IEEE Trans Pattern Anal Mach Intell. 2013 Aug;35(8):1847-71.

Deep Hierarchies in the Primate Visual Cortex: What Can We Learn For Computer Vision?

Norbert Krüger, Peter Janssen, Sinan Kalkan, Markus Lappe, Aleš Leonardis, Justus Piater, Antonio J. Rodríguez-Sánchez, Laurenz Wiskott

Abstract—Computational modeling of the primate visual system yields insights of potential relevance to some of the challenges that computer vision is facing, such as object recognition and categorization, motion detection and activity recognition or vision-based navigation and manipulation. This article reviews some functional principles and structures that are generally thought to underlie the primate visual cortex, and attempts to extract biological principles that could further advance computer vision research. Organized for a computer vision audience, we present *functional principles* of the *processing hierarchies* present in the primate visual system considering recent discoveries in neurophysiology. The hierarchal processing in the primate visual system is characterized by a sequence of different levels of processing (in the order of ten) that constitute a *deep hierarchy* in contrast to the *flat* vision architectures predominantly used in today's mainstream computer vision. We hope that the functional description of the deep hierarchies realized in the primate visual system provides valuable insights for the design of computer vision algorithms, fostering increasingly productive interaction between biological and computer vision research.

Index Terms—Computer Vision, Deep Hierarchies, Biological Modeling

1 Introduction

The history of computer vision now spans more than half a century. However, general, robust, complete satisfactory solutions to the major problems such as large-scale object, scene and activity recognition and categorization, as well as vision-based manipulation are still beyond reach of current machine vision systems. Biological visual systems, in particular those of primates, seem to accomplish these tasks almost effortlessly and have been, therefore, often used as an inspiration for computer vision researchers.

Interactions between the disciplines of "biological vision" and "computer vision" have varied in intensity throughout the course of computer vision history and have in some way reflected the changing research focuses of the machine vision community [32]. Without any doubt, the groundbreaking work of Hubel and Wiesel [72] gave a significant impulse to the computer vision community via Marr's work on building visual hierarchies analogous to the primate visual system [109]. However, the insufficient computational resources that were available at that time and the lack of more detailed understanding of the processing stages in the primate visual system presented two insurmountable obstacles to further progress in that direction.

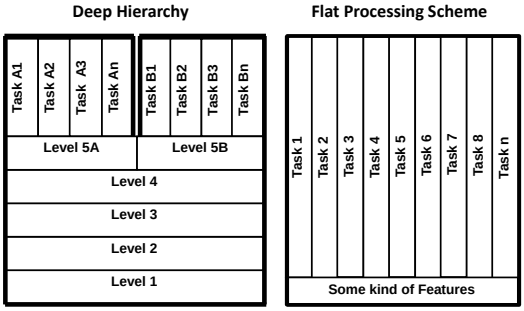


Fig. 1. Deep hierarchies and flat processing schemes

What followed was a reorientation of mainstream computer vision from trying to solve general vision problems to focusing more on specific methods related to specific tasks. This has been most commonly achieved in *flat processing schemes* (see figure 1, right) in which rather simple feature-based descriptors were taken as an input and then processed by the task-dependent learning algorithms. The ties with the biological vision faded, and if there were some references to biological-related mechanisms they were most commonly limited to individual functional modules or feature choices such as Gabor wavelets.

While the progress on some specialized machine vision problems and problem domains has been enormous (on some tasks, these systems can easily surpass human capabilities), artificial systems still lack the generality and robustness inherent in the primate visual system. As we are gaining

- *N. Krüger is first author since he initiated and organized the writing of this paper, all other authors are ordered alphabetically.*
- *N. Krüger is with the Maersk Mc-Kinney Moller Institute at the University of Southern Denmark.*
- *P. Janssen is with the Division of Neurophysiology at the KU Leuven.*
- *S. Kalkan is with the Dept. of Computer Engineering, Middle East Technical University, Turkey.*
- *M. Lappe is with the Institute for Psychology of the University of Muenster, Germany.*
- *A. Leonardis is with the Faculty of Computer and Information Science, University of Ljubljana, Slovenia, and with the Centre for Computational Neuroscience and Cognitive Robotics, University of Birmingham, United Kingdom.*
- *J. Piater and A. Rodríguez-Sánchez are with the Intelligent and Interactive Systems group at the University of Innsbruck, Austria.*
- *L. Wiskott is with the Institut für Neuroinformatik at the Ruhr-Universität Bochum, Germany.*

more and more insight into the functional mechanisms of the visual cortex (largely due to the advanced imaging techniques used in neuroscience), the time may be ripe to make a new attempt at looking at the mechanisms that could bring the capabilities of artificial vision, primarily in terms of generality and robustness, closer to those of biological systems. This may be a feasible enterprise also from the computational point of view, particularly in the light of new developments of computer architectures such as GPUs and multi-core systems.

In this paper, we will primarily focus on hierarchical representations and functional mechanisms of primates. We will look at different hierarchical levels of processing as well as different information channels (e.g., shape, color, motion) and discuss information abstractions that occur throughout the hierarchy.

It is known that around 55% of the neocortex of the primate brain is concerned with vision [44] and that there is a hierarchical organization of the processing pipeline that spans 8 to 10 levels (see figure 2). There is clear evidence that neurons in the early visual areas extract simple image features (e.g., orientation, motion, disparity, etc.) over small local regions of visual space and that this information is then transmitted to neurons in higher visual areas which respond to ever more complex features with receptive fields[1] covering larger and larger regions of the visual field. Such hierarchical structures, to which we refer as *deep hierarchies* (see figure 1, left), exhibit a number of computational advantages compared to the so-called *flat processing schema* (see figure 1, right).

Two important aspects are computational efficiency and generalization: As the hierarchical levels build on top of each other, they exploit the *shareability* of the elements to efficiently arrive at more complex information units. Such a design principle also contributes to common computations (both during learning and inference) which results in highly efficient processing as well as in lower storage demands. Moreover, reusing commonalities that exist among different visual entities and which are important perceptual building blocks for achieving different tasks leads to generalization capabilities and transfer of knowledge. For example, there is strong neurophysiological evidence that a generic description in terms of a variety of visual properties is computed in areas V1–V4 and MT, covering around 60% of the volume of visual processing in the primate neocortex (see [44] and figure 2 where visual areas are drawn proportionally to their actual sizes). These areas carry necessary information for a completion of a number of different tasks, such as object recognition and categorization, grasping, manipulation, path planning, etc.

It is also evident that in the visual system of primates there are separate (though highly inter-connected) channels that process different types of visual information (color, shape, motion, texture, 3D information), which contribute to the efficiency of representation (avoiding the combinatorial explosion of an integrated representation) and robustness (with respect to the available information). These advantages cover multiple aspects and will be discussed in more detail in section 8.

However, although all neurophysiological evidence suggests that in the primate visual system quite a number of levels are realized, most existing computer vision systems are 'flat' and hence cannot make use of the advantages connected to deep hierarchies. Here in particular the generalization and scalability capabilities are crucial for any form of cognitive intelligence. In fact, there is overwhelming neurophysiological evidence that cognition and the concept of deep hierarchies are linked [178]. As a consequence, we see the issue of establishing deep hierarchies as one major challenge on our way towards artificial cognitive systems.

Bengio [9] discussed the potential of deep hierarchies as well as fundamental problems related to learning of deep hierarchies. In particular, he emphasizes the problem of the huge parameter space that has to be explored due the large number of hierarchical levels. This learning problem can be alleviated by (a) tackling intermediate representations as independent learning problems as well as (b) introducing bias in terms of basic connectivity structures expressed in the number of levels or the locality of connectivity of individual units of such deep structures. We believe that this paper can help to guide the learning process of deep hierarchies for vision systems by giving indications for suitable intermediate representations in the primate's visual system. In addition, we believe that useful guidelines for connectivity patterns can be derived from the biological model in terms of appropriate receptive field sizes of neurons, number of levels being processed in the biological model as well as the number of units in a certain hierarchical level as indicated by area sizes in the primate's visual cortex.

Despite the challenges connected to the learning of deep hierarchies, there exists a body of work in computer vision that made important contributions towards understanding and building hierarchical models. Due to lack of space, a more thorough review is outside the scope of this paper, and the following list is far from complete. From the computational complexity point of view, Tsotsos has shown that unbounded visual search is NP complete and that hierarchical architectures may be the most promising solution to tackle the problem [189]. Several works have shown that efficient matching can only be performed in several hierarchical stages, including Ettinger [42], Geman et al. [59], [60], Mel and Fiser [114], Amit [1] [2], Hawkins [69], Fidler et al. [45], Scalzo and Piater [153], Ullman and Epshtein [192], DiCarlo and Cox [31], Ommer and Buhmann [126], Serre and Poggio [157], Pugeault et al. [138], and Rodríguez-Sánchez [144]. Among the more known hierarchical models are the Neocognitron [54], HMAX [141], [158], LHOP [46], 2DSIL [145] and Convolutional Nets [101]. Recently, Bengio [9] published an exhaustive article on learning deep architectures for artificial intelligence.

In summary, in this article, we want to argue that deep hierarchies are an appropriate concept to achieve a general, robust, and versatile computer vision system. Even more importantly, we want to present relevant insights about the hierarchical organization of the primate visual system for computer vision scientists in an accessible way. We are aware that some of our abstractions are rather crude from the neurophysiological point of view and that we have left out important details of the

1. The *receptive field* of a neuron is the region where certain stimuli produce an effect on the neuron's firing.

processes occurring at the different levels[2], but we hope that such abstractions and the holistic picture given in this paper will help to foster productive exchange between the two fields.

The paper is organized as follows: In section 2, we will touch upon the aspects of the primate visual system that are relevant to understand and model the processing hierarchy. The hierarchy in the primate vision system is then outlined from two perspectives. In the *horizontal perspective* (sections 3–6) we give a description of processing in the different areas indicated in figure 2. In section 7, we give a *vertical perspective* on the processing of different visual modalities across the different areas. In section 8, we then draw conclusions for the modeling and learning of artificial visual systems with deep hierarchical structures.

2 Relevant Aspects of the Structure of the visual Cortex

In section 2.1, we provide a basic overview of the deep hierarchy in the primate visual system. In section 2.2, we also give an intuition of basic (mostly biological) terms used in the following sections. Most data we present in the following were obtained from macaque monkeys since most neurophysiological knowledge stems from investigations on these.

While the primate brain consists of approximately 100 cortical areas, the human brain probably contains as many as 150 areas.[3] There is a general consensus that the primary sensory and motor areas in the monkey are homologous to the corresponding areas in the human brain. Furthermore, several other cortical areas in the monkey have an identified homologue in the human (e.g. MT/MST, AIP). These areas can be viewed as landmarks which can be used to relate other cortical areas in the human to the known areas in the monkey.

It should be mentioned that a visual cortical area consists of six layers, which do not correspond to the layers in artificial deep models. In general, layer 4 is the input layer where the inputs from earlier stages arrive. The layers above layer 4 (layers 2 and 3) typically send feedforward connections to downstream visual areas (e.g. from V1 to V2), whereas layers 5 and 6 send feedback projections to upstream areas or structures (e.g. from V1 to the LGN and the Superior Colliculus – see also section 3.2). At higher stages in the visual hierarchy, the connectivity is almost always bidirectional. At present, detailed knowledge about the precise role of cortical microcircuits in these different layers is lacking.

2. For example, a heterogeneity of computations has been reported, including summation, rectification, normalization [19], averaging, multiplication, max-selection, winner-take all [150] and many others [89]. This is of great interest for addressing how neurons are inter-connected and the subject of much discussion but out of the scope of the present paper.

3. A region in the cerebral cortex can be considered to be an area based on four criteria: (1) cyto- and myeloarchitecture (the microscopic structure, cell types, appearance of the different layers, etc.), (2) the anatomical connectivity with other cortical and subcortical areas, (3) retinotopic organization, and (4) functional properties of the neurons. In far extrastriate cortex, where retinotopic organization is weak or absent, the specific functional properties of the neurons are an important characteristic to distinguish a region from the neighboring regions.

2.1 Hierarchical Architecture

Here we give a coarse and intuitive summary of the processing hierarchy realized in the primate visual system. A more detailed description can be found in sections 3 – 6. Basic data on the sizes of the different areas, receptive field sizes, latency, organization etc. are provided in table 1.

The neuronal processing of visual information starts in the retina of the left and right eye. Nearly all connections then project to a visual area called LGN before it reaches the visual cortex. We call these stages *precortical processing* and the processing in these areas is described in section 3. The visual cortex is commonly divided into three parts (figure 2 and table 1): the occipital part gives input to the dorsal and ventral streams. The occipital part covers the areas V1-V4 and MT. All areas are organized retinotopically, i.e., nearby neurons in the visual cortex have nearby receptive fields (see table 1, 6th column) and the receptive field size increases from V1 to V4 (see table 1, 3rd column). There are strong indications that these areas compute generic scene representations in terms of processing different aspects of visual information [84]. However, the complexity of features coded at the different levels increases with the level of the hierarchy as will be outlined in detail in section 4. Also it is worth noting that the size of the occipital part exceeds the other two parts occupying more than 62% of the visual cortex compared to 22% for the ventral and 11% for the dorsal pathway [44] (see table 1, 2nd column).[4] In the following, we call the functional processes established in the occipital part *early vision* indicating that a generic scene analysis is performed in a complex feature structure.

The ventral pathway covers the areas TEO and TE which are involved in object recognition and categorization. The receptive field sizes are in general significantly larger than in the occipital part. There is a weak retinotopic organization in area TEO which is not observed in area TE. Neurons' receptive fields usually include the fovea (the central part of the retina with the highest spatial resolution). In the ventral path, the complexity of features increases up to an object level for specific object classes (such as faces) [127], however most neurons are responsive to features below the object level indicating a coding scheme that uses multiple of these descriptors to code objects and scenes [173].

The dorsal pathway consists of the motion area MST and the visual areas in posterior parietal cortex. The dorsal stream is engaged in the analysis of space and in action planning. Similar to the ventral stream, the receptive field sizes increase along the dorsal pathway and the complexity of stimulus features increases progressively (e.g. from simple motion in MT to more complex motion patterns in MST and VIP). Moreover, the relation of receptive fields to retinal locations weakens. Instead, higher areas encode the location of stimuli in spatial or head fixed coordinates.

Besides the division into two pathways (ventral and dorsal) it is worth noting that there are also two streams to be

4. These proportions are unknown for the human visual cortex because in both the temporal and the parietal lobe new areas have probably evolved in humans compared to monkeys.

