University of the Algarve

Faculty of Sciences and Technology

Integrated multi-scale architecture of the cortex with application to computer vision

Ph.D. Thesis in Electronic and Computer Engineering,
Major in Computer Science

João Miguel Fernandes Rodrigues

Supervisor: Johannes Martinus Hubertina du Buf, Faculty of Science and Technology, University of the Algarve, Portugal

Juri:
President:
João Pinto Guerreiro, Reitor, University of the Algarve, Portugal

Vogals:
Rolf Wuertz, Institut fur Neuroinformatik, University of Bochum, Germany
Gustavo Deco, Department of Technology, University Pompeu Fabra, Barcelona, Spain
Aurélio Campilho, Faculty of Engineering, University of Porto, Portugal
Alexandra Reis, Faculty of Human and Social Sciences, University of the Algarve, Portugal
Johannes du Buf, Faculty of Science and Technology, University of the Algarve, Portugal
Miguel Castelo-Branco, Faculty of Medicine, University of Coimbra, Portugal
Hamid Shahbazkia, Faculty of Science and Technology, University of the Algarve, Portugal
Marina Graça, School of Education, University of the Algarve, Portugal

Faro
(July, 2007)

University of the Algarve

Faculty of Sciences and Technology

Integrated multi-scale architecture of the cortex with application to computer vision

Ph.D. Thesis in Electronic and Computer Engineering,
Major in Computer Science

João Miguel Fernandes Rodrigues

Faro
(July, 2007)

Acknowledgments

First many thanks go to Prof. Hans du Buf, for directing my attention to the challenging but also inspiring field of human vision, for his excellent supervision based on his broad experience, for providing a stimulating and cheerful research environment in his Vision Laboratory, for his encouraging support and his way to always find time for discussions, and last but not the least for the numerous hours that he spent in correcting my English.
Many thanks go also to my laboratory colleagues, to Roberto Lam who implemented Fig. 2.3, and to Samuel Nunes, Daniel Almeida, Vera Brito and João Carvalho for all the work that contributed to Chapter 7 and the related papers on painterly rendering.
To all the "inhabitants" of the laboratory, permanent or visiting, especially those with whom I have worked with, almost on a daily basis, discussing scientific and technical problems, but also almost all problems in the world.
A special thanks to the Escola Superior de Tecnologia at UAlg and my colleagues at the Department of Electrical Engineering, for reducing my teaching duties during the last 3 years.
Some Portuguese words for my parents: "um obrigado especial para os meus pais, Flávia e José Rodrigues, pelo apoio, carinho e motivação que sempre me deram ao longo de toda a minha vida académica, sem os quais nunca teria sido possível chegar até aqui."
To my own family, first to my wife Celia Ramos, for all the love and support, encouragements and suggestions, and for the home environment that allowed me to work many times until late into the night, but also to my daughter Joana, who was born at about the same time that this work started, for all the affection, not forgetting the interesting talks about how she saw the world as a two- and three-year old child. Many of these talks were very inspiring and contributed to the ideas about object categorizations.

NOME: João Miguel Fernandes Rodrigues
FACULDADE: Faculdade de Ciências e Tecnologia
ORIENTADOR: Professor Doutor Johannes Martinus Hubertina du Buf
DATA: Julho de 2007
TITULO DA TESE: Arquitectura integrada multi-escala do córtex visual com aplicações na visão por computador

Resumo

O foco principal desta dissertação é compreender o desenvolvimento e a funcionalidade do córtex visual através de modelos computacionais. Na camada de entrada V1 do córtex visual, existem células *simple*, *complex* e *end-stopped*. Estas permitem uma representação multi-escala de objectos ou de cenas em termos de linhas, arestas e pontos-chave. Nesta dissertação, são combinados os progressos mais recentes no desenvolvimento de modelos computacionais destas e de outras células com os processos que decorrem em áreas superiores do córtex V2, V4 etc. Três desafios pertinentes são estudados: (i) o reconhecimento de objectos embebido numa arquitectura cortical; (ii) a percepção do brilho, e (iii) a renderização de pinturas usando a visão humana. Aspectos específicos são Foco-de-Atenção baseado em mapas de saliência criados a partir de pontos-chave, o reencaminhamento dinâmico de atributos a partir de V1 para áreas superiores do córtex de forma a obter invariância à translação, à rotação e ao tamanho, e a construção de modelos canónicos das vistas dos objectos na memória visual. As nossas simulações mostram que as representações multi-escala podem ser integradas numa arquitectura cortical, de forma a modelar os seguintes passos: segregação, diferentes níveis de categorização e o reconhecimento final de objectos. Relativamente ao processamento cortical real, o sistema começa com a informação das escalas grosseiras, refina a categorização usando escalas intermédias, e utiliza todas as escalas para o reconhecimento. Também apresentamos um modelo de brilho em 2D, baseado na representação simbólica de linhas e arestas, combinado com um canal passa-baixo e com funções de transferência não lineares, de tal forma que o reconhecimento de objectos e a percepção de brilho são processos integrados e baseados na mesma informação. O modelo de brilho consegue prever efeitos tais como bandas Mach, a ilusão Craik-O'Brien-Cornsweet e a indução de *gratings* e de brilho, mais concretamente os efeitos opostos de assimilação (efeito White) e contraste simultâneo de brilho. Por fim, introduzimos uma nova aplicação: a renderização da pintura tem estado ligada à visão computacional, mas nós propomos a ligação desta com a visão humana, porque a percepção e a pintura são dois processos interligados.

PALAVRAS-CHAVE: Córtex visual, Foco-de-Atenção, categorização, reconhecimento, brilho, renderização.

Integrated multi-scale architecture of the cortex with application to computer vision

Abstract

The main goal of this thesis is to try to understand the functioning of the visual cortex through the development of computational models. In the input layer V1 of the visual cortex there are simple, complex and end-stopped cells. These provide a multi-scale representation of objects and scene in terms of lines, edges and keypoints. In this thesis we combine recent progress concerning the development of computational models of these and other cells with processes in higher cortical areas V2 and V4 etc. Three pertinent challenges are discussed: (i) object recognition embedded in a cortical architecture; (ii) brightness perception, and (iii) painterly rendering based on human vision. Specific aspects are Focus-of-Attention by means of keypoint-based saliency maps, the dynamic routing of features from V1 through higher cortical areas in order to obtain translation, rotation and size invariance, and the construction of normalized object templates with canonical views in visual memory. Our simulations show that the multi-scale representations can be integrated into a cortical architecture in order to model subsequent processing steps: from segregation, via different categorization levels, until final object recognition is obtained. As for real cortical processing, the system starts with coarse-scale information, refines categorization by using medium-scale information, and employs all scales in recognition. We also show that a 2D brightness model can be based on the multi-scale symbolic representation of lines and edges, with an additional low-pass channel and nonlinear amplitude transfer functions, such that object recognition and brightness perception are combined processes based on the same information. The brightness model can predict many different effects such as Mach bands, grating induction, the Craik-O'Brien-Cornsweet illusion and brightness induction, i.e. the opposite effects of assimilation (White effect) and simultaneous brightness contrast. Finally, a novel application is introduced: painterly rendering has been linked to computer vision, but we propose to link it to human vision because perception and painting are two processes which are strongly interwoven.

