Workshops*, volume 8926 of *Lecture Notes in Computer Science*, pages 406–412. Springer International Publishing, 2015.

[43] T. Kluth and C. Zetzsche. Spatial numerosity: A computational model based on a topological invariant. In C. Freksa, B. Nebel, M. Hegarty, and T. Barkowsky, editors, *Spatial Cognition IX*, volume 8684 of *Lecture Notes in Computer Science*, pages 237–252. Springer International Publishing, 2014.

[44] O. Koehler. Vom erlernen unbenannter anzahlen bei vögeln. *Die Naturwissenschaften*, 29(14–15):201–218, 1941.

[45] G. Krieger, I. Rentschler, G. Hauske, K. Schill, and C. Zetzsche. Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial vision*, 13(2-3):201–214, 2000.

[46] N. Krüger and M. Felsberg. A continuous formulation of intrinsic dimension. In *Proc. British Machine Vision Conference*, Norwich, September 2003.

[47] S. W. Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16(1):37–68, 1953.

[48] W. Kühnel. *Differentialgeometrie: Kurven - Flächen - Mannigfaltigkeiten.* Vieweg+Teubner Verlag, Wiesbaden, 2010.

[49] J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, and W. H. Pitts. What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11):1940–1951, 1959.

[50] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London B: Biological Sciences*, 207(1167):187–217, 1980.

[51] D. Marr and S. Ullman. Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London B: Biological Sciences*, 211(1183):151–180, 1981.

[52] S. Marčelja. Mathematical description of the responses of simple cortical cells∗. *J. Opt. Soc. Am.*, 70(11):1297–1300, 1980.

[53] R. Mehrotra, K. R. Namuduri, and N. Ranganathan. Gabor filter-based edge detection. *Pattern Recognition*, 25(12):1479–1494, 1992.

[54] D. Nakath, T. Kluth, T. Reineking, C. Zetzsche, and K. Schill. Active sensorimotor object recognition in three-dimensional space. In C. Freksa, B. Nebel, M. Hegarty, and T. Barkowsky, editors, *Spatial Cognition IX*, volume 8684 of *Lecture Notes in Computer Science*, pages 312–324. Springer International Publishing, 2014.

[55] A. Nieder and E. K. Miller. Coding of cognitive magnitude: compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, 37(1):149–157, 2003.

[56] U. Nuding and C. Zetzsche. Learning the selectivity of v2 and v4 neurons using non-linear multi-layer wavelet networks. *Biosystems*, 89(1-3):273–279, 2007.

[57] G. A. Orban. *Neuronal operations in the visual cortex*. Studies of brain function; 11. Springer, Berlin, 1984.

[58] J. K. O'Regan, , and A. Noë. What it is like to see: A sensorimotor theory of perceptual experience. *Synthese*, 129(1):79–103, 2001.

[59] J. K. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–973, 2001.

[60] G. Osterberg. *Topography of the layer of rods and cones in the human retina*. Nyt Nordisk Forlag, 1935.

[61] I. M. Pepperberg. Grey parrot numerical competence: a review. *Animal cognition*, 9(4):377–391, 2006.

[62] V. H. Perry and A. Cowey. Retinal ganglion cells that project to the superior colliculus and pretectum in the macaque monkey. *Neuroscience*, 12(4):1125–1137, 1984.

[63] A. J. Rockel, R. W. Hiorns, and T. P. S. Powell. The basic uniformity in structure of the neocortex. *Brain*, 103(2):221–244, 1980.

[64] H.-a. Saito, K. Tanaka, Y. Fukada, and H. Oyamada. Analysis of discontinuity in visual contours in area 19 of the cat. *The Journal of Neuroscience*, 8(4):1131–1143, 1988.

[65] M. Schetzen. *The Volterra and Wiener theories of nonlinear systems*. John Wiley & Sons, 1980.

[66] B. Schiele and J. L. Crowley. Transinformation for active object recognition. In *Computer Vision, 1998. Sixth International Conference on*, pages 249–254. IEEE, 1998.

[67] K. Schill. Decision support systems with adaptive reasoning strategies. In C. Freksa, M. Jantzen, and R. Valk, editors, *Foundations of Computer Science*, volume 1337 of *Lecture Notes in Computer Science*, pages 417–427. Springer Berlin Heidelberg, 1997.

[68] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetzsche. Scene analysis with saccadic eye movements: top-down and bottom-up modeling. *Journal of Electronic Imaging*, 10(1):152–160, 2001.

[69] A. M. Schmid. The processing of feature discontinuities for different cue types in primary visual cortex. *Brain research*, 1238:59–74, 2008.

[70] E. L. Schwartz. Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. *Vision Research*, 20(8):645–669, 1980.

[71] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 994–1000. IEEE, 2005.

[72] Z.-M. Shen, W.-F. Xu, and C.-Y. Li. Cue-invariant detection of centre–surround discontinuity by v1 neurons in awake macaque monkey. *The Journal of Physiology*, 583(2):581–592, 2007.