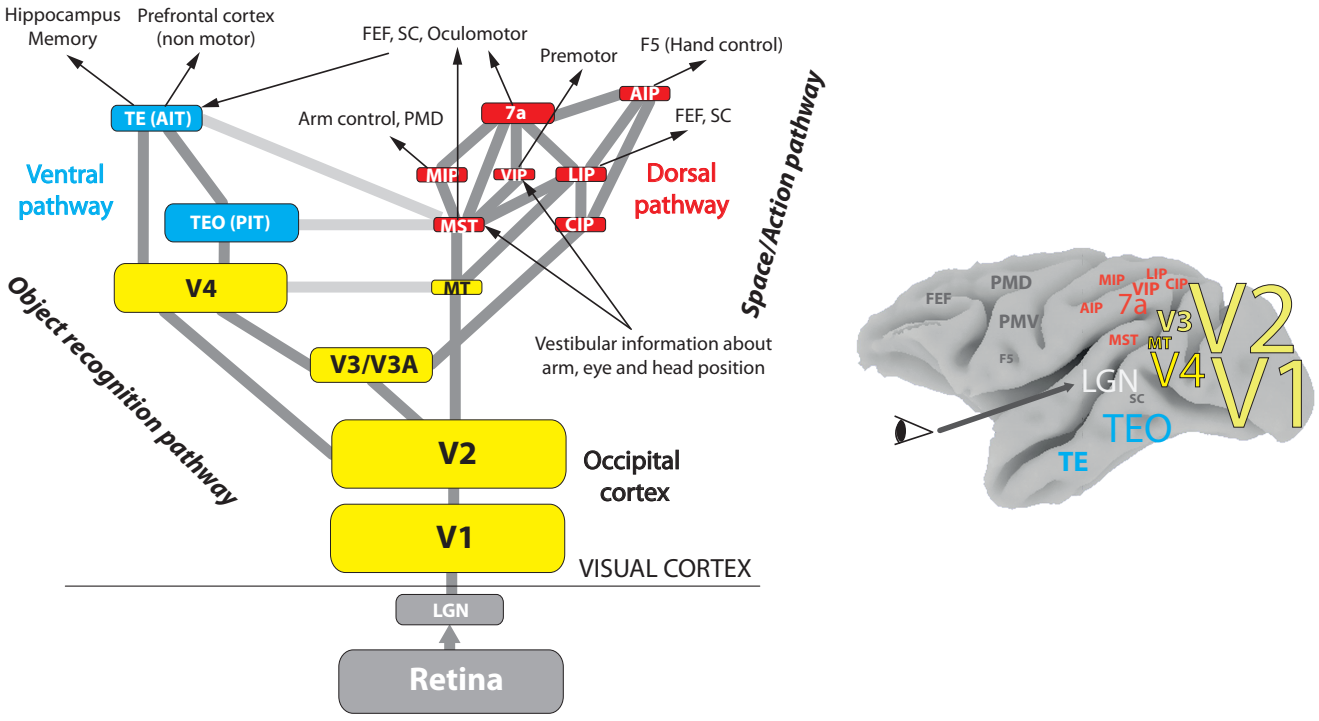


Fig. 2. Simplified hierarchical structure of the primate's visual cortex and approximate area locations (summarized from [44]). Box and font sizes are relative to the area size.

distinguished, the magnocellular (M–) and parvocellular (P–) stream [73]. This distinction is already present at the ganglion cell level, i.e., at the level of the output of the retina. P ganglion cells are color sensitive, have a small receptive field and are responsible for the high visual acuity in the central visual field. M ganglion cells have lower spatial but higher temporal resolution than P ganglion cells. The distinction between P and M cells carries through LGN to the whole visual cortex. To a first approximation, the P path is believed to be responsible for shape and object perception while the M path can account for the perception of motion and sudden changes [84]. Also the strongly space-variant resolution from the fovea to the visual periphery carries through most regions of the visual cortex.

It is worth noting that at every stage in the visual hierarchy, neurons also exhibit selectivities that are present at earlier stages of the hierarchy (e.g. orientation selectivity can be observed up to the level of TEO).

It is in general acknowledged that the influence of extrinsic information on the visual representations in the brain increases with its level in the hierarchy. For example, there is no report on any learning or adaptation processes in the retina and also quite some evidence on a high influence of genetic pre-structuring for orientation maps in V1 (see, e.g., [63]). On the other hand, it has also been shown that learning can alter the visual feature selectivity of neurons. However, the measurable changes at the single-cell level induced by learning appear to be much smaller at earlier levels in the visual hierarchy such as V1 [155] compared to later stages such as V4 [140] or IT [104].

2.2 Basic Facts on Different Visual Areas

Table 1 gives basic data on the different areas of the visual system. The first column indicates the name of the area, the second column the size in mm^2 (see also figure 2 where areas are drawn proportionally to their area size). The third column indicates the average receptive field size at 5 degrees of eccentricity. The fourth column indicates the latency to the first response to a stimuli at the retina.

Figure 3 provides a summary of most of the terms that follow in columns 5 through 7. The fifth column distinguishes between contra- and bilateral receptive fields. Contralateral (co in table 1) receptive fields only cover information from one hemifield while bilateral (bl in table 1) receptive fields cover both hemifields (figure 3b). The sixth column indicates different schemas of organization: Retinotopic organization (rt) indicates that the spatial arrangement of the inputs from the retina is maintained which changes every time we move our eyes, spatiotopic (st) indicates the representation of the world in real-world coordinates (see figure 3a), clustered organization (cl) indicates that there are larger subareas with similar functions, columnar organization (co) indicates that there is a systematic organization in columns according to some organizational scheme (mostly connected to visual features or retinotopy). The seventh column indicates different kinds of invariances (see figure 3c-f): cue invariance (CI) refers to the ability to obtain the same type of information from different cues, a cell that responds to an object independently of its size is called size invariant (SI), similarly for position

Area	Size (mm^2)	RFS	Latency (ms)	co/bi lat.	rt/st/cl/co	CI/SI/PI/OI	Function
	Sub–cortical processing						
Retina	1018	0.01	20-40	bl	+/-/-/-	-/-/-/-	sensory input, contrast computation
LGN		0.1	30-40	co	+/-/-/-	-/-/-/-	relay, gating
	Occipital / Early Vision						
V1	1120	3	30-40	co	+/-/-/+	-/-/-/-	generic feature processing
V2	1190	4	40	co	+/-/-/+	-/-/-/-	generic feature processing
V3/V3A/VP	325	6	50	co	+/-/-/+	-/-/-/-	generic feature processing
V4/VOT/V4t	650	8	70	co	+/-/-/+	+/-/-/-	generic feature processing / color
MT	55	7	50	co	+/-/-/+	+/+/-/+	motion
Sum	3340						
	Ventral Pathway / What (Object Recognition and Categorization)						
TEO	590	3-5	70	co	(+)/-/-/+	?/-/-/?	object recognition and
TE	180	10-20	80-90	bl	-/-/+/+	+/+/+/+(-)	categorization
Sum	770						
	Dorsal Pathway / Where and How (Coding of Action Relevant Information)						
MST	60	>30	60-70	bl	+/-/+/-	I	optic flow, self-motion, pursuit
CIP	?	?	?		+/-/?/?	+/?/?/?	3D orientation of surfaces
VIP	40	10-30	50-60	bl	-/+/-/-	I	optic flow, touch, near extra personal space
7a	115	>30	90	bl	(+)/-/-/-	?/?/+/?	Optic flow, heading
LIP	55	12-20	50	cl	+/-/-/-	?/-/-/-	salience, saccadic eye movements
AIP	35	5-7	60	bl	?/+/+/?	?/+/+/?	grasping
MIP	55	10-20	100	co	+/-/?/?	I	reaching
Sum	585						

TABLE 1

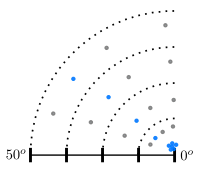
Basic facts on the different areas of the macaque visual cortex based on different sources [44], [28], [95], [142], [162] *First column:* Name of Area. *Second column:* Size of area in mm^2. '?' indicates that this information is not available. *Third column:* Average receptive field size in degrees at 5 degree of eccentricity. *Fourth column:* Latency in milliseconds. *Fifth Column:* Contra versus bilateral receptive fields. *Sixth Column:* Principles of organization: Retinotopic (rt), spatiotopic (st), clustered (cl), columnar (co). *Seventh Column:* Invariances in representation of shape: Cue Invariance (CI), Size Invariance (SI), Position Invariance (PI), Occlusion Invariance (OI). 'I' indicates that this entry is irrelevant for the information coded in these areas. *Eighth Column:* Function associated to a particular area.

invariance (PI). Finally, a cell that responds similarly to an object irrespective of whether it is completely or partially present is invariant to occlusions (OI).

3 Sub-cortical vision

In this section, we describe the primate sub-cortical vision system. We begin with the retinal photoreceptors as the first stage of visual processing (section 3.1), and follow the visual signal from the eye through the Lateral Geniculate Nucleus (LGN) (section 3.2). For all areas, we first give a neurophysiological and then a functional perspective.

3.1 Base Level: Retinal Photoreceptors



The retina is located in the inner surface of the eye and contains photoreceptors that are sensitive only to a certain interval of the electromagnetic spectrum, as well as cells that convert visual information to neural signals. The pictogram on the left illustrates the space-variant retinal density of rods (gray) and cones (blue) as described below, as well as the uniformly-small receptive field sizes (around 0.01° of visual angle). Compare this to the corresponding pictograms we consistently give in the following sections.

Neurophysiological view: There are two kinds of photoreceptors, rods and cones. Rods have a high sensitivity to low levels of brightness (see icons at the left). Cones, on the other hand, require high levels of brightness. We can classify the cones as a function of their wavelength absorbency as S (short wavelength = blue), M (middle wavelength = green) and L (long wavelength = red) cones. These three cone types allow for the perception of color [13]. The resolution (i.e., the number of receptors per mm^2) decreases drastically with the distance from the fovea. This holds for both rods and cones, except that there are no rods in the fovea. Most cones are concentrated in and around the fovea, while rods constitute the bulk of the photoreceptors at high eccentricities.

Functional view: Because only a small part of the retina has a high spatial resolution (the fovea), gaze control is required to direct the eyes such that scene features of interest project onto the fovea. Therefore, primates possess an extensive system for active control of eye movements (involving the FEF in the frontal lobe, LIP in the parietal lobe and the Superior Colliculus in the midbrain). It is influenced both by reflexive, signal-driven and by intentional, cognitively-driven attentional mechanisms, and involves the entire visual hierarchy. Attention models compute where to fixate [135], [143] and some work even addresses learning to control gaze, e.g., to minimize tracking uncertainty [6]. However, in computer vision cognitively-driven attentional mechanisms remain largely unexplored.

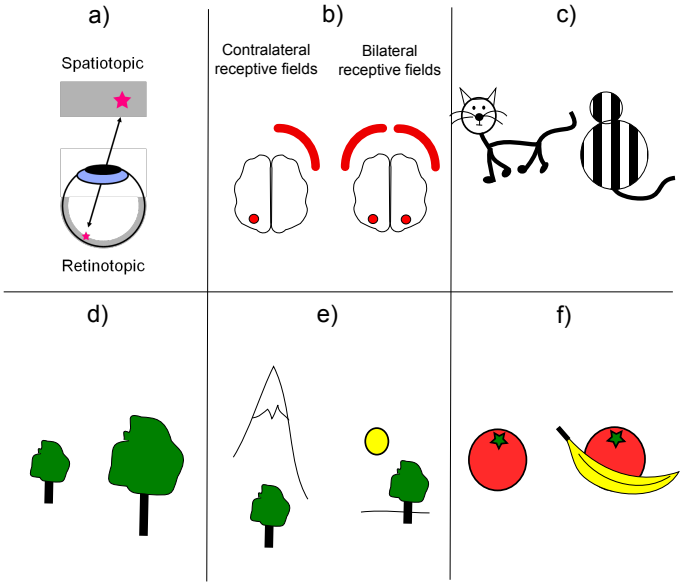
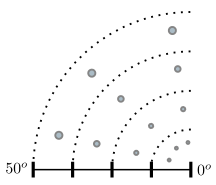


Fig. 3. Summary of table 1 concepts: a) Retinotopic (rt) and spatiotopic (st) organization; b) Contra- (co) versus bilateral (bl) receptive fields; c) Cue Invariance (CI); d) Size Invariance (SI); e) Position Invariance (PI); f) Occlusion Invariance (OI)

3.2 Ganglion Cells and LGN



From the photoreceptors of the retina information is passed through ganglion cells and LGN to the primary visual cortex. The left LGN receives input of the right visual hemifield from both eyes, and the right LGN receives input of the left visual hemifield from both eyes. However, the information from the two eyes remains still entirely separate in six different neuronal layers (four P- plus two M-layers, three layers receive input from the left eye, the other three layers from the right eye) of the LGN; no binocular integration is done at this level. Regarding spatial analysis, there are no significant differences between retinal ganglion cells and their LGN counterparts (there is even almost a one-to-one correspondence between retinal ganglion and LGN cells [95]). In motion analysis, LGN ganglion cells have lower optimal temporal frequencies, 4–10 Hz vs. 20–40 Hz in retinal ganglion cells, which indicates the presence of some low-pass filtering over retinal ganglion cells [95]. The two prominent new features emerging at this level are center-surround receptive fields and color opponency. The visual cortex is also organized into layers, where most of the feedforward connections (i.e. connections to a higher stage in the hierarchy) originate from the superficial layers and most of the feedback connections originate from the deeper layers. However, virtually nothing is known about the role of these different cortical layers in stimulus processing.

3.2.1 Center-Surround Receptive Fields

Neurophysiological view: Luminance sensitive cells with a center-surround receptive field come in two types: on-center/off-surround cells are sensitive to a bright spot on a dark background; off-center/on-surround cells are sensitive to the inverse pattern. Both are insensitive to homogeneous luminance. These cells are magnocellular (M) neurons and are involved in the temporal analysis.

Functional view: Center-surround receptive fields can be modeled by a difference of Gaussians and resemble a Laplace filter as used for edge detection [68]. They thus emphasize *spatial change* in luminance. These cells are also sensitive to *temporal changes* and form the basis of motion processing. Notably, the transformation into a representation emphasizing spatial and temporal change is performed at a very early stage, immediately following the receptor level, before any other visual processing takes place.

Most of the current computer vision techniques also involve in the earliest stages gradient-like computations which are essential parts of detectors / descriptors such as SIFT, HOG/HOF, etc.

3.2.2 Single-Opponent Cells

Neuropsysiological view: Single-opponent cells are color sensitive and compute color differences, namely L-M (L for long wavelength and M for middle wavelength, symbol "-" stands for opponency) and S-(L+M) (S stands for short wavelength), thereby establishing the red-green and the blue-yellow color axes. They have a band-pass filtering characteristic for luminance (gray value) stimuli but a low-pass characteristics for monochromatic (pure color) stimuli. These cells are parvocellular (P) neurons and are somewhat slower but have smaller receptive fields, i.e. higher spatial resolutions, than the magnocellular neurons. They are particularly important for high acuity vision in the central visual field.

Functional view: Single-opponent cells can be modeled by a Gaussian in one color channel, e.g. L, and another Gaussian of opposite sign in the opposing color channel, i.e. -M. This results in low-pass filtering in each color channel. The color opponency provides some level of invariance to changes in brightness and is one step towards color constancy.

4 Generic Scene Representation in the Occipital Cortex

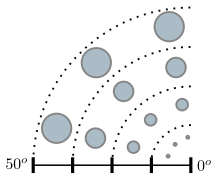
All areas in the occipital cortex (except MT) are organized retinotopically with orientation columns as basic units (see table 1, 6th column). MT is also organized retinotopically, but with depth and motion columns. Note that the visual system is not organized in a strictly sequential hierarchy but there are shortcuts between levels of the hierarchy. There is a stream V1 $\rightarrow$ V2 ($\rightarrow$ V3[5]) $\rightarrow$ V4 to the ventral pathway and another stream V1 $\rightarrow$ V2 $\rightarrow$ MT to the dorsal pathway (figure 2). However, there also exist cross connections between V4 and MT.

The latency of the visual signal increases with each level by approximately 10 ms, and the receptive field sizes increase gradually (see table 1, 3rd and 4th column). In general, the

5. Not much is known about the role of V3, therefore we have not given any detailed information in this paper about V3.

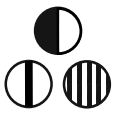
magnocellular pathway provides most of the input to the dorsal visual stream and the parvocellular pathway provides most of the information to the ventral pathway, but this is certainly not an absolute distinction.

4.1 Area V1



V1 is the first cortical area that processes visual information. Thus, the features it is sensitive to are more complex than in LGN but remain relatively simple: edges, gratings, line endings, motion, color, and disparity.

4.1.1 Edges, Bars, and Gratings



Neurophysiological view: V1 contains cells that respond preferentially to edges, bars, and gratings, i.e. linear oriented patterns. They are sensitive to the orientation of the patterns and, in case of gratings, to their spatial frequency (for a review, see [127]). Some cells are more sensitive to edges or single bars while others prefer gratings. There are two types of such cells, simple and complex cells. The former are sensitive to the phase of a grating (or exact position of a bar), the latter are not and have a larger receptive field.