KEYWORDS: Visual cortex, Focus-of-Attention, categorization, recognition, brightness, rendering.

Contents

Chapter 1

Introduction

Abstract: This chapter introduces the scope of the thesis as well as the two major problems that will be studied, namely object recognition and brightness perception.

1.1 Scope of the thesis

Imagine: you are going to see a movie with your daughter Joana, you are in the line in front of the entrance of the theater talking to one of your friends, and she enters the room first taking with her the tickets with the seat numbers. When you enter you don't know where she is. A small embedded system in your coat connected to a few button-sized cameras tells you "Joana is third to the right," "partly occluded by blond woman." When you start walking towards her, the light is dimmed and the system alerts you "attention handbag on floor," "attention cane between seats," "attention popcorn bag on seat."

From an engineering point of view, you will think that the implementation of such a system involves methods from Computer Vision (CV). When you try to join all the pieces, you find that even state-of-the-art CV methods, which are very good at solving restricted problems like object detection (floor, seats), categorization (face, handbag, cane) and identification (daughter Joana), are not able to categorize all types of objects in complex scenes nor recognize individual objects like faces when partly occluded, especially with additional complications like different illuminations and viewpoints etc. Just imagine for instance the same scene as above but at an airport lounge or in a disco. Not surprisingly, such a general and flexible system still belongs to science fiction. Nevertheless, we know very well one system that can cope with all such complications—our visual system. So, HV (Human Vision) will provide a solution, or CV based on HV. There's only one small problem left: we need to know first how HV works.

When analyzing the performance of contemporary systems based on HV—often referred to as biologically inspired systems—one must conclude that they fall behind CV systems, despite some very promising results. This means that "science fictional" systems based on HV have a long road to go, but we are certain that they will work at the end, simply because we see them working every day. In addition, all information exchanged between vision researchers, neurophysiologists, psychologists and engineers may also help to treat

vision deficiencies and diseases, and to create new experiments to better understand how our brain works.

There are many reasons for creating an HV model, but they can be summarized as follows:

1. The human visual system is the best vision system "on the market," not claiming that other systems of birds or mammals or other primates are inferior; we know what we see and we can only guess what a chimp sees. (After reading the following chapters of this thesis the reader should understand that the last part of the last sentence should read "we think we know what we see").

2. We believe that enough computational ideas and experimental data are now available. On the basis of these it is possible to begin the development of an integrated theory of the ventral and dorsal data streams in the brain, focusing on an explanation of visual object recognition. This theory as a whole or parts of it may be incorrect, but at least it represents a skeletal set of claims and ideas which can be tested, confirmed or rejected, and an integrated architecture will be modular such that parts can be replaced or improved.

3. With ever increasing performance of modern computers (Moore's Law) we are going to have the necessary power to create realistic models. To give an idea: using two graphics boards with GPUs optimized for vectorized multiplication and accumulation (multiply-add or MADD) operations, one can obtain a performance of 1 TFLOPS on a normal personal computer. 1 TFLOPS means 10^{12} or one million million of floating point operations per second. Our entire brain counts 10^{12} neurons. This does not mean that it will be possible to create a dynamic model of the entire brain at intervals of one second, because most neurons have between 100 and 1000 interconnections. Apart from this limitation, the real bottleneck is still storage capacity, both memory and disk space (and disk access time).

Of course, instead of modeling the entire brain or "only" the entire visual system we intend to focus on a few cortical areas and a relatively small group of cells in the early layers, because (1) the HV system is far too complex to try to model everything in a single step (this even applies to the biggest research groups), (2) cortical areas V1 and V2 etc. are the major processing areas, (3) they are the best known and investigated areas, and (4) like a house has to be constructed starting with its foundations, a cortical model/architecture has to be started from the best known and lowest layers, and with time more cells and layers can be added.

Despite the above arguments, many questions remain open. For instance: how is the information really processed in the cortex? Is it only done bottom-up, also called data-driven? Or are there also top-down feedback loops, and if so, only a few or many? Between which layers and/or which areas? What do they serve for? Even the interconnections between neurons are not well known, i.e., there is no consensus between the principal groups working in this field about a neural architecture that could unify all processing steps into a single structure. For example, at the lowest level one can ask how and when each cell is activated in each layer, which cells combine within one single layer, and which cells activate the next layer. At the highest level the same questions are related to how object representations are stored in visual memory, to when and how scene and object recognition start, etc.

Since there are so many open questions, we are convinced that many groups will be working on cortical models in the future, each time trying to take the knowledge one step

further, but progress will be slow because right now only few groups are developing cortical cell models, less groups are developing an integrated cortical architecture, and even fewer groups are combining object recognition with other aspects like brightness perception. That there are not more groups developing an integrated architecture has two main reasons: it is a very interdisciplinary field and it does not return "excellent" results very fast (excellent in terms of data suitable for publications in a "publish or perish" academic society).

The research groups working on cortical models or on HV are all looking for the Holy Grail: an approach, probably multi-scale, that can yield a complete characterization of an image. There are many practical applications that can benefit from advanced cortical models: object and face recognition, texture analysis, image segmentation, motion and depth prediction, and image enhancement and coding. In addition, a good two-dimensional brightness model can replace human observers, for example in estimating image quality in coding, by comparing the (subjective) brightness of a coded-decoded image with the (subjective) brightness of the original image. There also are other scientific and technological areas where HV-related knowledge can be applied, from engineering (better ways to recognize persons) to medicine (how to treat some illnesses), from education (more efficient ways to teach the brain) to arts (new ways to study paintings and painters).

Specifically, the main focus of this thesis is on the visual cortex, exploring a possible integrated architecture, but always having in mind practical applications. Main topics are:

- the development of computational cell models on the basis of cortical simple, complex and end-stopped cells for the explicit extraction of lines, edges and keypoints,
- to incorporate these models into a multi-scale approach,
- to extract the most accurate and reliable information by coarse-to-fine scale processing,
- to study multi-scale approaches for object categorization and recognition,
- to complement multi-scale image representations for object recognition with brightness perception, and
- to propose an integrated model or architecture of the cortex, relating features with cells, cell layers and with the information pathways of the visual system.

In the next sections, the modeling of object recognition and brightness perception are discussed in more detail and the structure of the thesis will be presented.

1.1.1 Object recognition

Object recognition is a classical problem which is addressed in any book on computer vision, image processing and machine vision. It can be loosely defined by determining whether or not an image (or video) contains one or several specific objects, features or activities. Despite this very generic definition, it is quite difficult to define the term recognition in the context of HV, because each author applies his own definition, and it even may change within one publication. For instance, we can refer to recognition as to recognize one or several pre-specified or learned objects or object classes (e.g. this is a coffee mug), it may be the identification of a specific object (this is Paul's coffee mug), or even detection (there are two coffee mugs in this image). It must follow from the context what is really meant, detection, categorization or identification. The same applies to this thesis.