[73] I. Shevelev, N. Lazareva, G. Sharaev, R. Novikova, and A. Tikhomirov. Selective and invariant sensitivity to crosses and corners in cat striate neurons. *Neuroscience*, 84(3):713–721, 1998.

[74] S. Shushruth, P. Mangapathy, J. M. Ichida, P. C. Bressloff, L. Schwabe, and A. Angelucci. Strong recurrent networks compute the orientation tuning of surround modulation in the primate primary visual cortex. *The Journal of Neuroscience*, 32(1):308–321, 2012.

[75] A. M. Sillito, K. L. Grieve, H. E. Jones, J. Cudeiro, and J. Davis. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, 378(6556):492–496, 1995.

[76] B. C. Skottun, R. L. De Valois, D. H. Grosof, J. A. Movshon, D. G. Albrecht, and A. B. Bonds. Classifying simple and complex cells on the basis of response modulation. *Vision Research*, 31(7–8):1078–1086, 1991.

[77] W. Strampp. *Mathematische Methoden der Signalverarbeitung.* Oldenbourg Lehrbücher für Ingenieure. Oldenbourg, München, 2010.

[78] R. Unbehauen. *Systemtheorie 1: Allgemeine Grundlagen, Signale und lineare Systeme im Zeit- und Frequenzbereich.* Naturwissenschaft und Technik 1/2010. Oldenbourg Wissenschaftsverlag, München, 2009.

[79] V. Walsh. A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, 7(11):483–488, 2003.

[80] B. A. Wandell. *Foundations of vision.* Sinauer Associates, Sunderland, MA, 1995.

[81] E. H. Weber. *Die Lehre vom Tastsinne und Gemeingefühle auf Versuche gegründet.* Friedrich Vieweg und Sohn, Braunschweig, 1851.

[82] B. Wegmann and C. Zetzsche. Feature-specific vector quantization of images. *Image Processing, IEEE Transactions on*, 5(2):274–288, 1996.

[83] J. Whalen, C. R. Gallistel, and R. Gelman. Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, 10(2):130–137, 1999.

[84] K. Wynn. Addition and subtraction by human infants. *Nature*, 358(6389):749–750, 1992.

[85] F. Xu and E. S. Spelke. Large number discrimination in 6-month-old infants. *Cognition*, 74(1):B1–B11, 2000.

[86] C. Zetzsche and E. Barth. Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision Research*, 30(7):1111–1117, 1990.

[87] C. Zetzsche and E. Barth. Image surface predicates and the neural encoding of two-dimensional signal variations. In *Human Vision and Electronic Imaging: Models, Methods, and Applications*, volume 1249 of *Proc. SPIE*, pages 160–177, 1990.

[88] C. Zetzsche, E. Barth, and B. Wegmann. The importance of intrinsically two-dimensional image features in biological vision and picture coding. In A. B. Watson, editor, *Digital images and human vision*, pages 109–138. MIT Press, Cambridge, MA, 1993.

[89] C. Zetzsche, K. Gadzicki, and T. Kluth. Statistical invariants of spatial form: From local and to numerosity. In O. Kutz, M. Bhatt, S. Borgo, and P. Santos, editors, *Proceedings of the Second Interdisciplinary Workshop The Shape of Things*, volume 1007 of *Workshop Proceedings*, pages 163–172, Aachen, Apr. 2013. CEUR-WS.org.

[90] C. Zetzsche and G. Krieger. Intrinsic dimensionality: nonlinear image operators and higher-order statistics. In S. K. Mitra and G. L. Sicuranza, editors, *Nonlinear image processing*, pages 403–441. Academic Press, San Diego, CA, 2000.

[91] C. Zetzsche and G. Krieger. Nonlinear mechanisms and higher-order statistics in biological vision and electronic image processing: review and perspectives. *Journal of Electronic Imaging*, 10(1):56–99, 2001.

[92] C. Zetzsche and U. Nuding. Natural scene statistics and nonlinear neural interactions between frequency-selective mechanisms. *Biosystems*, 79(1–3):143–149, 2005.

[93] C. Zetzsche and U. Nuding. Nonlinear and higher-order approaches to the encoding of natural scenes. *Network: Computation in Neural Systems*, 16(2-3):191–221, 2005.

[94] C. Zetzsche and U. Nuding. Nonlinear encoding in multilayer lnl systems optimized for the representation of natural images. In *Human Vision and Electronic Imaging XII*, volume 6492 of *Proc. SPIE*, pages 649204–649225, 2007.

[95] C. Zetzsche, K. Schill, H. Deubel, G. Krieger, E. Umkehrer, and S. Beinlich. Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach. In R. Pfeifer, editor, *From animals to animats 5: proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*, pages 120–126. MIT Press, Cambridge, MA, 1998.

[96] C. Zetzsche, J. Wolter, and K. Schill. Sensorimotor representation and knowledge-based reasoning for spatial exploration and localisation. *Cognitive processing*, 9(4):283–297, 2008.