Functional view: The original proposal by Hubel and Wiesel to achieve the phase-invariant orientation tuning characteristic of complex cells was simply to add the responses of simple cells along the axis perpendicular to their orientation, see [167] for a computational model. Later authors have attributed the behavior of complex cells to a MAX-like operation [48] (producing responses similar in amplitude to the larger of the responses pertaining to the individual stimuli – see, e.g., [141]) or to a nonlinear integration of a pool of unoriented LGN cells [115]. In computational models, simple cells can self-organize from natural images by optimizing a linear transformation for sparseness, i.e. only few units should respond strongly at any given time [125], or statistical independence [8] – however, it has been noted that linear models may not be sufficient for modeling simple cells [149]. Complex cells can be learned from image sequences by optimizing a quadratic transformation for slowness, i.e. the output of the units should vary as slowly over time as possible [38], [10]. On a more technical account it has been shown that Gabor wavelets are a reasonable approximation of simple cells while the magnitude of a Gabor quadrature pair resembles the response of complex cells [82]. Gabor wavelets have also been very successful in applications such as image compression [29], image retrieval [108], and face recognition [202]. In fact, it has been shown using statistics of images that Gabor wavelets (and the simple cells in V1) construct an efficient encoding of images [164].

4.1.2 Point Features

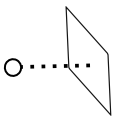
Neurophysiological view: V1 also contains cells that are sensitive to the end of a bar or edge or the border of a grating. Such cells are called end-stopped or hypercomplex [127].

Functional view: In V1, end-stopped cells might help to solve the aperture problem the system is faced with in motion as well as disparity processing (see section 4.1.3) since they can detect displacement also in the direction of an edge [127]. Like complex cells, hypercomplex cells can be learned from image sequences by optimizing slowness [10].

In computer vision, interest point detectors (which are not subject to the aperture problem due to the fact that they analyze local regions with occurrence of different orientations) of various kinds [107], [116] have been used since these features have turned out to be discriminative and stable, which is important for matching tasks and fundamental in many computer vision problems (pose estimation, object recognition, stereo, structure from motion, etc.). In this regard, it is interesting that V1 is dominated by detectors (simple and complex cells) for *linear* features (edges, bars, gratings). A possible reason might be that most meaningful features in natural scenes are actually edges which also allow for a complete reconstruction of the input signal (see, e.g., [39]).

The rather infrequent occurrence of neurons sensitive to point features at this low-level stage of visual processing suggests that primate vision does not necessarily rely on point features for bottom-up visual processing. Stereo and motion processing on the basis of edge and line features further suggests that the aperture problem is not solved by V1, but involves subsequent cortical layers for spatial integration.

4.1.3 Absolute Disparity



Neurophysiological view: V1 is the first area containing neurons that receive input from both eyes [84] (neurons in LGN are still monocular) and are able to compute disparity. In V1, this is still absolute disparity (i.e., the angular difference between the projections of a point onto the left and right retinas with reference to the fovea). Calculating disparity and thereby depth can be done in V1 without monocular contours in the image, as it is evident from our ease at interpreting random-dot stereograms [83]. There are also neurons in V1 that are sensitive to disparity in anticorrelated stereograms [26], in which the contrast polarity of the dots in one eye is reversed compared to the other eye. However, these neurons do not contribute to the depth perception and may have other functions.

Functional view: A prominent model for disparity estimation in V1 is the energy model, which is based on Gabor wavelets with slight phase or positional shifts [50]. Disparity is, of course, only one cue for depth perception, although an early one (in terms of processing and development, see [87]) and operational at close distance. On higher levels and at farther distances, cues such as occlusion, motion parallax etc. are used [84] which however are processed in higher-level areas of the primate brain's dorsal and ventral visual streams (see section 4.4). Also from a developmental perspective there are significant differences with pictorial depth cues developing only after approx. 6 months [87]. This is very much linked to the observation that statistics of natural scenes are linked to laws of perceptual organization, an idea first formulated by Brunswick [17] which has then later been confirmed computationally (see [200] for a review). This line of thought opens the perspective to formulate the problem of deriving

pictorial depth cues in computer vision systems as a statistical learning problem. Disparity is not only important for depth perception but also for gaze control [84], object grasping and object recognition. It has been shown that disparity tuned units can be learned from stereo images by maximizing mutual information between neighboring units, because depth is a feature that is rather stable across space [7].

In computer vision, stereo is a whole field of research, with many methods based on point features, which are convenient since their matches fix all degrees of freedom (see, e.g., [16]). However, there are approaches in computer vision that also use phase-differences of Gabor wavelets [49].

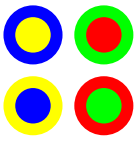
4.1.4 Local Motion

Neurophysiological view: Neurons in areas V1 and V2 are not only involved in static scene analysis but also in motion analysis. A fraction of simple and complex cells in V1 are direction selective, meaning that they respond only if the stimulus pattern (grating) moves in one direction and not the other [127]. However, only complex cells have spatio-temporal frequency tuning. The direction selective cells belong to the M-pathway and project mostly to area MT [118]. The aperture problem is not solved at that stage of processing.

Functional view: Estimating motion, or optic flow, is actually quite related to estimating disparity, since the latter can be viewed as a special case of the former with just two frames that are displaced in space rather then in time. The algorithms in computer vision as well as models of V1 are in general correspondingly similar to those discussed for estimating disparity (see section 4.1.3). For V1 (mainly simple cells), motion processing is usually conceptualized and modeled by spatiotemporal receptive fields [195], [179]. Complex cell-like units learned by optimizing slowness are motion direction selective, much like physiological neurons [10].

It is interesting to note that spatiotemporal features such as motion have been demonstrated to be the first features developmentally present in humans for recognizing objects (even sooner than color and orientation) [204].

4.1.5 Double-Opponent Cells

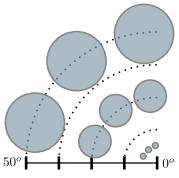


Neurophysiological view: About 5–10% of V1 cells are dedicated color-coding cells (for reviews see [23], [161]). In addition to single-opponent cells similar to those in LGN, which respond to local color (on a blue-yellow or red-green axis), V1 has double-opponent cells. These cells, whose existence used to be debated and is now supported with growing evidence (e.g., [22]), have a spatial-opponency structure within each color channel in addition to the opponency between different color channels. Such cells respond particularly well to a spot of one color on a background of its opponent color, and are thought to play a crucial role in perceptual color constancy. It is therefore not surprising that color contrast effects, i.e. a shift of perceived color of a stimulus away from the color of the background, have been observed in V1 [127]. The receptive fields of these cells are rarely circularly symmetric and therefore also show some orientation tuning, but their spatial resolution is low. Some double-opponent cells are also orientation selective. On the other hand, simple and complex cells, although not considered as coding color, are often sensitive to the orientation of equiluminant stimuli, i.e. edges or gratings defined only by color contrast and not luminance contrast. This shows that they are sensitive to color, but they do not code the color polarity but only orientation. We therefore see that color and form processing are largely (but not completely) separated in V1.

Functional view: Double-opponent cells form the basis of color contrast and color constancy, because they allow the system to take the color context into account in determining the perceived color [23]. It is interesting that double-opponent receptive fields can be learned from natural color images by optimizing statistical independence [20], which suggests that they are organized by an information optimization process and are therefore functionally driven.

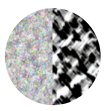
In contrast to low-level color normalization in computer vision, which is based primarily on operations applied the same way to each pixel (see, e.g., [47]), it is evident from human color perception that the achievement of color constancy involves local and global processes spanning all levels of the hierarchy, as already indicated by Helmholtz (see [199] and section 7.1).

4.2 Area V2



V2 is a retinotopically-organized area that mostly receives its input from V1. In V2, the segregation between M and P pathways is largely preserved although not complete [84]. Like V1, V2 contains cells tuned to orientation, color, and disparity. However, a fraction of V2 cells are sensitive to relative disparity (in contrast to absolute disparity arising in V1), which means that they represent depth relative to another plane rather than absolute depth. The main new feature of V2 is the more sophisticated contour representation including texture-defined contours, illusory contours, and contours with border ownership.

4.2.1 Texture-Defined and Illusory Contours



Neurophysiological view: Some V2 cells are sensitive to texture-defined contours, with an orientation tuning that is similar to that for luminance-defined contours [127]. V2 cells are also sensitive to illusory contours [84]. These can arise in various contexts, including texture or disparity discontinuities, or in relation to figure-ground effects such as the Kanizsa triangle (see icons at the left).[6]

Functional view: This is a step towards greater invariance of shape perception, since contours can be defined by a greater variety of cues.

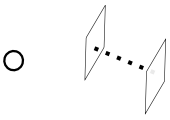
6. V1 also responds to illusory contours but has longer latencies and might be driven by feedback from V2.

4.2.2 Border Ownership

Neurophysiological view: Borders (i.e., contours) are mostly formed by the projections of two or more surfaces that either intersect or have gap between them in 3D. In most cases, such borders belong only to one of the surfaces that meet at the border, and border ownership pertains to the assignment of which surface (or region) a border belongs to. Border ownership was already identified as an important visual information by [90], although with a different term, *belongingness*. Border ownership, which was largely neglected in computational approaches to vision, is especially crucial for diffusion and filling-in mechanisms with which missing and ambiguous visual information can be reduced and rectified to a great extent. Discovery of cells sensitive to border ownership was quite recent. In 2000, Zhou et al. [206] found that 18% of the cells in V1 and more than 50% of the cells in V2 and V4 (along the ventral pathway) respond or code according to the direction of the owner of the boundary. However, the mechanisms by which neurons determine the ownership is largely unclear.

Functional view: The fact that border ownership sensitive neurons differentiate the direction of the owner 10–25 ms after the onset of the response and that border ownership sensitivity emerges as early as V1 (although to a lesser extent) suggests that border ownership can be determined using local cues that can be integrated by lateral long-range interactions along a boundary. However, as shown recently by Fang et al. [43], the process might also be modulated or affected from higher-level cortical areas with attention.

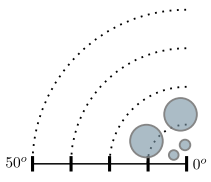
4.2.3 Relative Disparity



Neurophysiological view: V2 also includes disparity-sensitive cells. However, contrary to disparity-sensitive cells in V1, those in V2 are sensitive to relative disparity, which is the difference between the absolute disparities of two points in space. Relative disparity is for example the difference in disparity between a point at the fixation plane (zero disparity) and a point closer to the observer (near disparity). It is known that stereopsis relies mostly on the processing of relative disparity [130].

Functional view: With sensitivity to relative disparity in V2, it becomes possible to compare depth of objects and reason about their 3D spatial relationships.

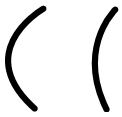
4.3 Area V4



In contrast to MT (see section 4.4) which seems to be dominated by M-pathway input, V4 seems to combine input from the M as well as the P pathway since blocking either M or P pathway reduces activity of most cells in V4 [84].

V4 neurons respond selectively to orientation, color, disparity and simple shapes. They continue the process of integrating lower-level into higher-level responses and increasing invariances. For instance, V4 cells respond to contours defined by differences in speed and/or direction of motion with an orientation selectivity that matches the selectivity to luminance-defined contours [127] (a few such cells are also found in V1 and V2 but with longer latencies, which again suggests that they are driven by feedback from V4). Prominent new features in V4 are curvature selectivity and luminance-invariant coding of hue.

4.3.1 Curvature Selectivity



Neurophysiological view: Some V4 cells are tuned to contours with a certain curvature (with a bias towards convex contours [131]) or vertices with a particular angle [127]. This selectivity is even specific to the position of the contour segment relative to the center of the shape considered, thus yielding an object centered representation of shape. V2 also has cells that respond to curves (contours that are not straight lines), but their response can be explained by their tuning to edges alone, which is not the case for V4 neurons.

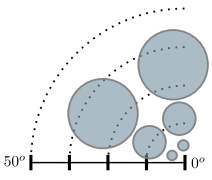
Functional view: Experiments in monkeys where area V4 was ablated showed that V4 is important for the perception of form and pattern/shape discrimination. V4 neuronal responses represent simple shapes by a population code that can be fit by a curvature-angular position function [131]. In this representation, the object's curvature is attached to a certain angular position relative to the object center of mass. Most V4 neurons represent individual parts or contour fragments.

4.3.2 Color Hue and Luminance Invariance

Neurophysiological view: Color coding cells in V4 differ from those in V2 in that they code for hue, rather than color opponency along the two principal color axes, and that the tuning to hue is invariant to luminance [24]. Even though specialized to color, many of these cells also show a prominent orientation tuning.

Functional view: Luminance invariant tuning to hue is already a form of color constancy, and the orientation tuning of color coding cells indicates some level of integration between color and form perception, although V4 neurons are clearly segregated into two populations, one for color and one for form processing [177].

4.4 Area MT

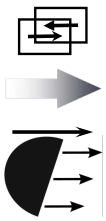


The middle temporal (MT) area is dedicated to visual motion and binocular depth processing. The vast majority of neurons in area MT are sensitive to moving stimuli. Neurons are tuned to direction and speed of motion [112]. Receptive fields are about 10 times larger than in V1 so that MT neurons integrate a set of motion signals from V1 over a larger area. The receptive fields show characteristic substructures of different motion sensitivity in different parts of the receptive field [106]. Many MT neurons are also sensitive to binocular disparity [30]. Activity in MT directly relates to perceptual motion [152] and depth [14] judgments. Area MT is retinotopically organized with motion and depth

columns similar to orientation and ocular dominance columns in V1.

MT is not only important for perception but also for motor control, particularly for smooth pursuit eye movements. MT together with MST provides the main velocity signal in the feedback control loop [37], [94] through output connections into oculomotor structures in the brain stem.

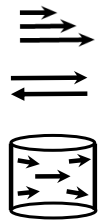
4.4.1 2D motion



Neurophysiological view: MT neurons compute a mid-level representation of motion by combining inputs from V1 neurons that respond to local motion [165], [118]. Some MT cells solve the aperture problem and encode the direction of motion independent of the orientation of the moving stimulus [117]. MT cells encode the speed rather than spatiotemporal frequency as V1 cells do [133]. In calculating motion signals, MT neurons follow a coarse-to-fine strategy in which responses to moving stimuli are fast, but imprecise, and become more refined over time [129].

Functional view: After initial measurements of local spatiotemporal energy (in V1), the combination of motion measurements is required to solve the aperture problem, derive 2D motion direction, and estimate speed. This results in a mid-level representation of motion in the visual field that is more faithful to the true motion and more robust against noise than earlier visual areas such as V1 and V2. The spatial smoothing that is inherent in the combination of motion over large receptive fields is partially reduced by disparity information in the combination of motion signals [99].

4.4.2 Motion Gradients and Motion-Defined Shapes



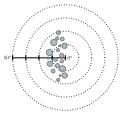
Neurophysiological view: Some MT cells are selective to higher order features of motion such as motion gradients, motion-defined edges, locally opposite motions, and motion-defined shapes [127]. These selectivities are aided by disparity sensitivity. Disparity helps to separate motion signals from objects at different distances, retain motion parallax and compute transparent motion and three-dimensional motion surfaces.

Functional view: MT constructs a representation of motion-defined surfaces and motion on surfaces.

5 Object Recognition and Categorization: The Ventral Stream

Lesion studies have demonstrated that the ventral pathway is critical for object discrimination [193], whereas the posterior parietal cortex is important for spatial vision. The most widely used partitioning of the inferior temporal cortex (IT) is between the more posterior part, TEO, and the more anterior part, area TE, based on the presence of a coarse retinotopy in TEO but not in TE (see table 1) as well as a larger receptive field size of neurons in the latter area over the former.[7] Two types of neurons have been identified in IT [174]: Primary cells respond to simple combinations of features and are a majority in TEO; Elaborate cells respond to faces, hands and complex feature configurations and have a high presence in area TE.