Computationally, recognition is one of the most difficult tasks, but the difficulty depends on the task: it is quite easy for an average computer-vision student to detect and recognize objects after a very few lessons, i.e., if we put a few and very distinct objects on the top of a white table with good illumination. But if we put more and less distinct objects, some objects partly occluding other objects, and cover the table with a cloth with some complex texture, the task becomes more difficult, recognition performance decreases, and much more effort will be required to boost performance to an acceptable level.

Most real-world applications are not trivial, even the ones that appear relatively easy. For instance, counting how many people there are waiting on a sidewalk to automate the control of a zebra crossing, or reading number plates of cars passing a toll gate at 100 km/h, are already quite complicated. And then there are the very complicated applications, like recognizing a person after changing the hair style, after growing or shaving a beard, or after having grown old and gray. The extreme case is spotting in CCTV video, in real-time or in logged video, someone who does not want to be recognized, who therefore may use all disguisement tricks and even plastic surgery.

In neurosciences the concept of object recognition is even more difficult since it involves several levels of understanding, from the information processing or computational level to the level of circuits and cellular and biophysical mechanisms. After decades of research effort, neuroscientists working on functions in striate and extrastriate cortical areas have produced a huge and still rapidly increasing amount of data, and the emerging picture of how cortex performs object recognition is in fact becoming too complex for any simple model [Serre et al., 2005]. Recognition turns out to be a delicate compromise between selectivity and invariance. Therefore, the key computational issue in object recognition is the specificity-invariance trade-off: the system must be able to finely discriminate between different objects or object classes, while at the same time be tolerant to sometimes big object transformations which include scaling, translation and (2D) rotation, also changes of illumination, (3D) viewpoint, context and clutter, non-rigid transformations such as a change of facial expression and, in the case of categorization, also shape variations within a class [Serre et al., 2005].

Another problem that increases difficulty in modelling "biological recognition" is the definition of the instant when it all starts. Psychologists and psychophysicists, who study how we perceive patterns and images, used to think that, before the processes of object recognition and categorization could begin, the brain must first isolate a figure in the image—such as a tree or a piece of fruit—from its background (this process is called object segregation). However, recent research suggests that we actually categorize objects before we have segregated them, or that both processes occur in parallel. This means that by the time you realize that you are looking at something, your brain already knows what that thing is [Oliva and Torralba, 2006]. Such topics even relate to consciousness, which will not be stressed in this thesis.

Grill-Spector and Kanwisher [2005] tested three types of visual recognition by briefly flashing images before the eyes of human observers. The first type, object detection, was tested by showing images that may or may not have contained figures. Participants had to quickly judge whether or not there was a figure present against a background. The second type concerned categorization, where participants were shown images of figures and they had to indicate what type of figure they saw, such as bird, car, or food. In the third part of the test, more specific images were shown in order to test identification. Participants had to identify figures within categories, such as parrot or pigeon in the category "bird." It turned out that the participants were as fast and accurate in naming the category that an object belonged to as they were at saying whether or not they had seen an object at all. The ability

of the subjects to process the images in such a short time proved that, by the time they knew an image contained some sort of object, they already knew its category.

Grill-Spector and Kanwisher [2005] concluded that "There are two main processing stages in object recognition: categorization and identification, with identification following categorization," also "Overall, these findings provide important constraints for theories of object recognition," and "Rapid categorization obviously facilitates our survival and interaction with the environment on an everyday level." This built-in human process of rapid categorization before identification restricts the brain's search for a match between the visual input (the picture you looked at) and internal category-relevant representations (stored images of other objects you have seen and identified prior to today).

From these conclusions it follows that recognition/identification should not be studied or modeled as a single-level task, but as a multi-level task where one or several levels of categorization should be performed. In addition, categorization should start at the same time as detection or segregation.

1.1.2 Modeling brightness perception

Visual psychophysics is a scientific area concerned with developing a complete understanding of how it works: from the physical input (the light flux entering the eye) to the output (the subjective image that we perceive). There are many aspects like brightness, contrast, color, shape, shading and texture [du Buf, 2001]. One chapter of this thesis concentrates on brightness, i.e., the relation between (physical) luminance and (subjective) brightness of many spatial patterns. The goal is the construction of a generally applicable brightness model, which can predict most if not all known brightness effects.

Developing brightness models is perhaps one of the most difficult aspects of quantitative visual psychophysics, and this subject has not been very popular in vision research [du Buf and Fischer, 1995]. It requires knowledge about published data and experiments, as there is no standardized database that joins all available experimental results, also knowledge about signal and image processing, a lot of programming and, again, really fast computers.

du Buf [2001] proposed that semantic processing, wherever it may be done, obtains input from lower-level syntactical processing layers, probably providing a multi-scale line, edge and vertex representation. This actually is the same representation that will be exploited in this thesis for object recognition. In other words, object recognition and brightness perception are related processes which can be integrated: seeing an object implies seeing its brightness pattern but also knowing what it is. This, again, relates to consciousness: we open our eyes and we see the world around us, we become conscious of the world and our position in it. Here there are four main observations: (1) the entire idea is based on the fact that simple cells do not allow to discriminate between lines and ramp edges, which explains the appearence of Mach bands at ramp edges (see Chapter 6), (2) brightness perception being related to multi-scale detection and processing of lines and edges in area V1 and beyond, higher-level cognitive effects such as change blindness imply that brightness is based on low- and high-level processing, (3) the previous point implies that consciousness too is a holistic process which may involve the entire brain, and (4) the image that we perceive is not a straightforward reconstruction because we think in terms of semantics, where objects are not represented any more by some sort of stored "pictograms" but in the form of functional descriptors. Some of these points lead to a nice paradox: opening our eyes is like switching on a TV set, but where are the electrons and the phosphor atoms?

Despite the difficulties and complications referred to above, some brightness models have

been published recently, either as computational models or as theoretical explanations (see e.g. [du Buf and Fischer, 1995; Blakeslee et al., 2005; Logvinenko and Ross, 2005]). There are good reasons to pursue research in this field: (1) There are real applications in areas like image processing and computer graphics, such as image enhancement for pattern detection in medical imaging. (2) A model based on psychophysical data can be tested against these and model predictions, in particular inaccurate ones, lead to a better insight into the process of visual perception, i.e., feedback leading to additional psychophysical experiments concerning unclear aspects of spatial interactions in brightness perception. (3) A good brightness model can serve as the basis for codecs (coding/decoding) schemes with high compression rates because those may cause image deformations that are more natural and therefore more difficult to perceive if compared to standard codecs based on straightforward subband decomposition and quantization schemes.

1.2 Overview of the thesis

Chapter 2 presents a small overview of the visual cortex, its architecture and functionality. It explains generically most known cells, the most significant visual areas and visual pathways. It finalizes by presenting some initial conclusions that will guide us towards developing an invariant object categorization and recognition architecture.

Chapter 3 introduces the multi-scale keypoint representation. It shows that this provides very important information for object and face detection. It also shows that saliency maps for Focus-of-Attention can be constructed on the basis of this representation, and that such maps can be employed for the detection of facial landmarks and faces.

Chapter 4 introduces the multi-scale line and edge representation. It illustrates visual reconstruction, and how object segregation can be achieved with coarse-to-fine-scale groupings. A two-level object-categorization scenario is tested and also a multi-scale object-recognition model. A new disparity model based on the multi-scale line and edge coding is presented, such that depth from stereo can be attributed to lines and edges.