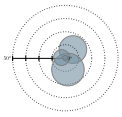
5.1 Area TEO



Neuropsysiological view: TEO (also known as PIT for *Posterior IT*) neurons are orientation- and shape-selective. It has been shown that TEO neurons mostly respond to very simple shape elements. The main difference between TEO and TE is the coarse retinotopic organization in TEO, which is absent in TE. The receptive fields of TEO neurons are still relatively small (3-5 deg) and located around the fovea or in the contralateral hemifield.

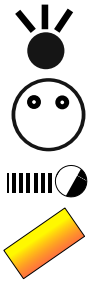
Functional view: TEO is responsible for medium complexity features and it integrates information about the shapes and relative positions of multiple contour elements. TEO integrates contour elements but with a higher degree of complexity over V4. This integration is non-linear and it includes inhibitory inputs (in addition to the excitatory ones). Shape tuning is position and size invariant, and it supports part-based shape theories [127].

5.2 Area TE



Neuropsysiological view: Area TE (also known as AIT for *Anterior IT*) can be characterized by a marked increase in the complexity of the visual features that drive the neurons with respect to the previous areas in the ventral pathway (Sec. 4). It is suggested that shape-selective TE neurons integrate the output from the previous areas. The receptive fields of visual neurons in TE range from 10 to 20 degrees of visual angle, and the average response latencies are around 70–80 ms.

Although 2D shape is the primary stimulus dimension to which TE neurons respond, other object attributes are encoded in TE as well: color [183], disparity [183], texture [183], and 3D shape [81]. At least for color and 3D shape it has been demonstrated that the processing of these object properties is largely confined to specific subregions in TE [80], [184].



Tanaka and co-workers [174] made a critical contribution by developing the stimulus-reduction method (see figure 4). After having measured the responses of TE neurons to real-world objects, they systematically reduced the image of the most effective object in an effort to identify the critical feature to which the TE neurons were responding. For many TE neurons, the critical feature was moderately complex, i.e. less complex than the entire image but more complex than simple bars or spots (figure 4). In some cases, the neurons driven by the critical features were clustered in what might be considered cortical columns [183]. These findings have led to the hypothesis that TE neurons do not explicitly code for entire objects but only for object parts. Therefore, the read-out of TE needs to combine

7. Many more functional subdivisions have been proposed for IT, including separate regions encoding information about faces, color or 3D shape, but the correspondence with the anatomical subdivisions is unclear at present.

information from many TE neurons to build an explicit object representation.

Functional view: Many properties of TE neurons (e.g. invariances, see table 1, 7th column) correspond well with the properties of visual object recognition. Several studies have demonstrated that the trial-to-trial variations in the firing rate of TE neurons correlate with the perceptual report of rhesus monkeys in various tasks, including object recognition [119], color discrimination [111] and 3D shape discrimination [196].

A neural system capable of object recognition has to fulfill two seemingly conflicting requirements, i.e. selectivity and invariance. On the one hand, neurons have to distinguish between different objects in order to provide information about object identity (and object class in the case of categorization) to the rest of the system, by means of sensitivity to features in the retinal images that discriminate between objects. On the other hand, this system also has to treat highly dissimilar retinal images of the same object as equivalent, and must therefore be insensitive to transformations in the retinal image that occur in natural vision (e.g. changes in position, illumination, retinal size, etc.). This can be achieved by deriving invariant features that are highly robust towards certain variations by discarding certain aspects of the visual data (as, e.g., SIFT descriptors [107]). From a systematic point of view, it would be however advantageous to not discard information, but to represent the information such that the aspects that are invariant are separated from the variant parts such that both kinds of information can be used efficiently (see, e.g., [31] and section 8.2).

TE neurons generally show invariance of the shape preference to a large range of stimulus transformations (though in general not in the absolute response levels). The most widely studied invariances of TE neurons include invariance for position (PI, cf. table 1, 7th column) and size (SI), but other stimulus transformations can also evoke invariant shape preferences: the visual cue defining the shape (cue invariance CI; [183]), partial occlusion (occlusion invariance OI; [183]), position-in-depth [79], illumination direction [92] and clutter (overlapping shapes, [183]). Rotation in depth evokes the most drastic changes in the retinal image of an object, and also the weakest invariance in TE, since most TE neurons show strongly view-dependent responses even after extensive training. The only exception might be faces, for which both view-dependent and view-invariant responses have been documented [183].

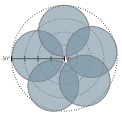
TE neurons typically respond to several but not all exemplars of the same category, and many TE neurons also respond to exemplars of different categories [198]. Therefore object categories are not explicitly represented in TE. However, recent *readout* experiments have demonstrated that statistical classifiers (e.g. support vector machines) can be trained to classify objects based on the responses of a small number of TE neurons [183], [88]. Therefore, a population of TE neurons can reliably signal object categories by their combined activity.[8] It is surprising that relatively little visual training has noticeable physiological effects on visual perception, on a single cell level as well as in fMRI [93]. For instance morphing objects into each other increases their perceived similarity, which is thought to be a useful mechanism for learning invariances [51].

Fig. 4. TE neurons respond to critical features of objects that can be quite complex; more complex than edges or bars but less complex than objects [174].

6 Vision for Action: The Dorsal Stream

The dorsal visual stream (see Figure 2) contains a number of areas that receive visual information from areas such as MT and V3A, and project mostly to the premotor areas in the frontal lobe, bridging between the visual and motor systems. The areas located in the dorsal stream are functionally related to different effectors: LIP is involved in eye movements, MIP in arm movements, AIP in hand movements (grasping) and MST and VIP in body movements (self-motion).[9]

6.1 MST



Neurophysiological view: Area MST receives its major input from area MT (see figure 2). Like MT, MST has many neurons that respond to visual motion. Receptive fields in MST are much larger than those of MT, often covering substantial portions of the visual field without a clear retinotopic arrangement. Many MST neurons respond selectively to global motion patterns such as large-field expansions or rotations [176]. Thus, MST neurons integrate motion in different directions from within the visual field. The structure of the receptive fields, however, is very complex and often not intuitively related to the pattern selectivity [34]. MST neurons are tuned to the direction of self-motion, or heading, in an optic flow field [132], [100]. MST neurons carry disparity signals [148] and receive vestibular input [15], [67] both consistent with their involvement in self-motion estimation.

Area MST is also involved in smooth pursuit eye movement [37], where it employs non-visual (extraretinal) input [122]. Using this extraretinal information, some MST neurons cancel the retinal effects of eye movements and respond to motion in the world rather than to motion on the retina [41]. This is also seen in Area V3A [57].

Functional view: Area MST is concerned with self-motion, both for movement of the head (or body) in space and

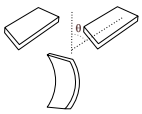
8. In contrast, an explicit category representation is present in the prefrontal cortex [53] and surprisingly also in the posterior parietal cortex (area LIP, [52]). Category information even occurs earlier, is stronger and more reliable in parietal cortex than in the prefrontal cortex [169].

9. Note that since not much is known about area 7a we have not discussed this area in detail.

movement of the eye in the head. The selectivity of MST neurons to optic flow patterns generates a population-based map of heading in MST [100]. Rather than representing the distribution of particular features in a retinotopic map of the visual field, as lower areas such as V1, V2, V4 or MT do, MST creates a new reference frame that represents self-motion in different directions in space. The organization is not retinotopic, but heading is represented in retinal coordinates, i.e, left or right with respect to the direction of gaze. The access to extraretinal eye movement information enables MST to estimate heading during combinations of body movement and eye movement.

The estimation of self-motion from optic flow is a common requirement in robotics. Solutions to this problem rely on the combination of many motion signals from different parts of the visual field as well as from non-visual areas relevant for heading estimation.

6.2 Caudal Intraparietal Area (CIP)

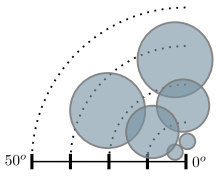


Neuropsysiological view: CIP[10] receives strong projections from area V3A and projects to LIP and AIP [121]. In [163], [170] it was reported that CIP neurons respond selectively to tilted planar surfaces defined by binocular disparity (first-order disparity). Some CIP neurons are also selective for the 3D orientation of elongated stimuli [151]. CIP neurons can show cue invariance for the tilt of planar surfaces, which means that the preference for a particular tilt is preserved when different depth cues signal the tilt (disparity, texture and perspective [190]). Results [188] suggest selectivity for zero-order disparity (position in depth) of CIP neurons. More recently, [86] also reported selectivity for curved surfaces (second-order disparity) in one monkey. CIP neurons do not respond during saccadic eye movements. No data exist on the size and shape of the CIP receptive fields nor on response latencies of CIP neurons.

Functional view: It is convenient to make a distinction between different orders of depth information from disparity [71]. Zero-order disparity refers to position-in-depth of planar surfaces (or absolute disparity, no disparity variation along the surface, see section 7.3) first-order disparity refers to inclined surfaces (tilt and slant, linear variations of disparity along the surface), and second-order disparity refers to curved surfaces (concave or convex, a change in the variation of disparity over the surface). CIP contains neurons that encode zeroth-, first- and possibly second-order disparities, which suggests that it is an important visual intermediate area that may provide input to visuomotor areas such as LIP and AIP. Not much is known about about the internal organization of CIP.

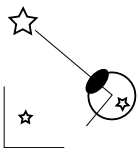
10. Note that since receptive field sizes of CIP neurons are unknown we have not drawn a corresponding figure as for the other regions.

6.3 Lateral Intraparietal Area (LIP)



Neurophysiological view: LIP is situated between visual areas and the motor system, receiving information from the dorsal and the ventral stream and projecting to other oculomotor control centers in the frontal lobe (FEF) and the superior colliculus [103]. LIP neurons respond before saccadic eye movements into the receptive field, and electrical microstimulation of LIP can evoke saccadic eye movements [181].

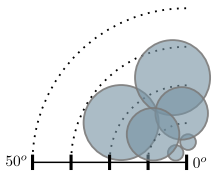
The visual responses in LIP are related to the salience of the stimulus [65], which led to the suggestion that LIP contains a salience map of the visual field, that guides attention and decides about saccades to relevant stimuli [11]. Moreover, LIP has been implicated in several other cognitive processes: decision formation [160], reward processing [136], timing [76] and categorization [52]. A more recent series of studies has also demonstrated that LIP neurons can respond selectively to simple two-dimensional shapes during passive fixation [156], a property that had been primarily allocated to the ventral visual stream.



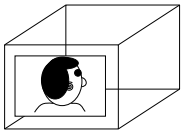
Functional view: The representation of space in LIP exemplifies several key properties of spatial processing in the dorsal stream. LIP neurons have visual receptive fields that represent locations on the retina, i.e. they represent stimuli in a retino-centric coordinate system. However, a few milliseconds before a saccadic eye movement, some LIP neurons become sensitive to stimuli at locations where their receptive field will be after the saccade [36]. This remapping of activity between the current and the future receptive field seems like a transient shift of the receptive field before a saccade. Moreover, although LIP receptive fields are basically in retino-centric coordinates, the activity of the cells is modulated by eye position, i.e., some cells respond more strongly to stimuli in their receptive field when the animal looks to the right than when it looks to the left, and vice versa [4]. The combination of retino-centric receptive fields and eye position modulation provides a population code in LIP that can represent the location of a stimulus in head-centric coordinates, i.e. can perform a coordinate transformation [207], [137]. This transformation allows, for example, for a combination of visual with auditory spatial input for the localization of sights and sounds [3].

LIP is one of the most studied areas in the dorsal stream. Despite more than two decades of single-cell studies, a considerable controversy exists with respect to the role of area LIP in high-level cognitive control processes such as motor planning, attention, decision formation, etc. However, LIP is believed to be a core area for spatial representation of behaviorally relevant stimuli. Visual (and auditory) input is transformed into a spatial representation in which each neuron uses eye-centered coordinates but in which the entire population forms a head-centric representation that encodes stimulus location even when the eye position changes. At the single neuron level, remapping of activity across saccades ensures continuity of the visual representation despite the eye movement.

6.4 Ventral Intraparietal Area (VIP)

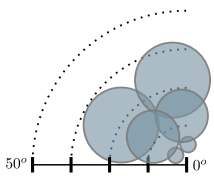


Neuropsysiological view: Area VIP is connected with a wide range of visual, somatosensory, and premotor (mouth representation) areas. VIP neurons are multi-modal, in the sense that they can be activated by visual, tactile, vestibular and auditory stimulation, and smooth pursuit eye movements [21]. The tactile receptive fields are generally located on the skin of the head and face, and visual and tactile receptive fields frequently match in size and location: a neuron that responds to tactile stimulation of an area around the mouth will also respond to visual stimuli approaching the mouth. It has been proposed that VIP encodes near-extrapersonal space [21]. The receptive fields of VIP neurons vary from purely retinocentric to purely head-centered [35], including also receptive fields that are intermediate between retinocentric and head-centered. Furthermore, some VIP neurons respond to complex motion stimuli, such as the direction of heading in optic flow displays.



Functional view: Area VIP is likely to be involved in self-motion, control of head movements, and the encoding of near-extrapersonal (head-centered) space which link tactile and visual fields.

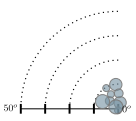
6.5 Medial Intraparietal Area (MIP)



Neuropsysiological view: MIP mainly projects to the dorsal premotor cortex (PMd). Neurons in this area typically respond selectively during a delayed reach task, in which monkeys are instructed to reach to a target on a touch screen after a certain time delay in order to receive a reward. MIP neurons will respond to particular reaching directions but not to others, and this neural selectivity is primarily eye-centered. When monkeys are free to choose the target, the MIP and PMd show increased spike-field coherence, suggesting direct communication between these brain areas [134].

Functional view: The activity of MIP neurons mainly reflects the movement plan towards the target, and not merely the location of the target or visual attention evoked by the target appearance [55]. MIP neurons also respond more when the animal chooses a reach compared to when the animal chooses a saccade towards a target, indicating that MIP encodes autonomously selected motor plans [25].

6.6 Anterior Intraparietal Area (AIP)



Neuropsysiological view: The main inputs to AIP arise in LIP, CIP and the ventral pathway [12], whereas the output from AIP is directed towards the ventral premotor area F5, which is also involved in hand movements. Reversible inactivation of AIP causes a profound grasping deficit in the contralateral hand [56]. Sakata and co-workers showed that AIP neurons frequently discharge during object grasping [151], with a preference for some objects over other objects. Some AIP neurons respond during object fixation and grasping, but not during grasping in the dark (visual-dominant neurons), other AIP neurons do not respond during object fixation but only when the object is grasped, even in the dark (motor-dominant neurons), whereas a third class of AIP neurons responds during object fixation and grasping, and during grasping in the dark (visuo-motor neurons, [120]). AIP encodes the disparity-defined 3D structure of curved surfaces [168]. However, experiments with monkeys indicate that the neural coding of 3D shape in AIP is not related to perceptual categorization of 3D shape [196]. In contrast, most 3D-shape-selective AIP neurons also respond during object grasping [180], suggesting that AIP represents 3D object properties for the purpose of grasping (i.e., grasping affordances).

Functional view: Neurons in AIP are sensitive to the 2D and 3D features of the object and shape of the hand (in a light or dark environment) relevant for grasping. In other words, area AIP might be involved in linking grasping affordances of objects with their 2D and 3D features. The extraction of grasping affordances from visual information is also currently a highly researched area in robotics since picking up unknown objects is a frequent task in autonomous and service robotics.

7 The Vertical View: Processing of Different Visual Modalities

Based on the knowledge we gained in sections 3 – 6 on the brain areas involved in the processing of visual information, we can now summarize the processing of different visual modalities such as color (section 7.1), 2D and 3D shape (section 7.2 and 7.3), motion (section 7.4) as well as the processing for object recognition (section 7.5) and actions (section 7.6) in a 'vertical view', emphasizing the hierarchical aspects of processing of visual information. Figure 5 gives an overview of this vertical (per modality) as well as the horizontal (per area) view.

7.1 Color

Color can be an extremely informative cue and has always been used as one of the basic features in psychophysical visual search experiments. Efficient search can be performed with heterogeneous colors (up to nine distractors) as soon as they are widely separated in color space [203].

Neurophysiologically color processing is characterized by a steady progression towards color constancy (see figure 5, 3rd column). The three cones types (L, M, S) have a broad and largely overlapping wavelength tuning, and their firing rate is heavily affected by luminance. The single-opponent cells in LGN establish the two color axes red-green and blue-yellow, thereby sharpening the wavelength tuning and achieving some invariance to luminance. Double-opponent cells provide the means to take nearby colors into account for color contrast. In V4 hue is encoded, which spans the full color space. The final step is IT where there exists an association of color with form [205]. In TEO (closer to V4) most of the neurons are activated

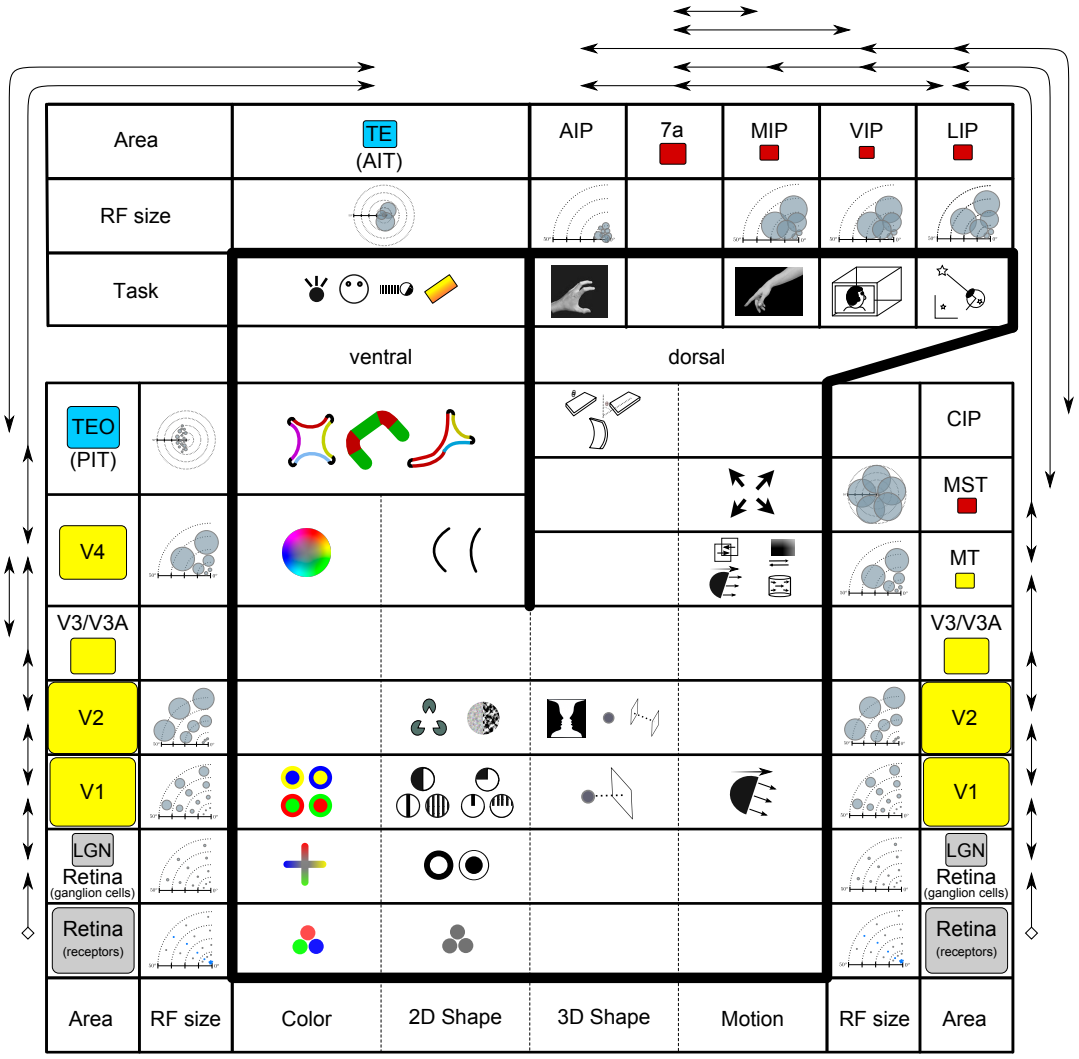


Fig. 5. Overview over the processing in the visual system. The icons of the text are arranged according to areas and modalities. The general layout follows figure 1, left. The yellow, blue and red rectangles have a size proportional to those of the corresponding areas. Connectivity is indicated in the vicinity, with the downwards arrowheads indicating the source area and the upwards arrowheads indicating the destination area. All but the retina → LGN connection are mirrored by feedback connections.

maximally by a simple combination of features such as bars or disks varying in size, orientation and color [175]. Elaborate cells (a majority in sub-area TE) respond to combinations of different features (shape and texture, shape and color, texture and color, texture and color and shape) [175].

There are a number of relevant insights that can be drawn from the neurophysiological evidence presented in the sections before. Color processing is taking place in a, to a large degree separated, pathway that only merges in general shape representations on the level of TE. Color is a cheap but also brittle feature for computer vision purposes. Its efficient use for object recognition depends on achieving color constancy which can still be seen as a challenge in computer vision applications. In the primate visual system, this is only achieved at rather late stages (V4 and beyond), hence involving a large part of the visual hierarchy. This is very different from color normalization schemes on a pixel level predominant in computer vision. A hierarchical representation might be able to provide means to provide mid– and high level cues for achieving color constancy.

7.2 Two-Dimensional Shape

Processing of 2D shape is characterized throughout the visual system by increasing receptive field sizes, increasing complexity of relevant features, and increasing degree of invariance (see figure 5, 4th column).

The receptive field sizes are tiny in the retina and can be as large as half the visual field in IT (see Table 1, second column). But it is not only the size that increases, the receptive fields also get more complex and dynamic. In the early areas, receptive fields tend to show a linear response. Beginning in V1, cells have a non-classical receptive field, i.e. the response

of these cells is modulated by the surrounding, which implies contextual effects. In V4, strong attentional effects have been shown, resulting in different responses of a cell for identical stimuli, if the task and thereby the attentional state changes [84]. In IT, receptive fields are very large but the selectivity can be influenced by clutter and other objects in the scene [147]. For isolated objects on a blank background, receptive fields are very large; for an object on a cluttered background or among many distractors, receptive fields are relatively small which indicates a tight connection between detection and segmentation.

The features that drive cells in the visual system also gradually increase in complexity. They are simply spots in retina and LGN, primarily bars or edges in V1, particular curves in V4, then more complex patterns and object parts in TEO and TE. The general notion is that the more complex features are built up from the simpler ones, e.g. simple cells that are sensitive to bars can be combined from on- and off-center cells that are sensitive to spots.

Early on does the visual system try to make responses invariant to frequently occurring but irrelevant variations [201]. That starts already in the retina where several mechanisms are in place to achieve a high degree of luminance invariance, so that we can see in the darkness of the night and the brightness of a sunny day. Some position invariance is first achieved in V1 by the complex cells, which are sensitive to bars and edges of certain orientations, like simple cells, but which are less sensitive to the position of the stimuli. This position invariance increases throughout the visual system and in IT, objects can be moved around by 10 degrees or even more without degrading the selectivity of some of the IT cells [185], [75]. There is also increasing size invariance. In addition to invariances to illumination and to geometrical transformations, invariance is also achieved with respect to the cues used to define objects (see table 1, 7^{th} column). Edges are the primary features used to represent objects, it seems. In V1 they are defined as boundaries between dark and light, or between different color hues; in V2 contours may also be defined by texture boundaries and these cells respond to illusory contours; in V4 contours may even be defined by differences in motion.

Representing and recognizing a 2D shape requires more than a collection of edges. The edges must be integrated somehow into one coherent percept. This is known in neuroscience as the binding problem [186], [84]. It is thought that there must be a mechanism that binds together the elementary features to one object, because otherwise one would mix the features of one object with those of another one and perceive something that is not there (this actually happens in humans in case of fast presentation times [187]). Possible solutions to the binding problem are tuning of cells to conjunctions of features, spatial attention, and temporal synchronization. The latter idea assumes that somehow the visual system manages to synchronize the firing of those neurons that represent the same object and desynchronize them from others [166], which could also explain the fairly limited number of objects we can process simultaneously. The binding problem is related to segmentation. Responses that represent also border ownership, like in V2, and responses that are specific to the relative position of an edge with respect to the object center, like in V4, are probably relevant for both processes.

7.3 Three-Dimensional Shape

The brain computes the third dimension (depth) from a large number of depth cues. Binocular disparity is one of the most powerful depth cues. Importantly, only second-order disparities (see section 6.2) are independent of eye position (vergence angle) and distance [71], thereby constituting a very robust parameter to estimate the three-dimensional layout of the environment.

The neural representation of 3D shape emerges gradually in the visual system (see figure 5, '3D shape' column). A few general principles can be identified. First, at progressively higher stages in the visual system, the neurons become tuned to more complex depth features, starting with absolute disparity in V1 [27]. Along the ventral stream, new selectivity emerges for relative disparity in V2 [182], first-order disparity in V4 [70] and finally second-order disparity in IT [70], [80]. Along the dorsal stream areas V3 and V3A encode primarily absolute disparities [5], area MT encodes absolute, relative and first-order disparity [97], [191], [123], area CIP encodes primarily first-order disparity [190], and AIP second-order disparities [168]. As with every other visual feature representation, the receptive fields of the neurons become larger and the latencies become longer. Secondly, at every level in the hierarchy the neural selectivity of the previous level(s) is reiterated such that at the highest levels in the hierarchy (e.g. IT cortex) selectivity for zero-, first- and second-order disparities can be measured [81].

Thirdly, in the visual hierarchy there seems to be a considerable amount of parallel processing of 3D shape information. Thus the end-stage areas of both the ventral and the dorsal visual stream (area AIP), each contain a separate representation of 3D shape [79], [168]. These representations are distinct because the properties of 3D-shape selective neurons differ markedly between IT and AIP: the coding of 3D shape in AIP is faster (shorter latencies), coarser (less sensitivity to discontinuities in the surfaces), less categorical and more boundary-based (less influence of the surface information) compared to IT [180], [77]. Finally, the two neural representations become more tailored towards the behavioral goal that the two processing streams support: in IT the 3D-shape representation subserves categorization of 3D shapes [197], but in AIP most 3D-shape selective neurons also respond during grasping [180]. In contrast, selectivity for anticorrelated disparities (in which each black dot in one eye corresponds to a white dot in the other eye and no depth can be perceived) is present in V1 [26], MT [96] and MST [171], weak in V4 [172] but absent in IT [78] or AIP [180], presumably because the latter areas are not involved in eye movements, which are strongly modulated by anticorrelated disparity [110].

7.4 Motion

The pattern of motion that is induced on the retina when one moves through the environment provides information about one's own motion and about the structure of the environment

[61]. The motion pathway extracts this information from the optic flow.

The first steps of motion analysis in V1 involve the computation of local spatiotemporal motion energy from the dynamics of the retinal image [195], [179]. A mid-level representation in area MT computes basic motion features such as 2D direction and speed based on the V1 inputs [165]. This computation needs to solve several difficult problems. First, local motion energy calculation by spatiotemporal receptive fields in V1 measures only the direction normal to the orientation of a moving bar or grating (the aperture problem). Secondly, spatiotemporal receptive fields cannot calculate speed but only spatiotemporal frequency. Speed tuning corresponds to orientation in spatiotemporal frequency. Most V1 neurons respond to a specific combination of spatial and temporal frequency whereas truly speed-tuned neurons respond to a preferred speed v over a range of spatial and temporal frequency s.t.: $v = df/dt$. Both problems are solved in MT by combining signals from many different V1 cells [165], [133]. However some complex cells in V1 have also been found to already solve these problems [128].

As indicated in figure 5 (6th column), the spatial integration that is needed to perform this integration leads to larger receptive fields in MT and thus has the effect of spatially smoothing the motion pattern. However, this smoothing is well adapted to the structure of the optic flow and preserves self-motion information [18]. Different weighting of inputs within the MT receptive fields moreover allows new motion features to be computed such as differential motion (differences between motion directions at adjacent positions in the visual field), motion edges, and gradients of the motion field [127]. These higher order motion signals are directly related to properties of surfaces in the scene. An important signal that carries this information is motion parallax, i.e. the difference in speed of two objects at different distances from a moving observer. The sensitivity of MT neurons to motion edges and locally opposite motion can be used to extract motion parallax from the optic flow. Motion processing is combined with disparity analysis in MT in order to separate motion signals from different depths [99].

The extraction of information about self-motion is a function of area MST. MST neurons have very large, bilateral receptive fields and respond to motion patterns. The patterns include expansion, contraction, rotation, and more generally speaking, spirals [176], [34]. One way to look at MST is thus in terms of pattern analysis. However, MST is better understood in terms of self-motion analysis [100]. Self-motion describes the translation and rotation of the eye of the observer in space, i.e. the 6 degrees of freedom of any rigid body motion. Single MST neurons are tuned to particular self motions, i.e. to particular translation directions (e.g. forward or rightward) and to rotations as well as to combinations of rotation and translation [100], [67]. MST thus contains a representation of self-motion.

Motion processing is linked to smooth pursuit eye movements. When one tracks a moving target with the eyes, the target is stable on the retina while the background is sweeping across the retina. The target, however, is perceived to move and the background is perceived as stable. Some cells in MST respond to motion in the world rather than motion on the retina, by combining visual information with extraretinal information about ongoing eye movements [41]. This combination of visual and extraretinal signals is also useful for self-motion analysis when one does not look in the direction of movement but fixates and tracks an object in the visual field [99]. Vestibular input about the state of self-motion is also combined with vision in MST [67].

In summary, the analysis of motion in the primate visual system proceeds in a hierarchy from V1 (local spatiotemporal filtering) to MT (2D motion) to MST (self-motion, motion in world coordinates). Along this hierarchy several computational problems are solved, the features become more complex, receptive fields become larger, and spatial integration of motion signals increases. The representation shifts from one of motion in the visual field (V1, MT) to one of motion in the world and motion of oneself in the world (MST). Also along this hierarchy, visual motion processing is combined with disparity (MT, MST), eye movement information (MST), and vestibular signals (MST). The representation becomes thus less tied to the image and more to the action of the body.

7.5 Object Recognition

Object recognition goes beyond simple 2D-shape perception in several aspects: integration of different cues and modalities, invariance to in-depth rotation and articulated movement, use of context. It is also important to distinguish between-class discrimination (object categorization) and within-class discrimination of objects.

Some integration of different cues is done already for 2D-shape perception. For instance, edges can be defined by luminance in V1, by textures in V2 and by differences in motion in V4. However, color and shape seems to be processed rather independently until high up in the hierarchy. Motion is processed early on, but it is used for object recognition in a different way than for shape perception. For instance, one can recognize familiar people from great distance by their characteristic gait. Other modalities, such as sound and odor, obviously also contribute to object recognition.

It appears that the units in IT pull together various features of medium complexity from lower levels in the ventral stream to build models of object parts. Precise granularity of these parts has not been established at present time, although there are indications that they span different sizes of receptive fields and are possibly tuned to different levels of feature invariance (abstraction) [174]. Computational models that can predominantly be described as compositional hierarchies (the hierarchical organization of categorical representations) define/learn units that are not inconsistent with these findings. For example, it has been shown that features that have been learned (in an unsupervised manner) at the level that roughly corresponds to IT contain sufficient information for reliable classification of object categories (this can be related to readout experiments [74]). Some of the related computational models could also help in making predictions regarding the need for massive feedback (from IT to LGN/V1) and alleviate the problems

with the stimulus-reduction method as these stimuli could be generated through a learning procedure [46].