Chapter 5 extends the multi-scale representations into an integrated, invariant architecture with dynamic routing of object features throughout the cortex and the construction of normalized object and group templates.

Chapter 6 presents a two-dimensional brightness model. This model is calibrated using psychophysical data, and it is shown that it can predict many brightness effects such as Mach bands, White's effect, simultaneous brightness contrast, grating induction and the Craik-O'Brien-Cornsweet illusion.

Chapter 7 presents a specific application: painterly rendering using human vision. Completely automatic rendering is obtained by applying the multi-scale line and edge representation that provides a very natural way to render broad and fine brush strokes, and the multi-scale keypoint representation serves to create saliency maps for Focus-of-Attention to render important structures (abstraction).

Final remarks and ideas for future research are presented in Chapter 8.

Parts of this thesis have already been published in journals and related work has also been presented at conferences. Chapters 3 and 7 were published in 2006 in BioSystems and Virtual, respectively. Chapter 5 has been submitted to Cognitive Processing. Chapters 4 and 6 are being prepared for submission to journals like BioSystems and Spatial Vision. Appendix A lists all publications.

Chapter 2

Overview: cortex, architecture and functionality

Abstract: This chapter presents a brief overview of the biological aspects of vision with special focus on the visual cortex. The view-based approach of how object invariance can be achieved is discussed. This chapter is concluded with a brief summary of conclusions that will guide us towards developing an invariant object categorization and recognition architecture.

2.1 Introduction

Intuition tells us that the brain is complicated. The brain contains about 10^{12} (one million million) cells, an astronomical number by any standard. In addition, a typical neuron receives information from hundreds to thousands of other neurons and in turn transmits information to the same number of neurons, so the total number of interconnections is between 10^{14} and 10^{15}. But complexity is not only defined by these numbers, even more important is the organization and functionality, aspects which are very hard to quantify [Hubel, 1995]. Hubel states that neurons are the basic structural components of the brain. A neuron is an individual cell, specialized by architectural features that enable fast changes in neighboring neurons. The brain is "just" an assembly of such cells, and while individual neurons do not see, reason or remember, the brain as a whole does.

A neuron, or nerve cell, consists of the cell body that has a globular shape and contains the nucleus, and from the cell body protrudes the output-signal transmitting nerve fiber called axon. Besides the axon, a number of other branching and tapering fibers are connected to the cell body, the dendrites. The entire cell, body, axon and dendrites, are enclosed by the cell membrane. The cell body and dendrites receive information from other cells, whereas the axon transmits information from the cell to other cells. Near the end an axon normally splits into many branches, whose terminal parts come very close to the cell bodies and/or dendrites of other cells. In these signal-transmission regions, called synapses, information is conveyed from one nerve cell (presynaptic) to the next (postsynaptic) one [Hubel, 1995].

Here we are interested in the visual pathways and brain regions involved in vision. The retina in an eye, which is considered part of the brain, is a thin laminar structure with several layers of cells, one of which containing the light-sensitive or photoreceptor cells, the rods and cones. The optic nerves of the two retinas pass through the optic chiasm, where about half of the fibers cross to the side of the brain opposite the eye of origin, left and right, and about half stays on the same side. From the chiasm the fibers lead to the lateral geniculate nucleus (LGN). The optic-nerve fibers have terminal synapses at cells in the LGN and axons of LGN cells terminate in the primary visual cortex, layers 4Cα and 4Cβ in area V1 [Hubel, 1995; Bruce et al., 2000]. In all these connections, from the retina via the LGN to the cortex, there are retinotopic projections. This means that the mapping of each structure to the next is systematic: as you move in the retina from one point to another, the corresponding points in the LGN and cortex also follow a continuous path. In other words, in retinotopic projections the neighborhood relations like left-right and up-down are preserved.

Another important concept is the receptive field (RF) of a neuron. The RF is defined by the spatial region at the retinal level in which the presence of a stimulus will affect the firing rate of that neuron. In the visual system, receptive fields are volumes in visual space. For example, the receptive field of a single photoreceptor is a cone-shaped volume comprising all the visual directions in which light will alter the response of that photoreceptor. In the case of binocular neurons in the visual cortex, it is necessary to specify the corresponding areas in both retinas. Although these can be mapped separately in each retina by shutting the one and then the other eye, the full influence on the neuron's firing is revealed only when both eyes are open [Hubel, 1995].

Hubel and Wiesel in 1963 advanced the theory that receptive fields of cells at one level of the visual system are formed from input by cells at a lower level (see [Hubel, 1995]). In this way, small and simple receptive fields could be combined to form big and complex receptive fields. Theorists later elaborated that this simple, hierarchical processing structure can be influenced by feedback from higher levels. Receptive fields have been mapped from cells at all levels of the visual system: photoreceptors, retinal ganglion cells, and cells in the lateral geniculate nucleus, the visual cortex cells and even in extrastriate cortical areas.

Before presenting a description of the visual cortex, it is useful to introduce the concept of cortical plasticity (or neuroplasticity), which refers to changes that occur in the organization of the brain, in particular the changes in the location of specific information processing functions, as a result of the effects of learning and experience. A surprising consequence of plasticity is that a specific function can "move" from one location to another after repeated learning or even brain traumas. This phenomenon is complex and involves many levels of organization. To some extent the term itself has lost its explanatory value because almost any changes in brain activity can be attributed to some sort of "plasticity." Cortical organization, especially for the sensory systems, is often described in terms of maps. For example, tactile information from the foot projects to one cortical site and information from the eyes (vision) projects to another site. As a result, the cortical representation of the body resembles a map, but this map is not "fixed" but rather plastic. Several groups began exploring the impacts of removing parts of the sensory inputs in the late 1970s. We now know that re-organization occurs at every level in the processing hierarchy in the cerebral cortex [Miikkulainen et al., 2005].

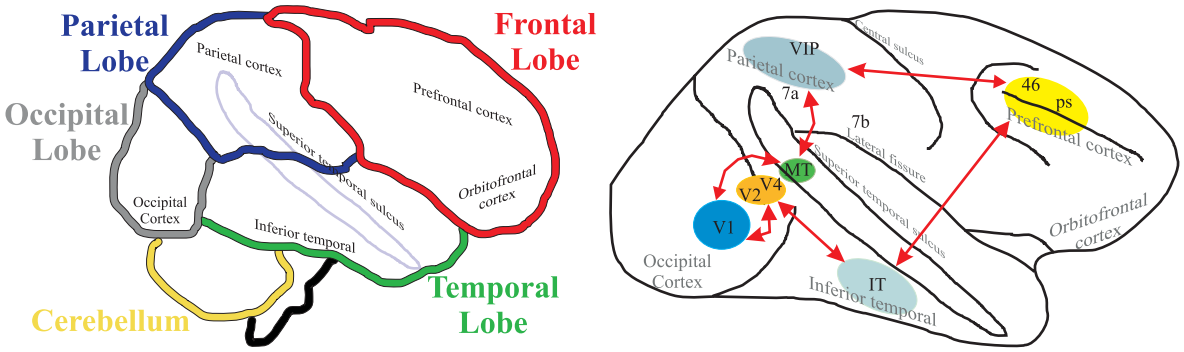

Figure 2.1: Left: the brain's anatomical areas. Right: data flow in Deco and Rolls' cortical architecture, adapted from Fig. 1 in Deco and Rolls [2005].

2.2 The visual cortex

In this thesis we concentrate on the architecture of the visual cortex. The primary visual (or striate) cortex or area V1 is a layer of cells which is 2 mm thick, with a surface area of a few tens of square centimeters. The other visual cortical areas (V2, V3, V4, MT) are not at the surface of the brain and are called extrastriate areas. A detailed diagram of the cortical areas and their connections in the macaque monkey can be found in Churchland and Sejnowski [1992] (pp. 22) and in Parasuraman [1998] (see pp. 309 for the color plates).

There are many visual pathways, but we concentrate on two called the ventral stream and the dorsal stream (ventral means belly and dorsal means back, common anatomical terms which also apply to the spinal chord and the forward bending brain). Both streams start at the level of retinal ganglion cells and continue through the LGN to V1. The ventral stream then goes via areas V2 and V4 to the inferior temporal cortex, IT (see Fig. 2.1). Based on physiological experiments in monkeys, IT has been postulated to play a central role in object recognition. IT cortex, in turn, is a major source of input to prefrontal cortex (PF), which is involved in linking perception to memory and action [Miller and Cohen, 2001]. The ventral stream, also called the "what" pathway, is associated with form recognition and object representation. It is also associated with storage in long-term memory. The dorsal stream goes from V1 via V2 and V3 to middle temporal area (MT) and to the inferior parietal lobule [Goodale and Milner, 1992]. The dorsal stream, also called the "where" pathway, is associated with motion, the representation of object locations, and control of the eyes and arms, especially when visual information is used to guide saccades. The dichotomy of the ventral/dorsal or what/where pathways (sometimes also referred to as the perception/action streams) was proposed (among others) by Goodale and Milner [1992] and is still being applied, but also disputed, by vision scientists and psychologists. It is probably an over-simplification of the real organization of the visual cortex.

Many neurons in the visual cortex only respond to a subset of stimuli within their receptive field. This property is called tuning. In the earlier visual areas, neurons are tuned to simpler patterns. For example, a neuron in V1 may fire to any vertical stimulus in its receptive field. In the highest visual areas, neurons are tuned to much more complex patterns. For example, in inferior temporal cortex (IT), a neuron may only fire when a certain face appears in its receptive field. Individual V1 neurons in primates and animals with binocular vision have ocular dominance, i.e., a preference for one of the two eyes.

In V1, and the primary sensory cortex in general, neurons with similar tuning proper-

ties tend to cluster together in cortical columns, spatially arranged following two tuning properties: ocular dominance and orientation [Hubel, 1995]. However, this model cannot accommodate color, spatial frequency and many other features to which neurons can be tuned. As mentioned above, the transformation of the visual image from retina to V1 is referred as retinotopic mapping. The correspondence between a given location in V1 and in the subjective visual field in the external environment is very precise: even the retinal blind spots are mapped into V1. Evolutionary, this correspondence is very basic and found in most animals that possess a V1. In man and animals with a fovea in the retina, a large portion of V1 is mapped to the small, central part of the visual field, a phenomenon known as cortical magnification.

2.2.1 Cortical areas

As already mentioned, the first cortical area is V1; see [Olshausen and Field, 2005] for a detailed discussion of V1. Current consensus seems to be that V1 consists of tiled sets of spatiotemporally selective filters. Theoretically, these filters together can carry out neuronal processing of spatial frequency, orientation, motion, direction, speed and many other spatiotemporal features. Many experiments with V1 neurons have lead to this insight. Visual information relayed to V1 is not coded in terms of a spatial (or optical) intensity image, but rather as local contrast. As an example, in the case of an image which is half black and half white, the dividing edge between black and white has a strong local contrast and this edge is encoded, while few neurons may code the brightness information. As information is further relayed to subsequent visual areas, it is coded as increasingly non-local frequency/phase signals.

Area V2 is the second major area in the visual cortex. It receives direct input from V1 and sends output to V3, V4 and MT. It also sends feedback signals to V1. Functionally, V2 has many properties in common with V1. Cells are tuned to simple features such as orientation, spatial frequency and color [Hubel, 1995]. Responses of many V2 neurons are also modulated by more complex features, such as the orientation of illusory contours and whether a stimulus is part of the figure or the ground, at least at the level of local occlusions [Qiu and von der Heydt, 2005].

Area V3 is part of the dorsal stream, receiving inputs from V2 and primary cortex. It projects to the posterior parietal cortex. Properties of cells in V3 offer few clues as to its function. Most cells are selective to orientation, and many are also tuned to motion and to depth. Relatively few are color sensitive, for more details see [Gegenfurtner et al., 1997; Kaas and Lyon, 2001].

Area V4 has been identified in the extrastriate visual cortex of the macaque. It is still unknown what the human homologue of V4 is; this issue is currently the subject of much scrutiny. V4 is the third cortical area in the ventral stream and the first one that shows strong attentional modulation [Chelazzi et al., 2001]. It receives strong feedforward input from V2 and sends strong output to the posterior inferotemporal cortex (PIT). It also receives direct input from V1. In addition, it has weaker connections to MT and visual area DP (the dorsal prelunate gyrus). Like V1, V4 is tuned to orientation, spatial frequency and color. Unlike V1, it is tuned to object features of intermediate complexity, like simple geometric shapes, but simpler than IT, although no one has yet developed a full parametric description of the tuning space of V4. Although first known for their color selectivity, neurons in V4 are selective to a wide variety of forms and shapes, such as bars, gratings, angles, closed contour features, sparse noise, etc.; see e.g. [Pasupathy and Connor, 2001; Chelazzi et al.,

2001]). Area V4 is not tuned to complex objects such as faces, in contrast to areas in the inferotemporal cortex. V4 is also known to have receptive fields of intermediate sizes (larger than V1 and smaller than IT on average), and invariance to small translations.

Area MT (middle/medial temporal) is a region in the extrastriate cortex that appears to process complex motion stimuli. It contains many neurons which are selective to the motion of complex features like line ends and corners [Hubel, 1995; Bruce et al., 2000]. Much work has been carried out on MT as it appears to integrate local motion signals into the global motion of complex objects, but some research suggests that motion information is in fact already available at lower levels of the visual system such as V1. There is still much controversy over the exact computations carried out in area MT. An updated overview of the rich literature on MT was recently presented by Born and Bradley [2005].