Rotation in depth usually changes the shape of an object quite dramatically. However, a small fraction of IT neurons can exhibit some rotation invariance and speed of recognition of familiar objects does not depend on the rotation angle [79]. A particular case are face sensitive neurons which can show a rather large invariance to rotations in depth. Representations of the same object under different angles are presumably combined into a rotation invariant representation like simple cell responses might be combined into a complex cell response. Comparing unfamiliar objects from different perspectives seems to require mental rotation and requires extra time that is proportional to the rotation angle [154].

Context plays a major role in object recognition [124] and can be of different nature – semantic, spatial configuration or pose – and is, at least partially, provided by higher areas beyond IT. A simple example are the words 'THE' and 'CAT', which can be written with an identical character in the center with a shape somewhere between an 'H' and an 'A'. We recognize this very same shape immediately in the appropriate way depending on the context of the surrounding two letters. But we are also faster to recognize a sofa in a living room than floating in the air or a street scene. Interestingly, objects also help to recognize the context and context may be defined on a crude statistical level [124].

Some people have perfectly good object recognition capabilities but cannot recognize faces, a deficit known as *prosopagnosia*, although they can recognize people by their clothes or voices. The FFA (*fusiform face area*) seems the brain structure for face recognition [85]. There is evidence that prosopagnosia not only affects face processing but that it is a deficit in telling apart instances from the same category. For instance bird-watcher with prosopagnosia cannot tell birds apart anymore and car experts cannot make fine car distinctions [58].

It is interesting that in human subjects highly selective neurons have been described that may support object recognition. For example, recordings from epileptic patients in the medial temporal lobe have shown that single neurons reliably respond to particular objects, like the tower of Pisa, in whatever image [139].

7.6 Action Affordances

To supply visual information to the planning and control of action, the visual system extracts specific action-relevant features in hierarchical processing along the occipital and dorsal pathways. This processing is characterized by successively increasing complexity, multi-sensory integration, and a shift from general visual representations to representation specific for particular effectors and actions. Moreover, this processing is to some degree independent of conscious perception, such that lesion patients may be able to interact correctly with objects they fail to recognize and vice versa [64].

Early stages in the dorsal stream hierarchy (V1, V2, MT) are concerned with visual feature extraction (location, orientation, motion,) and the estimation of action-relevant objects features, such as surface orientation, from different cues (motion: MT, stereo: CIP). These features are encoded in a retinotopic frame of reference. Hierarchically higher areas encode information in spatiotopic or head-centric reference frames, sometimes at the single cell level (as in area VIP [35]) and often in a population code (areas MST, LIP, 7A, MIP) [137]. A major function of the dorsal stream thus lies in coordinate transformations.

These transformations are necessary because the planning of action with different effectors needs to consider targets in different reference frames. Eye movements are best encoded in a retinocentric representation but reach movements need a transformation to arm coordinates, and hence a representation of the target in space. It is not always clear what the best encoding for a particular action is, but the areas in the parietal cortex provide a number of parallel encodings for different tasks.

A further issue for these transformations lies in the combination of vision with other sensory or motor signals. Along the processing in the dorsal stream visual information is combined with vestibular (in MST, VIP), auditory (in LIP), somatosensory (in VIP), and proprioceptive or motor feedback signals (MST and VIP for smooth eye movements, LIP for saccades, MST/VIP/7A/MIP for eye position). Since these signals come in different sensory representations, the combination with vision requires extensive spatial transformations.

Eventually, higher areas in the dorsal stream construct spatial representations that are specialized to provide information for specific actions: LIP represents salience in the visual scene as a target signal for eye movements, MIP and AIP provide information for reaching (target signals) and grasping (shape signals). LIP and VIP provide information for the control of self-motion. Therefore, the processing of action-relevant visual information in the dorsal stream is characterized by a separation of functions, unlike processing in the ventral stream, which is focused on the perception of objects.

8 What Can We Learn from the Visual System for Computer Vision?

What can we learn from the primate visual system for computer vision systems as well as the learning of deep hierarchies? We believe that there are at least four design principles of the former that could be advantageous also for the latter: hierarchical processing[11], separation of information channels, feedback and an appropriate balance between prior coded structure and learning.

8.1 Hierarchical Processing

One prominent feature of the primate visual system is its hierarchical architecture consisting of many areas that can roughly be ordered in a sequence with first a common early processing and then a split into two interacting pathways, see Figure 2 and 5. Each pathway computes progressively more complex and invariant representations. What are the possible advantages of such an architecture?

11. In Introduction, we listed several authors who have in various ways studied and demonstrated this principle.

Computational efficiency: The brain is a machine with an enormous number of relatively simple and slow processing units, the neurons. Thus, significant performance can only be achieved if the computation is distributed efficiently. A visual hierarchical network does this spatially as well as sequentially. The spatial partitioning results in localized receptive fields, and the sequential partitioning results in the different areas that gradually compute more and more complex features. Thus, computation is heavily parallelized and pipelined. On a PC, this is less of an issue because it has only one or few but very fast processing units. However, this might change with GPUs or other computer architectures in the future and then the high degree of parallelization of hierarchical networks might be a real plus.

Computational efficiency in the primate visual system also arises from the fact that a lot of processing is reused for several different purposes. The occipital part, which constitutes most of the visual cortex, provides a generic representation that is used for object recognition, navigation, grasping, etc. This saves a lot of computation.

Learning efficiency: Equally important as the computational efficiency during the inference process is the learning efficiency. Hierarchical processing helps in that it provides several different levels of features that already have proven to be useful and robust in some tasks. Learning new tasks can build on these and can be fast because appropriate features at a relatively high level are available already. For instance invariance properties can simply be inherited from the features and do not have to be learned again.

Hierarchical processing, in particular in conjunction with the progression of receptive field sizes (see Table 1, column 3), offers mechanisms that may alleviate the overfitting problem. Namely, small size receptive fields in the lower hierarchical layers limit the potential variability of the features inside the receptive fields and consequently confine the units to learn low dimensional features, which can be sampled with relatively few training examples [46]. The process is recursively applied throughout the hierarchy resulting in a controlled progression in the overall complexity of units on the higher layers. This corresponds to an implicit regularization.

It is important to note that biological visual systems mature in complexity and sophistication in an intertwined process of development (through growing neural substrate) and learning (tuning of neural units) in a sequence of stages. From the computational point of view, this has important implications that deserve more attention in the future.

The world is hierarchical: Even within the brain is the visual system extreme in that has such a deep hierarchy. This may have to do with the complexity of the vision problem or the importance vision has for us. But it might also be a consequence of the fact that the (visual) world around us is spatially laid out and structured hierarchically. Objects can be naturally split into parts and subparts, complex features and simple features, which makes hierarchical processing useful. Nearby points in the visual field are much more related than distant points, which makes local processing within limited receptive fields effective at lower levels.

8.2 Separation of Information Channels

Another prominent feature of the visual system is the separation of information channels. Color, motion, shape etc. are processed separately, even in separate anatomical structures, for quite some time before they are integrated in higher areas. Some of these features are even duplicated in the dorsal and the ventral pathway but with different characteristics and used for different purposes. We believe this has at least two reasons: availability of information and efficiency of representation.

Availability of information: Depending on the circumstances, some of the information channels may not be available at all times. If we look at a photograph, depth and motion are not available. If it is dark, color is not available. If it is foggy high resolution shape information is not available, and motion and color might be the more reliable cues. A representation that would integrate all cues at once would be seriously compromised if one of the cues is missing. Separating the information channels provides robustness with respect to the availability of the different information cues.

Efficiency of representation: Separating the information channels naturally results in a factorial code; an integrated representation would yield a combinatorial code, which is known to suffer from the combinatorial explosion and also does not generalize well to new objects. If we represent four colors and four shapes separately, we can represent 16 different object more efficiently, i.e. with fewer units, than if we would represent each object as a unique color/shape combination. And also if we have seen only a few of the 16 possible combinations, we still can learn and represent unseen combinations easily.

It has been suggested that the binding problem, which arises because different neurons process different visual features of the same object (e.g. color and shape), is solved by means of neuronal synchronization in the temporal domain [40], [146]. In this 'binding by synchronization' hypothesis, neurons throughout the cortex encoding features of the same object would show synchronous activity, which would act as a 'label' that would indicate that the different features belong to the same object. However, experimental support for the synchronization hypothesis has been mixed [98], [33], [159], and no experiment has unambiguously proven that synchrony is necessary for binding.

8.3 Feedback

While we have outlined in this paper a hierarchical feedforward view on visual processing, it is important to remember that within the visual cortex there are generally more feedback connections than forward connections. Also lateral connections play an important role. This hints at the importance of processes like attention, expectation, top-down reasoning, imagination, and filling in. Many computer vision systems try to work in a purely feed-forward fashion. However, vision is inherently ambiguous and benefits from any prior knowledge available. This may even imply that the knowledge of how the tower of Pisa looks influences the perception of an edge on the level of V1. It also means that a system should be able to

produce several hypotheses that are concurrently considered and possibly not resolved [102].

8.4 Development and Learning of Visual Processing Hierarchies

In this paper, we focused on a description of and lessons to be learned from the end product, the functional visual system of the adult primate. We do not have the space here to discuss what is known about the development [194] and learning of biological visual processing hierarchies (e.g., [105], [113]). However, there are some fairly obvious conclusions relevant to computer vision.

First, in contrast to most deeply hierarchical computer vision systems, the primate visual processing hierarchy does not consist of a homogeneous stack of similar layers that are trained either bottom-up or in a uniform fashion. Rather, it consists of heterogeneous and specialized (horizontal) layers and (vertical) streams that differ considerably in their functions. Thus, a conceptually simple, generic vision system may not be achievable. It may be that biology has instead chosen to optimize specialized functions and their integration into a perceptual whole. It remains to be seen however, whether the specialization of cortical areas is due to fundamentally different mechanisms or to differences in the input and the particular combination of a very small set of learning principles (see, e.g, [31], [91]).

An aspect of these heterogeneous layers and streams that should be of interest to computer vision is that these distinct functional units and intermediate representations provide structural guidance for the design of hierarchical learning systems. As discussed by Bengio [9], this constitutes a way of decomposing the huge end-to-end learning problem into a sequence of simpler problems (see also p. 2).

Secondly, biological vision systems arise due to interactions between genetically-encoded structural biases and exposure to visual signals. One might argue that this is precisely how today's computer vision systems are conceived: The computational procedure is designed by hand, and its parameters are tuned using training data. However, inhomogeneous processing hierarchies require dedicated learning methods at various stages. Mounting evidence for adult cortical plasticity suggests that the influence of learning on cortical processing is much more profound than the tuning of synaptic strengths within fixed neural architectures [66], [62].

9 Conclusion

We have reviewed basic facts about the primate visual system, mainly on a functional level relevant for visual processing. We believe that the visual system still is very valuable as a proof of principle and a source of inspiration for building artificial vision systems. We have in particular argued for hierarchical processing with a separation of information channels at lower levels. Moreover, concrete design choices which are crucial for or potentially facilitate the learning of deep hierarchies (such as the structure of intermediate representations, the number of layers and the basic connectivity structure between layers) can be motivated from the biological model. Main stream computer vision, however, seems to follow design principles that are quite different from what we know from primates. We hope that the review and the thoughts presented here help in reconsidering this general trend and encourage the development of flexible and multi-purpose vision modules that can contribute to a hierarchical architecture for artificial vision systems.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 (Specific Programme Cooperation, Theme 3, Information and Communication Technologies) under grant agreement no. 270273, Xperience. We would also like to thank Michael D'Zmura for fruitful discussions.

REFERENCES

[1] Y. Amit. *2D Object Detection and Recognition: Models, Algorithms and Networks.* MIT Press, Cambridge, 2002.

[2] Y. Amit and D. Geman. A computational model for visual selection. *Neural Comp.*, 11(7):1691–1715, 1999.

[3] R. Andersen, A. Batista, L. Snyder, C. Buneo, and Y. Cohen. Programming to look and reach in the posterior parietal cortex. In M. Gazzaniga, editor, *The New Cognitive Neurosciences*, chapter 36, pages 515–524. MIT Press, 2 edition, 2000.

[4] R. Andersen and V. B. Mountcastle. The influence of the angle of gaze upon the excitability of the light-sensitive neurons of the posterior parietal cortex. *J. Neurosci.*, 3(3):532–548, 1983.

[5] A. Anzai, S. Chowdhury, and G. DeAngelis. Coding of stereoscopic depth information in visual areas v3 and v3a. *The Journal of Neuroscience*, 31(28):10270–10282, 2011.

[6] L. Bazzani, N. Freitas, H. Larochelle, V. Murino, and J.-A. Ting. Learning attentional policies for tracking and recognition in video with deep networks. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 937–944, New York, NY, USA, June 2011. ACM.

[7] S. Becker and G. E. Hinton. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

[8] A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Res.*, 37(23):3327–3338, 1997.

[9] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[10] P. Berkes and L. Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, 5(6):579–602, 2005. http://journalofvision.org/5/6/9/, doi:10.1167/5.6.9.

[11] J. W. Bisley and M. E. Goldberg. Attention, intention, and priority in the parietal lobe. *Annu. Rev. Neurosci.*, 33:1–21, 2010.

[12] E. Borra, A. Belmalih, R. Calzavara, M. Gerbella, A. Murata, S. Rozzi, and G. Luppino. Cortical connections of the macaque anterior intraparietal (aip) area. *Cerebral Cortex*, 18(5):1094, 2008.

[13] J. Bowmaker and H. Dartnall. Visual pigments of rods and cones in a human retina. *The Journal of Physiology*, 298:501–511, 1980.

[14] D. C. Bradley, G. C. Chang, and R. A. Andersen. Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature*, 392:714–717, 1998.

[15] F. Bremmer, M. Kubischik, M. Pekel, M. Lappe, and K.-P. Hoffmann. Linear vestibular self-motion signals in monkey medial superior temporal area. *Ann. N.Y. Acad. Sci.*, 871:272–281, 1999.

[16] M. Brown, D. Burschka, and G. Hager. Advances in computational stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(8):993–1008, 2003.

[17] E. Brunswik and J. Kamiya. Ecological cue–validity of 'proximity' and of other Gestalt factors. *American Journal of Psychologie*, LXVI:20–32, 1953.

[18] D. Calow, N. Krüger, F. Wörgötter, and M. Lappe. Biologically motivated space-variant filtering for robust optic flow processing. *Network*, 16(4):323–340, 2005.

[19] M. Carandini and D. J. Heeger. Normalization as a canonical neural computation. *Nature neuroscience*, 13:51–62, 2012.

[20] M. S. Caywood, B. Willmore, and D. J. Tolhurst. Independent components of color natural scenes resemble V1 neurons in their spatial and color tuning. *J. Neurophysiol.*, 91(6):2859–2873, Jun 2004.

[21] C. Colby and M. Goldberg. Space and attention in parietal cortex. *Annual Review of Neuroscience*, 22(1):319–349, 1999.

[22] B. Conway. Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (v-1). *The Journal of Neuroscience*, 21(8):2768–2783, 2001.

[23] B. R. Conway. Color vision, cones, and color-coding in the cortex. *Neuroscientist*, 15:274–290, Jun 2009.

[24] B. R. Conway, S. Moeller, and D. Y. Tsao. Specialized color modules in macaque extrastriate cortex. *Neuron*, 56:560–573, Nov 2007.

[25] H. Cui and R. Andersen. Posterior parietal cortex encodes autonomously selected motor plans. *Neuron*, 56(3):552–559, 2007.

[26] B. Cumming and A. Parker. Responses of primary visual cortical neurons to binocular disparity without depth perception. *Nature*, 389(6648):280–283, 1997.