Area IT (inferior temporal) is one of the highest levels of the ventral stream, with representations of visual shapes and objects. In Logothetis et al. [1995] monkeys were trained to recognize a set of novel "paperclip" objects and some neurons in anterior IT were found to be tuned to the trained views of those objects, but invariant to changes in size, translation and 3D rotation. These view-tuned neurons responded more strongly to scaled, translated and rotated (in depth) images of the preferred paperclip than to a large number of distractor paperclips, even though these objects had been previously presented with just one size, position and viewpoint. A later study systematically looked at the effect of adding one or two distractor objects within the receptive field of an IT neuron [Zoccolan et al., 2005]. Most recorded neurons showed an average-like behavior. That is, the response to the cluttered condition, containing two or three objects, was close to the average of responses to the individual objects if presented alone. There are several potential explanations listed by Zoccolan et al. [2005]. One explanation is to assume a normalization stage by the overall activation of the entire IT cell population. This is quite feasible, since such a normalization would make learning easier for the next layer.

Area PF (prefrontal), which receives most input from IT, is involved in linking perception to memory and to action [Miller, 2000]. IT is also the last purely visual area which is task independent. Responses of PF cells are much more task dependent than responses of IT cells [Freedman et al., 2003]. Recent recordings [Freedman et al., 2002, 2003] revealed that neurons in PF are often "category-tuned," conveying reliable information about category membership, learned in a supervised way, and relatively little information about individual stimuli within each category. By contrast, the majority of neurons in IT showed shape-tuning, i.e., they tended to show selectivity to individual stimuli (for example faces, see [Afraz et al., 2006]) and little evidence for selectivity to category membership.

2.2.2 Cells: simple, complex, end-stopped and more

Receptive fields of cells in the visual cortex are larger and more complex than retinal ganglion and LGN cells. Hubel and Wiesel first classified cells into three types: simple, complex and hypercomplex cells [Hubel, 1995].

Receptive fields of simple cells are elongated, for example with an excitatory central oval region and an inhibitory surrounding region, or approximately rectangular, with one long side being excitatory and the other being inhibitory. Cells with receptive fields with their long axis rotated to any angle have been found. Excitatory and inhibitory domains are always separated by a straight line or by two parallel lines. Some of the cells have an excitatory or an inhibitory region that is positioned exactly in the center of the receptive field, resulting in a symmetric receptive field; these are called even simple cell because of

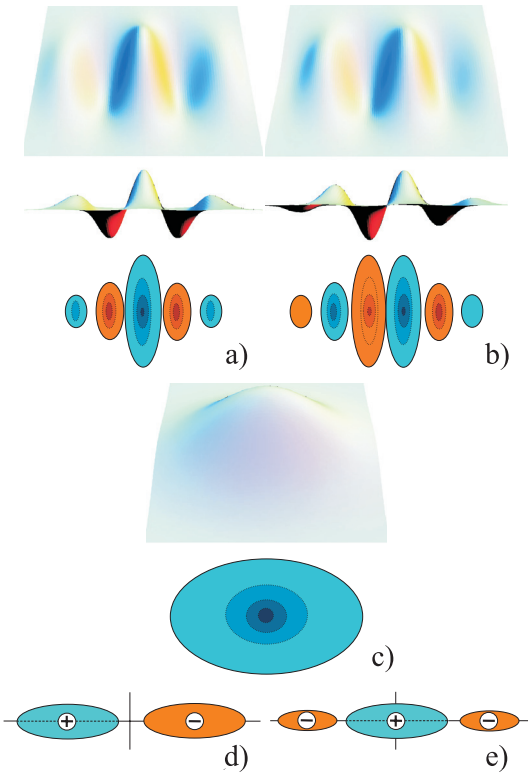

Figure 2.2: 2D and 3D receptive-field representations of even (a) and odd (b) simple cells and complex (c) cells. Single (d) and double (e) end-stopped cells,

even symmetry, see Fig. 2.2a. Others, have an asymmetric receptive field profile, as the striped regions are positioned with a certain offset which respect to the center of the field: odd simple cells, see Fig. 2.2b. The size of the receptive field depends on its corresponding position in the retina relative to the fovea, but even at a given position in the retina there is some variation in size. In general, simple cells with smallest receptive fields are found in and near the fovea [Hubel, 1995]. Simple cells must be built up from preceding cells, probably from retinal ganglion cells with circular receptive fields.

Complex cells represent the next step in the analysis. Their receptive fields are also elongated but simpler than those of simple cells, because there are no sub-regions; see Fig. 2.2c. Complex cells are the most common cells in the primary cortex. Hubel [1995] guesses that they make up to 3/4 of the entire cell population. Complex cells share with simple cells the property that they respond only to specifically oriented structures. Like simple cells they respond to a limited region of the visual field, but unlike simple cells they cannot be explained by a neat subdivision of the receptive field into excitatory and inhibitory regions. Also, complex cells tend to have larger receptive fields than simple cells, but not much larger. For building complex cells on top of simple cells, Hubel [1995] proposed several possible schemes, one of them being that the activation of a complex cell requires successive activations of simple cells. The current mathematical model is rather simple and explained in Chapter 3).

Hypercomplex cells form the third category of striate cells initially identified by Hubel and Wiesel. These possess inhibitory zones at one or both ends of oriented excitatory regions, thereby responding to bars of preferred orientation only if they are not too long. Many of them respond more to the end of an oriented edge, i.e., if the edge does not extend beyond a

specific part of the receptive field. Such cells are therefore called end-stopped cells, and there are single and double end-stopped cells with receptive fields as shown in Fig. 2.2d and e. The fields are composed of a activation regions and regions at one or both ends called inhibitory regions. The simplest scheme for modeling such a cell consists of assuming excitation by one or a few complex cells with fields in the activation region in combination with inhibition by other complex cells with similarly oriented fields situated at the neighboring regions [Hubel, 1995]. The entire next chapter is devoted to end-stopped cells, improved models and applications.

Disparity-tuned cells have also been identified [Hubel, 1995]. These can account for a horizontal displacement, or disparity, which can be tolerated, the maximum displacement being a fraction of the width of the receptive field. Responses of such cells are a function of the distance of an object, which translates into the relative positions of a stimulus pattern in the two eyes. There is evidence that disparity-tuned cells exist in V1 of monkeys [Cumming and Parker, 2000]. The fact that many simple and complex cells are also tuned to disparity opens the possibility that, at a very early processing stage, depth is attributed to lines and edges. In other words, the visual system might use some sort of "wireframe" representation of 3D objects, like the ones used in the modeling of solid objects in computer graphics.

There are many other types of cells. For example, there are grating cells that were discovered in areas V1 and V2 of the monkey visual cortex by von der Heydt et al. [1992]. Such cells respond vigorously to grating patterns of appropriate orientation and periodicity, but very weakly or not at all to isolated bars. On the other hand, bar cells, which are found in the same areas of the visual cortex [von der Heydt et al., 1992], have a functional behavior which is less well explored and documented in the literature. In general, bar cells respond to single bars and their responses decrease when further bars are added in the form of a periodic pattern [Petkov and Kruizinga, 1997]. Computational models inspired by bar and grating cells were used in pattern recognition, for example in texture analysis, see e.g. [Kruizinga and Petkov, 1999; du Buf, 2007].

Figure 2.3 shows one scheme for visualizing activities of even and odd simple cells, complex cells, single and double end-stopped cells, plus a saliency map (for an explanation and all details see Chapter 3), using different colors with the saturation corresponding to a cell's response strength. In addition, the dominant local orientation, which corresponds to the orientation of the complex cell with maximum amplitude, is coded by rotating the "color wheels." The center-left panel in Fig. 2.3 shows the scheme in the case that the dominant local orientation is horizontal. The colored circle is subdivided into four quadrants, and one quadrant is further divided into two octants. The red and blue quadrants show responses of even (B) and odd (F) simple cells, the green quadrant (A) shows responses of complex cells. The line (D) separating the two pinkish octants shows the dominant orientation (here horizontal). The upper and lower octants show responses of single (C) and double (E) end-stopped cells. The black dot (G) in the center shows the information in the saliency map related to Focus-of-Attention based on end-stopped cells (see Section 3.7). The advantage of using color saturation is that complex cells (green) with very low activity are displayed as white; hence, areas with no green component do not contain significant lines and edges, and therefore also no keypoints. In contrast, responses of even and odd simple cells can be positive or negative (white indicates a large but negative amplitude). Finally, the colored circles can be superimposed on the input image in order to see (part of) the underlying image structure. Figure 2.3 shows a real example, Fiona image, at the finest scale (top-left) and at a coarse scale (top-right). The bottom-right image shows a zoomed version of the area around the pupil. Such images, if printed at bigger size, are aesthetically appealing, but

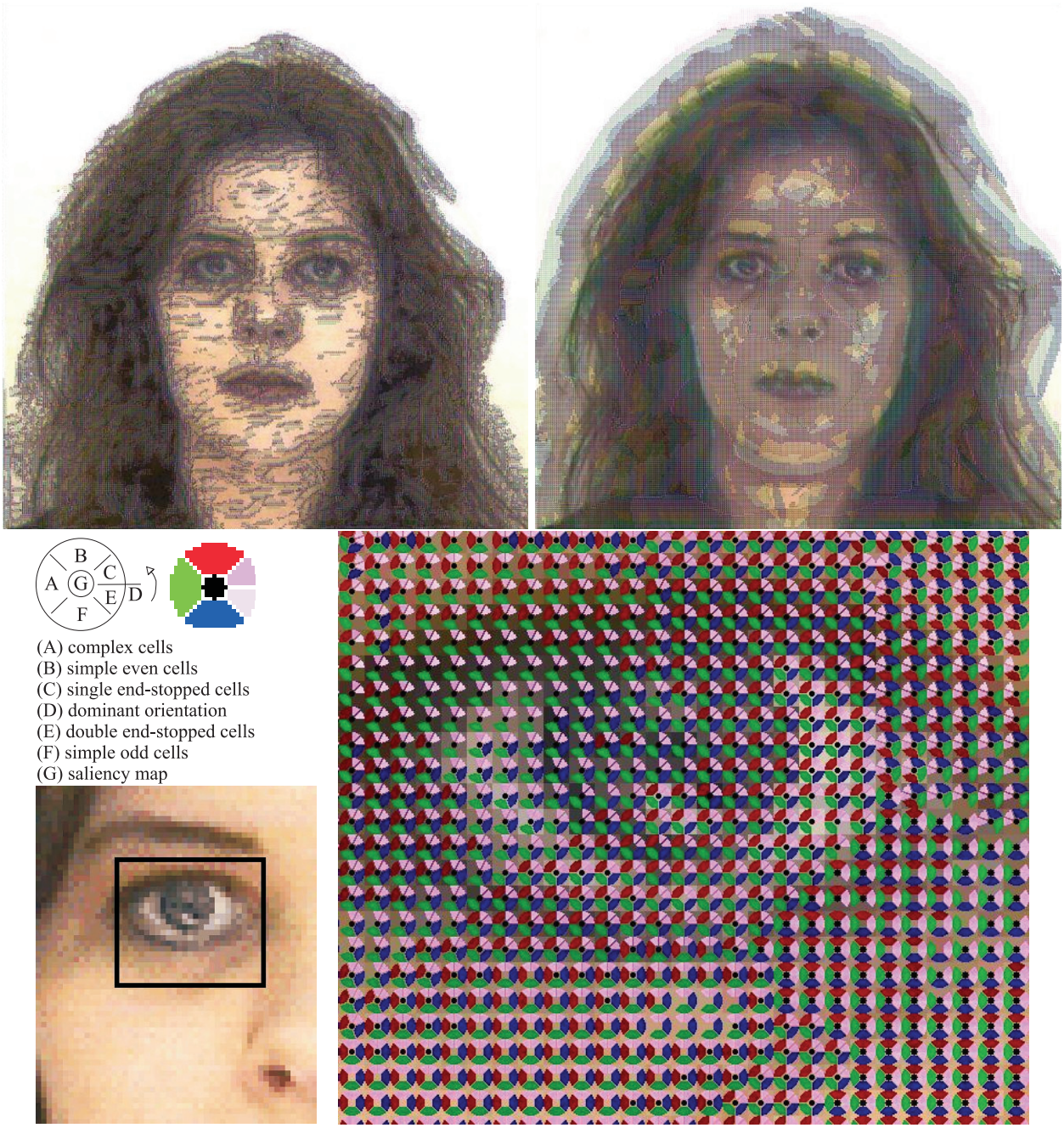

Figure 2.3: Color wheel visualization of cell responses. At the top a fine-scale (left) and a coarse-scale (right) representation of Fiona image. At bottom-right a zoomed area at the left eye.

more important is that we can analize the local image structure and we can see the responses of the cells in order to optimize the basic detection schemes of lines, edges and keypoints.

2.2.3 Modeling simple cells by Gabor functions

For simulating simple cells in a computational model we will use 2D Gabor functions as models of their receptive fields. Gabor functions or filters are also used in image processing and computer vision. The goal of this section is not to present in detail all the properties of

these functions, nor to explore all the existing computational models (for modelling simple cells). For this we refer to Chapter 2 of Peter Kruizinga's PhD thesis [Kruizinga, 1999], where this matter is exhaustively studied and described. Here we only briefly expose the reasons to use these functions.

Different models can be used to model simple cells. Kruizinga [1999] compares several models—Difference-of-Gaussians, Difference-of-offset-Gaussians, Sum-of-offset-DoGs, Derivatives of Gaussians, Hermite polynomials and Gabor functions—using three criteria: (a) the ability of a model to cover the properties of different cells, each one having different preferred orientations, spatial frequencies, phases and bandwidths; (b) the number of model parameters and their relations to the relevant proprieties of real cells; and (c) the biological plausibility of the scheme. Kruizinga states that, despite existing differences between the receptive field profiles of the models listed above, the differences are so subtle that no model can be rejected bacause of bad fits to neurophysiological data. Nevertheless he concludes that there are some differences between the models in their ability to model a large variety of types of simple cells, which is related to the number and nature of the parameters. He therefore chooses the Gabor model, because it is easier to relate parameters to essential receptive field proprieties.

Although 2D Gabor functions are now generally accepted as an appropriate model of receptive fields of simple cells, there still are some criticisms (see [Kruizinga, 1999]): the even-symmetric Gabor function has a non-zero DC response, or there are more than three side lobes, or there are too many parameters, but all this can be corrected or minimized. For the mathematical model see Chapter 3 of this thesis and see also [Lee, 1996; Kruizinga, 1999; Bruce et al., 2000; Grigorescu et al., 2003]. The mathematical models for complex and end-stopped cells are also presented in Chapter 3.