[27] B. Cumming and A. Parker. Binocular neurons in v1 of awake monkeys are selective for absolute, not relative, disparity. *The Journal of Neuroscience*, 19(13):5602–5618, 1999.

[28] C. Curcio and K. Allen. Topography of ganglion cells in human retina. *The Journal of Comparative Neurology*, 300(1):525, 1990.

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

[30] G. C. De Angelis, B. G. Cumming, and W. T. Newsome. Cortical area MT and the perception of stereoscopic depth. *Nature*, 394:677–680, 1998.

[31] J. J. DiCarlo and D. D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, 2007.

[32] S. Dickinson. The evolution of object categorization and the challenge of image abstraction. In S. Dickinson, A. Leonardis, B. Schiele, and T. M., editors, *Object Categorization: Computer and Human Vision Perspectives*, pages 1–37. Cambridge Univ. Press Cambridge, UK, 2009.

[33] Y. Dong, S. Mihalas, F. Qiu, R. von der Heydt, and E. Niebur. Synchrony and the binding problem in macaque visual cortex. *Journal of Vision*, 8(7), 2008.

[34] C. J. Duffy and R. H. Wurtz. Sensitivity of MST neurons to optic flow stimuli. II. mechanisms of response selectivity revealed by small-field stimuli. *J. Neurophysiol.*, 65:1346–1359, 1991.

[35] J. Duhamel, F. Bremmer, S. BenHamed, and W. Graf. Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*, 389(6653):845–848, 1997.

[36] J. Duhamel, C. Colby, and M. Goldberg. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, 255(5040):90, 1992.

[37] M. R. Dürsteler and R. H. Wurtz. Pursuit and optokinetic deficits following chemical lesions of cortical areas MT and MST. *J. Neurophysiol.*, 60:940–965, 1988.

[38] W. Einhäuser, C. Kayser, P. König, and K. P. Körding. Learning the invariance properties of complex cells from their responses to natural stimuli. *Euro J Neurosci*, pages 475–486, February 2002.

[39] J. H. Elder. Are edges incomplete ? *International Journal of Computer Vision*, 34:97–122, 1999.

[40] A. Engel, P. Roelfsema, P. Fries, M. Brecht, and W. Singer. Role of the temporal domain for response selection and perceptual binding. *Cerebral Cortex*, 7(6):571–582, 1997.

[41] R. G. Erickson and P. Thier. A neuronal correlate of spatial stability during periods of self-induced visual motion. *Exp. Brain Res.*, 86:608–616, 1991.

[42] G. J. Ettinger. Hierarchical object recognition using libraries of parameterized model sub-parts. Technical report, MIT, 1987.

[43] F. Fang, H. Boyaci, and D. Kersten. Border ownership selectivity in human early visual cortex and its modulation by attention. *The Journal of Neuroscience*, 29(2):460–465, 2009.

[44] D. Felleman and D. V. Essen. Distributed hierarchical processing in primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.

[45] S. Fidler, M. Boben, and A. Leonardis. Similarity-based cross-layered hierarchical representation for object categorization. In *CVPR*, 2008.

[46] S. Fidler, M. Boben, and A. Leonardis. Learning hierarchical compositional representations of object structure. In S. Dickinson, A. Leonardis, B. Schiele, and T. M., editors, *Object Categorization: Computer and Human Vision Perspectives*, pages 196–215. Cambridge Univ. Press Cambridge, UK, 2009.

[47] G. Finlayson and S. Hordley. Color constancy at a pixel. *J. Opt. Soc. Am. A*, 18:253–264, 2001.

[48] I. M. Finn and D. Ferster. Computational diversity in complex cells of cat primary visual cortex. *Journal of Neuroscience*, 27(36):9638–9648, 2007.

[49] D. J. Fleet, A. D. Jepson, and M. R. Jenkin. Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198 – 210, 1991.

[50] D. J. Fleet, H. Wagner, and D. J. Heeger. Neural encoding of binocular disparity: energy models, position shifts and phase shifts. *Vision Res.*, 36(12):1839–1857, Jun 1996.

[51] M. Franzius, N. Wilbert, and L. Wiskott. Invariant object recognition and pose estimation with slow feature analysis. *Neural Computation*, 23(9):2289–2323, 2011.

[52] D. Freedman and J. Assad. Experience-dependent representation of visual categories in parietal cortex. *Nature*, 443(7107):85–88, 2006.

[53] D. Freedman, M. Riesenhuber, T. Poggio, and E. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316, 2001.

[54] K. Fukushima, S. Miyake, and T. Ito. Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Systems, Man and Cybernetics*, 13(3):826–834, 1983.

[55] A. Gail and R. Andersen. Neural dynamics in monkey parietal reach region reflect context-specific sensorimotor transformations. *The Journal of Neuroscience*, 26(37):9376–9384, 2006.

[56] V. Gallese, A. Murata, M. Kaseda, N. Niki, and H. Sakata. Deficit of hand preshaping after muscimol injection in monkey parietal cortex. *Neuroreport*, 5(12):1525–1529, 1994.

[57] C. Galletti, P. P. Battaglini, and P. Fattori. 'real-motion' cells in area v3a of macaque visual cortex. *Exp. Brain Res.*, 82:67–76, 1990.

[58] I. Gauthier, P. Skudlarski, J. C. Gore, and A. W. Anderson. Expertise for cars and birds recruits brain areas involved in face recognition. *Nat Neurosci*, 3(2):191–197, 2000.

[59] S. Geman. Hierarchy in machine and natural vision. In *Proceedings of the 11th Scandinavian Conference on Image Analysis*, 1999.

[60] S. Geman, D. Potter, and Z. Chi. Composition systems. *Quarterly of Applied Mathematics*, 60(4):707–736, 2002.

[61] J. Gibson. The perception of visual surfaces. *American Journal of Psychology*, 63(367–384.), 1950.

[62] C. Gilbert and W. Li. Adult visual cortical plasticity. *Neuron*, 75(2):250–264, 7 2012.

[63] I. Gödecke and T. Bonhoeffer. Development of identical orientation maps for two eyes without common visual experience. *Nature*, 379:251–255, 1996.

[64] M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends Neurosci.*, 15(1):20–25, 1992.

[65] J. P. Gottlieb, M. Kusunoki, and M. E. Goldberg. The representation of visual salience in monkey parietal cortex. *Nature*, 391:481–484, 1998.

[66] E. Gould, A. Reeves, M. Graziano, and C. Gross. Neurogenesis in the neocortex of adult primates. *Science*, 286(5439):548–552, 1999.

[67] Y. Gu, P. Watkins, D. Angelaki, and G. DeAngelis. Visual and nonvisual contributions to three-dimensional heading selectivity in the medial superior temporal area. *J. Neurosci.*, 26(1):73–85, 2006.

[68] M. Hawken and A. Parker. Spatial properties of neurons in the monkey striate cortex. *Proceedings of the Royal Society of London, series B, Biological Sciences*, 231:251–288, 1987.

[69] J. Hawkins and S. Blakeslee. *On Intelligence.* Times Books, 2004.

[70] D. Hinkle, C. Connor, et al. Three-dimensional orientation tuning in macaque area v4. *Nature Neuroscience*, 5(7):665–670, 2002.

[71] I. Howard and B. J. Rogers. *Seeing in depth, Vol. 1: Basic mechanisms.* University of Toronto Press, 2002.

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

[73] D. Hubel and T. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195(1):215–243, 1968.

[74] C. Hung, G. Kreiman, T. Poggio, and J. DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005.

[75] M. Ito, H. Tamura, I. Fujita, and K. Tanaka. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, 73(1):218–226, Jan. 1995.

[76] P. Janssen and M. Shadlen. A representation of the hazard rate of elapsed time in macaque area lip. *Nature Neuroscience*, 8(2):234–241, 2005.

[77] P. Janssen, R. Vogels, Y. Liu, and G. Orban. Macaque inferior temporal neurons are selective for three-dimensional boundaries and surfaces. *The Journal of Neuroscience*, 21(23):9419–9429, 2001.

[78] P. Janssen, R. Vogels, Y. Liu, and G. Orban. At least at the level of inferior temporal cortex, the stereo correspondence problem is solved. *Neuron*, 37(4):693–701, 2003.

[79] P. Janssen, R. Vogels, and G. Orban. Macaque inferior temporal neurons are selective for disparity-defined three-dimensional shapes. *Proceedings of the National Academy of Sciences*, 96(14):8217, 1999.

[80] P. Janssen, R. Vogels, and G. Orban. Selectivity for 3d shape that reveals distinct areas within macaque inferior temporal cortex. *Science*, 288(5473):2054–2056, 2000.

[81] P. Janssen, R. Vogels, and G. Orban. Three-dimensional shape coding in inferior temporal cortex. *Neuron*, 27(2):385–397, 2000.

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

[83] B. Julesz. *Foundations of cyclopean perception*. U. Chicago Press, 1971.

[84] E. R. Kandel, J. H. Schwartz, and T. M. Jessel, editors. *Principles of Neural Science*. McGraw-Hill, 4th edition, 2000.

[85] N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11):4302–4311, 1997.

[86] N. Katsuyama, A. Yamashita, K. Sawada, T. Naganuma, H. Sakata, and M. Taira. Functional and histological properties of caudal intraparietal area of macaque monkey. *Neuroscience*, 167(1):1–10, 2010.

[87] P. Kellman and M. Arterberry. *The Cradle of Knowledge*. MIT-Press, 1998.

[88] R. Kiani, H. Esteky, K. Mirpour, and K. Tanaka. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97(6):4296–4309, 2007.

[89] C. Koch. *Biophysics of Computation: Information Processing in Single Neurons.*. Oxford University Press, New York., 1999.

[90] K. Koffka. *Principles of Gestalt Psychology*. New York, 1955.

[91] P. König and N. Krüger. Perspectives: Symbols as self-emergent entities in an optimization process of feature extraction and predictions. *Biological Cybernetics*, 94(4):325–334, 2006.

[92] K. Köteles, P. De Maziere, M. Van Hulle, G. Orban, and R. Vogels. Coding of images of materials by macaque inferior temporal cortical neurons. *European Journal of Neuroscience*, 27(2):466–482, 2008.

[93] Z. Kourtzi and J. J. DiCarlo. Learning and neural plasticity in object recognition. *Curr. Opin. Neurobiol.*, 16:152–158, Apr 2006.

[94] R. J. Krauzlis and S. G. Lisberger. A model of visually-guided smooth pursuit eye movements based on behavioral observations. *J. Comput. Neurosci.*, 1(4):265–283, 1994.

[95] J. Kremers. *The primate visual system: a comparative approach*. Wiley, 2005.

[96] K. Krug, B. Cumming, and A. Parker. Comparing perceptual signals of single v5/mt neurons in two binocular depth tasks. *Journal of Neurophysiology*, 92(3):1586–1596, 2004.

[97] K. Krug and A. Parker. Neurons in dorsal visual area v5/mt signal relative disparity. *The Journal of Neuroscience*, 31(49):17892–17904, 2011.

[98] V. Lamme, H. Spekreijse, et al. Neuronal synchrony does not represent texture segregation. *Nature*, 396(6709):362–366, 1998.

[99] M. Lappe. Functional consequences of an integration of motion and stereopsis in area MT of monkey extrastriate visual cortex. *Neural Comp.*, 8(7):1449–1461, 1996.

[100] M. Lappe, F. Bremmer, M. Pekel, A. Thiele, and K. P. Hoffmann. Optic flow processing in monkey STS: a theoretical and experimental approach. *J. Neurosci.*, 16(19):6265–6285, 1996.

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

[102] T. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America*, 20(7):1434–1448, 7 2003.

[103] J. Lewis and D. Van Essen. Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *The Journal of Comparative Neurology*, 428(1):112–137, 2000.

[104] N. Li and J. DiCarlo. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron*, 67(6):1062–1075, 2010.

[105] S. Li, S. D. Mayhew, and Z. Kourtzi. Learning shapes spatiotemporal brain patterns for flexible categorical decisions. *Cerebral Cortex*, In Press, 2011.

[106] M. S. Livingstone, C. C. Pack, and R. T. Born. Two-dimensional substructure of MT receptive fields. *Neuron*, 30(3):781–793, 2001.

[107] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 2(60):91–110, 2004.

[108] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[109] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. Freeman, 1977.

[110] G. Masson, C. Busettini, and F. Miles. Vergence eye movements in response to binocular disparity without depth perception. *Nature*, 389(6648):283–286, 1997.

[111] T. Matsumora, K. Koida, and H. Komatsu. Relationship between color discrimination and neural responses in the inferior temporal cortex of the monkey. *Journal of Neurophysiology*, 100(6):3361–3374, 2008.

[112] J. H. R. Maunsell and D. C. van Essen. Functional properties of neurons in middle temporal visual area of the macaque monkey. I. selectivity for stimulus direction, speed, and orientation. *J. Neurophysiol.*, 49(5):1127–1147, 1983.

[113] S. D. Mayhew, S. Li, and Z. Kourtzi. Learning acts on distinct processes for visual form perception in the human brain. *J. Neurosci.*, 32(3):775–786, 2012.

[114] B. W. Mel and J. Fiser. Minimizing binding errors using learned conjunctive features. *Neural Computation*, 12(4):731–762, 2000.

[115] B. W. Mel, D. L. Ruderman, and A. K. A. Translation-invariant orientation tuning in visual "complex" cells could derive from intradendritic computations. *Journal of Neuroscience*, 18(11):4325–4334, 1998.

[116] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[117] J. A. Movshon, E. H. Adelson, M. S. Gizzi, and W. T. Newsome. The analysis of moving visual patterns. In C. Chagas, R. Gattass, and C. Gross, editors, *Pattern Recognition Mechanisms*, pages 117–151, New York, 1985. Springer.

[118] J. A. Movshon and W. T. Newsome. Visual response properties of striate cortical neurons projecting to area mt in macaque monkeys. *Journal of Neuroscience*, 16(23):7733–7741, 1996.

[119] R. Mruczek and D. Sheinberg. Activity of inferior temporal cortical neurons predicts recognition choice behavior and recognition time during visual search. *The Journal of Neuroscience*, 27(11):2825–2836, 2007.

[120] A. Murata, V. Gallese, G. Luppino, M. Kaseda, and H. Sakata. Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area aip. *Journal of Neurophysiology*, 83(5):2580–2601, 2000.

[121] H. Nakamura, T. Kuroda, M. Wakita, M. Kusunoki, A. Kato, A. Mikami, H. Sakata, and K. Itoh. From three-dimensional space vision to prehensile hand movements: the lateral intraparietal area links the area v3a and the anterior intraparietal area in macaques. *The Journal of Neuroscience*, 21(20):8174–8187, 2001.

[122] W. T. Newsome, R. H. Wurtz, and H. Komatsu. Relation of cortical areas MT and MST to pursuit eye movements. II. differentiation of retinal from extraretinal inputs. *J. Neurophysiol.*, 60(2):604–620, 1988.

[123] J. Nguyenkim and G. DeAngelis. Disparity-based coding of three-dimensional surface orientation by macaque middle temporal neurons. *The Journal of Neuroscience*, 23(18):7117–7128, 2003.

[124] A. Oliva and A. Torralba. The role of context in object recognition. *Trends Cogn. Sci.*, 11:520–527, 2007.

[125] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996.

[126] B. Ommer and J. M. Buhmann. Learning the compositional nature of visual objects. In *CVPR*, 2007.

[127] G. A. Orban. Higher order visual processing in macaque extrastriate cortex. *Physiol. Rev.*, 88:59–89, Jan 2008.

[128] C. Pack, M. S. Livingstone, K. Duffy, and R. Born. End-stopping and the aperture problem: Two-dimensional motion signals in macaque v1. *Neuron*, 39:671680, 2003.

[129] C. C. Pack and R. T. Born. Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature*, 409(6823):1040–1042, 2001.

[130] A. Parker. Binocular depth perception and the cerebral cortex. *Nature Reviews Neuroscience*, 8(5):379–391, 2007.

[131] A. Pasupathy and C. Connor. Responses to contour features in macaque area v4. *Journal of Neurophysiology*, 82(5):2490, 1999.

[132] M. Pekel, M. Lappe, F. Bremmer, A. Thiele, and K.-P. Hoffmann. Neuronal responses in the motion pathway of the macaque monkey to natural optic flow stimuli. *NeuroReport*, 7(4):884–888, 1996.

[133] J. A. Perrone and A. Thiele. Speed skills: measuring the visual speed analyzing properties of primate MT neurons. *Nat. Neurosci.*, 4(5):526–532, 2001.

[134] B. Pesaran, M. Nelson, and R. Andersen. Free choice activates a decision circuit between frontal and parietal cortex. *Nature*, 453(7193):406–409, 2008.

[135] R. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *International Journal of Neural Systems*, 45(18):2397–2416, 2005.

[136] M. Platt and P. Glimcher. Neural correlates of decision variables in parietal cortex. *Nature*, 400(6741):233–238, 1999.

[137] A. Pouget and T. J. Sejnowski. Spatial transformations in the parietal cortex using basis functions. *J. Cog. Neurosci.*, 9(2):222–237, 1997.

[138] N. Pugeault, F. Wörgötter, and N. Krüger. Visual primitives: Local, condensed, and semantically rich visual descriptors and their applications in robotics. *International Journal of Humanoid Robotics*, 7(3):379–405, 2010.

[139] R. Q. Quiroga, L. Reddy, C. Koch, and I. Fried. Decoding visual inputs from multiple neurons in the human temporal lobe. *J Neurophysiol*, 98(4):1997–2007, 2007.

[140] S. Raiguel, R. Vogels, S. Mysore, and G. Orban. Learning to see the difference specifically alters the most informative v4 neurons. *The Journal of Neuroscience*, 26(24):6589–6602, 2006.

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

[142] K. Rockland, J. Kaas, and A. Peters. *Cerebral Cortex: Extrastriate Cortex in Primates*, volume 12. Springer, 1997.

[143] A. Rodríguez-Sánchez, E. Simine, and J. Tsotsos. Attention and visual search. *International Journal of Neural Systems*, 17(4):275–288, 2007.

[144] A. J. Rodriguez-Sanchez and J. K. Tsotsos. The importance of intermediate representations for the modeling of 2d shape detection: Endstopping and curvature tuned computations. *Proc. IEEE Computer Vision and Pattern Recognition*, pages 4321–4326, 2011.

[145] A. J. Rodriguez-Sanchez and J. K. Tsotsos. The roles of endstopped and curvature tuned computations in a hierarchical representation of 2d shape. *PLoS ONE*, 7(8):1–13, 2012.

[146] P. Roelfsema. Solutions for the binding problem. *Zeitschrift für Naturforschung. C, Journal of biosciences*, 53(7-8):691, 1998.

[147] E. Rolls, B. Webb, and M. C. A. Booth. Responses of inferior temporal cortex neurons to objects in natural scenes. *Society for Neuroscience Abstracts*, 26:1331, 2000.

[148] J.-P. Roy and R. H. Wurtz. Disparity sensitivity of neurons in monkey extrastriate area MST. *J. Neurosci.*, 12(7):2478–2492, 1992.

[149] N. Rust, O. Schwartz, J. Movshon, and E. Simoncelli. Spike-triggered characterization of excitatory and suppressive stimulus dimensions in monkey v1. *Neurocomputing*, 58:793–799, 2004.

[150] U. Rutishauser, R. J. Douglas, and J. J. Slotine. Collective stability of networks of winner-take-all circuits. *Neural Computation*, 23(3):735–773, 2011.

[151] H. Sakata, M. Taira, M. Kusunoki, A. Murata, and Y. Tanaka. The parietal association cortex in depth perception and visual control of hand action. *Trends in Neurosciences*, 20(8):350–357, 1997.

[152] C. D. Salzman, K. H. Britten, and W. T. Newsome. Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, 346(12):174–177, 1990.

[153] F. Scalzo and J. H. Piater. Statistical learning of visual feature hierarchies. In *Workshop on Learning, CVPR*, 2005.

[154] S. C. E. Schendan H. E. Mental rotation and object categorization share a common network of prefrontal and dorsal and ventral regions of posterior cortex. *Neuroimage*, 35:1264–1277, 2007.

[155] A. Schoups, R. Vogels, N. Qian, and G. Orban. Practising orientation identification improves orientation coding in v1 neurons. *Nature*, 412(6846):549–53, 2001.

[156] A. Sereno and J. Maunsell. Shape selectivity in primate lateral intraparietal cortex. *Nature*, 395(6701):500–503, 1998.

[157] T. Serre and A. T. Poggio. Neuromorphic approach to computer vision. *Communications of the ACM (online)*, 53, 2010.

[158] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

[159] M. Shadlen and J. Movshon. Synchrony unbound: review a critical evaluation of the temporal binding hypothesis. *Neuron*, 24:67–77, 1999.

[160] M. Shadlen and W. Newsome. Motion perception: seeing and deciding. *Proceedings of the National Academy of Sciences*, 93(2):628, 1996.

[161] R. Shapley and M. J. Hawken. Color in the cortex: single- and double-opponent cells. *Vision Res.*, 51(7):701–717, Apr 2011.

[162] X. Shi, N. Bruce, and J. K. Tsotsos. Fast, recurrent, attentional modulation improves saliency representation and scene recognition. *CVPR Workshop on Biologically-Consistent Vision*, pages 1–8, 2011.

[163] E. Shikata, Y. Tanaka, H. Nakamura, M. Taira, H. Sakata, et al. Selectivity of the parietal visual neurones in 3d orientation of surface of stereoscopic stimuli. *Neuroreport*, 7(14):2389–2394, 1996.

[164] E. Simoncelli. Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, 13(2):144–149, 2003.

[165] E. P. Simoncelli and D. J. Heeger. A model of neuronal responses in visual area MT. *Vis. Res.*, 38(5):743–761, 1998.

[166] W. Singer. Consciousness and the binding problem. *Annals of the New York Academy of Sciences*, 929(1):123–146, 2001.

[167] H. Spitzer and S. Hochstein. A complex-cell receptive-field model. *Journal of Neurophysiology*, 53(5):1266–1286, 1985.

[168] S. Srivastava, G. Orban, P. De Mazière, and P. Janssen. A distinct representation of three-dimensional shape in macaque anterior intraparietal area: fast, metric, and coarse. *The Journal of Neuroscience*, 29(34):10613–10626, 2009.

[169] S. Swaminathan and D. Freedman. Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. *Nat Neurosci*, 15:315–320, 2012.

[170] M. Taira, K. Tsutsui, M. Jiang, K. Yara, and H. Sakata. Parietal neurons represent surface orientation from the gradient of binocular disparity. *Journal of Neurophysiology*, 83(5):3140, 2000.

[171] A. Takemura, Y. Inoue, K. Kawano, C. Quaia, and F. Miles. Single-unit activity in cortical area mst associated with disparity-vergence eye movements: evidence for population coding. *Journal of Neurophysiology*, 85(5):2245–2266, 2001.

[172] S. Tanabe, K. Umeda, and I. Fujita. Rejection of false matches for binocular correspondence in macaque visual cortical area v4. *The Journal of Neuroscience*, 24(37):8170–8180, 2004.

[173] K. Tanaka. Neuronal mechanisms of object recognition. *Science*, 262:685–688, 1993.

[174] K. Tanaka. Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19(1):109–139, 1996.

[175] K. Tanaka, H. Saito, Y. Fukada, and M. Moriya. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, 66(1):170–189, 1991.

[176] K. Tanaka and H.-A. Saito. Analysis of motion of the visual field by direction, expansion/contraction, and rotation cells clustered in the dorsal part of the medial superior temporal area of the macaque monkey. *J. Neurophysiol.*, 62(3):626–641, 1989.

[177] H. Tanigawa, H. D. Lu, and A. W. Roe. Functional organization for color and orientation in macaque V4. *Nat. Neurosci.*, 13:1542–1548, Dec 2010.

[178] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331:1279–1285, 2011.

[179] F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L. Gallant. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, 12(3):289–316, 2001.

[180] T. Theys, S. Srivastava, J. van Loon, J. Goffin, and P. Janssen. Selectivity for three-dimensional contours and surfaces in the anterior intraparietal area. *Journal of Neurophysiology*, 107(3):995–1008, 2012.

[181] P. Thier and R. Andersen. Electrical microstimulation suggest two different kinds of representation of head-centered space in the intraparietal sulcus of rhesus monkeys. *Proc. Natl. Acad. Sci. 93*, pages 4962–4967, 1996.

[182] O. Thomas, B. Cumming, and A. Parker. A specialization for relative disparity in v2. *Nature Neuroscience*, 5(5):472–478, 2002.

[183] T. Tompa and G. Sáry. A review on the inferior temporal cortex of the macaque. *Brain Research Reviews*, 62(2):165–182, 2010.

[184] R. Tootell, K. Nelissen, W. Vanduffel, and G. Orban. Search for color center(s) in macaque visual cortex. *Cerebral Cortex*, 14(4):353–363, 2004.

[185] M. J. Tovée, E. T. Rolls, and P. Azzopardi. Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. *Journal of Neurophysiology*, 72(3):1049–1060, Sept. 1994.

[186] A. Treisman. The binding problem. *Current Opinion in Neurobiology*, 6(2):171 – 178, 1996.

[187] A. Treisman and H. Schmidt. Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14(1):107 – 141, 1982.

[188] D. Tsao, W. Vanduffel, Y. Sasaki, D. Fize, T. Knutsen, J. Mandeville, L. Wald, A. Dale, B. Rosen, D. Van Essen, et al. Stereopsis activates v3a and caudal intraparietal areas in macaques and humans. *Neuron*, 39(3):555–568, 2003.

[189] J. K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13(3):423–469, 1990.

[190] K. Tsutsui, H. Sakata, T. Naganuma, and M. Taira. Neural correlates for perception of 3d surface orientation from texture gradient. *Science*, 298(5592):409–412, 2002.

[191] T. Uka and G. DeAngelis. Linking neural representation to function in stereoscopic depth perception: roles of the middle temporal area in coarse versus fine disparity discrimination. *The Journal of neuroscience*, 26(25):6791–6802, 2006.

[192] S. Ullman and B. Epshtein. Visual classication by a hierarchy of extended features. In *Towards Category-Level Object Recognition.* Springer-Verlag, 2006.

[193] L. G. Ungerleider and M. Mishkin. Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, editors, *Analysis of Visual Behavior*, pages 549–586. MIT Press, Cambridge, MA, 1982.

[194] C. van den Boomen, M. J. van der Smagt, and C. Kemner. Keep your eyes on development: The behavioral and neurophysiological development of visual mechanisms underlying form processing. *Front Psychiatry*, 16(3), 2012.

[195] J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. Biological Sciences*, 265(1412):2315–2320, 1998.

[196] B. Verhoef, R. Vogels, and P. Janssen. Contribution of inferior temporal and posterior parietal activity to three-dimensional shape perception. *Current Biology*, 20(10):909–913, 2010.

[197] B. Verhoef, R. Vogels, and P. Janssen. Inferotemporal cortex subserves three-dimensional structure categorization. *Neuron*, 73:171–182, 2012.

[198] R. Vogels. Categorization of complex visual images by rhesus monkeys. part 2: single-cell study. *European Journal of Neuroscience*, 11(4):1239–1255, 1999.

[199] H. von Helmholtz, editor. *Handbuch der physiologischen Optic*. Hamburg & Leipzig: Voss, 1866.

[200] J. Wagemans, J. Elder, M. Kubovy, S. Palmer, M. Peterson, M. Singh, and R. von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure-ground organization. *Psychological Bulletin*, in press.

[201] L. Wiskott. How does our visual system achieve shift and size invariance? In J. L. van Hemmen and T. J. Sejnowski, editors, *23 Problems in Systems Neuroscience*, chapter 16, pages 322–340. Oxford University Press, New York, 2006.

[202] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):775–779, July 1997.

[203] J. M. Wolfe. Visual search. In H. Pashler, editor, *Attention*. Unversity College London Press, London, UK, 1998.

[204] F. Xu and S. Carey. Infants' metaphysics: The case of numerical identity. *Cognitive Psychology*, 30(2):111–153, APR 1996.

[205] S. Zeki, S. Aglioti, D. McKeefry, and G. Berlucchi. The neurobiological basis of conscious color perception in a blind patient. *Proceedings of the National Academy of Sciences*, 96:14124–14129, 1999.

[206] H. Zhou, H. Friedman, and R. Von Der Heydt. Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience*, 20(17):6594–6611, 2000.

[207] D. Zipser and R. A. Andersen. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331(6158):679–684, 1988.

Norbert Krüger is a professor at the Mærsk McKinney Møller Institute, University of Southern Denmark. He holds a M.Sc. degree from the Ruhr-Universität Bochum, Germany and his Ph.D. degree from the University of Bielefeld. He leads the Cognitive Vision Lab which focuses on computer vision and cognitive systems, in particular the learning of object representations in the context of grasping.

Ales Leonardis is a Chair of Robotics and a Co-Director of the Centre of Computational Neuroscience and Cognitive Robotics at the University of Birmingham. He is also a full professor at the Faculty of Computer and Information Science, University of Ljubljana, and an adjunct professor at the Faculty of Computer Science, Graz University of Technology.

Peter Janssen is professor of neurophysiology in the Laboratorium voor Neuro- en Psychofysiologie at the KU Leuven, Belgium. He holds an MD degree, a masters degree in psychology, and a PhD degree in Biomedical Sciences from the KU Leuven. His research interests are the processing of three-dimensional shape, object analysis in the dorsal visual stream, functional interactions between cortical areas, and the ventral premotor cortex.

Antonio J. Rodríguez-Sánchez is currently a senior research fellow in the department of Computer Science at the University of Innsbruck, Austria. He completed his Ph.D. at York University, Toronto, Canada on the subject modeling attention and intermediate areas of the visual cortex. He is part of the Intelligent and Interactive Systems group and his main interests are computer vision and computational neuroscience.

Sinan Kalkan received his M.Sc. degree in Computer Engineering from Middle East Technical University, Turkey in 2003, and his Ph.D. degree in Informatics from the University of Göttingen, Germany in 2008. He is currently an assistant professor at the Dept. of Computer Engineering, Middle East Technical University. Sinan Kalkan's research interests include biologically motivated Computer Vision and Cognitive Robotics.

Justus Piater is a professor of computer science at the University of Innsbruck, Austria. He holds a M.Sc. degree from the University of Magdeburg, Germany, and M.Sc. and Ph.D. degrees from the University of Massachusetts Amherst, USA. He leads the Intelligent and Interactive Systems group that works on visual perception and inference in dynamic and interactive scenarios, including applications in autonomous robotics and video analysis.

Markus Lappe received a PhD in physics from the University of Tübingen, Germany. He worked on computational and cognitive neuroscience of vision at the Max-Planck Institute of Biological Cybernetics in Tübingen, the National Institutes of Health, Bethesda, USA, and the Department of Biology of the Ruhr-University Bochum, Germany. In 1999 he was awarded the BioFuture prize of the German Federal Ministry of Education and Research. Since 2001 he is full professor of Experimental Psychology at the University of Muenster. He is also a member of the Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience at the University of Muenster.

Laurenz Wiskott is full professor at the Ruhr-Universität Bochum, Germany. He holds a Diploma degree in Physics from the Universität Osnabrück and a PhD from the Ruhr-Universität Bochum. The stages of his career include The Salk Institute in San Diego, the Institute for Advanced Studies in Berlin, and the Institute for Theoretical Biology, Humboldt-Universität Berlin. He has been working in the fields of Computer Vision, Neural Networks, Machine Learning and Computational Neuroscience.