2.3 Invariance

Until here we have presented a brief overview of cortical biology involved, i.e., cells and the basic functionality of the visual areas, but our goal is to develop an integrated architecture for recognizing objects or persons, and this involves the identification of similar, yet distinct, objects as members of the same class.

In object recognition, one form of invariance requires a many-to-one mapping between individual exemplars and object categories. At the same time, individual exemplars of three-dimensional objects rarely appear in the same form one moment after the other. Variations in the two-dimensional images falling on our retinas arise from almost any change in viewing conditions, including changes of position, pose, lighting and object configuration. Therefore invariance also requires a many-to-one mapping between all individual "views" of objects and their unique identities [Tarr, 2005].

There are several approaches for establishing object representations which are suitable for obtaining invariance in the brain, although so far none has been considered to be the correct or final one. Here we will only focus on the most common, the view-based approach, because it is the one which will be explored in this thesis. For the "Recognition By Components" approach (RBC) we refer to e.g. [Bierderman, 1987; Tarr and Bülthoff, 1995], and for mental rotation to e.g. [Zacks et al., 2003; Hegarty and Waller, 2004].

2.3.1 View-based approach

The terms view and view-based encompass a variety of specific theories and computational models. However, all view-based models share a common, defining assumption: they all assume that objects are represented and matched to memory in terms of their features in a spatial reference frame [Hummel, 2000]. The central point of the view-based approach is that we represent objects in long-term memory as views, and that by means of operations on the coordinates of the features in those views we bring new views into register with stored views, or into register with stored 3D models. The basic idea is that we recognize objects on the basis of stored views, by matching images to the templates stored in memory. This is the most common theory in object recognition; for variations on the same theme see [Lowe, 2004; Peters, 2004; Tarr, 2005], and see [Peters, 2000] for a survey on theories of three-dimensional object perception.

Tarr [1995] emphasizes that experience with particular views is a critical factor in achieving invariance. He found that when observers learn how to recognize novel objects from specific viewpoints, they are both faster and more accurate at recognizing the same objects from familiar viewpoints relative to unfamiliar viewpoints. Moreover, recognition performance in the case of unfamiliar viewpoints is systematically related to the views which are familiar: observers take progressively more time and are progressively less accurate when the distance between the unfamiliar and the familiar views increases. These and related results from Tarr and colleagues [Tarr et al., 1998] suggest that human object recognition relies on multiple views, where each view encodes the appearance of an object under specific viewing conditions, including viewpoint, pose, configuration and lighting, and that a collection of such views constitutes the mental representation of a given object.

In order to explain how view-based invariance can be achieved, Tarr [2005] refers to the work of Perrett, Oram and Ashbridge [Perrett et al., 1998], who found that individual object-selective neurons preferentially respond to particular object views. Invariance is then achieved by considering populations of such neurons as the actual neural code for objects. In this context, individual neurons may be considered as coding—from a familiar viewpoint—the complex features or parts of which objects are composed. Recognition then takes the form of "accumulation of evidence" across all neurons that are selective for some aspect of a given object. During recognition, the particular rate of accumulation will depend on the similarity between visible features/parts in the present viewpoint and the view-specific features/parts to which individual neurons are tuned [Perrett et al., 1998]. Across a population of object-selective neurons, sufficient neural evidence (summed activities of neurons) will accumulate more slowly when the current appearance of an object is dissimilar from all its learned appearances. Tarr [2005] himself concludes (assumes) from this that when an object's appearance is close to previously-experienced views, evidence across the appropriate neural population must accumulate more rapidly. Thus, systematic behavioral changes in recognition performance under changes of viewpoint may be explained as a consequence of how similarity is computed between new object percepts and previously learned neural representations.

Summarizing, Tarr [2005] concludes that recognition amounts to reaching a threshold of sufficient evidence in terms of activity across a neural population. One consequence of this is that unfamiliar views of objects will require more time to reach threshold, but will be successfully recognized given some similarity between input and known viewpoints. A second consequence is that unfamiliar exemplars within a familiar class will be likewise recognized given some similarity (similar configurations and viewpoints) with known exemplars from

within that class. One implication is that familiarity with individual objects should facilitate the viewpoint-dependent recognition of other, visually similar objects [Tarr and Gauthier, 1998]. A second implication is that object viewpoints or class exemplars that are significantly different from known views or objects should be represented as distinct representations; again, a prediction that seems to be supported.

2.4 Initial conclusions

From the relevant literature (see also Sections 3.1, 4.1 and 5.2) some further aspects may guide us toward developing an invariant object categorization and recognition architecture: (1) Extracted features play an important role in a biological model, both for characterizing the most significant aspects that are present and for abstraction of the scene [Heitger et al., 1992; Olshausen et al., 1993; van Deemter and du Buf, 2000; Corchs and Deco, 2005]. (2) The two major visual pathways consist of the dorsal "where" stream that runs from V1 via V2, V3 and MT to PP, and the ventral "what" stream that runs from V1 via V2 and V4 to IT [Goodale and Milner, 1992; Deco and Rolls, 2005]. (3) Two-dimensional Gabor functions are now generally accepted as an appropriate model of receptive fields of simple cells [Kruizinga, 1999; Grigorescu et al., 2003], and this model provides the basis to model other types of cells: complex and end-stopped cells, bar and grating cells, etc. [Heitger et al., 1992; Petkov and Kruizinga, 1997; du Buf, 2007]. (4) Object recognition is most probably a multi-level task which includes categorization and identification [Grill-Spector and Kanwisher, 2005], and it starts as soon as we have the "gist" of a scene [Rensink, 2000; Grill-Spector and Kanwisher, 2005; Oliva and Torralba, 2006]. (5) It is very likely that objects in memory are represented by templates in "view-based" form, i.e., one object must be represented by multiple, canonical views; e.g. [Tarr, 2005]. (6) There exist cortical top-down (and feedback) mechanisms which trigger and control visual attention, and which facilitate object recognition [Hupe et al., 2001; Bar et al., 2006; Oliva and Torralba, 2006]. (7) There is evidence for at least four, now generally accepted properties of the feedforward path of the ventral what stream [Serre et al., 2005]: (a) a hierarchical use of invariances, first to position and size (importantly, size and position invariance—over a restricted range—do not require learning specific for a given object), and then to viewpoint and other transformations (invariances to viewpoint, illumination etc. do require learning of several, different views of an object); (b) an increasing size of the receptive fields of cells coupled to an increasing complexity of their optimal stimuli throughout the cortical layers; (c) a basic feedforward processing of information for "immediate" recognition tasks; and (d) plasticity and learning, probably at all stages, but with a time scale that decreases from V1 to IT and PF cortex: fast adaptation to objects at high level and slower adaptation to local features at low level.

More specific details and references concerning recognition models are given in Chapter 5 Section 5.2. In the following three chapters we will work toward an integration of features into a computational recognition scheme. Chapters 3 and 4 are about individual features, i.e., the multi-scale keypoint and line/edge representations, and for which purposes they can be exploited. In Chapter 5 we combine and test all the available information in the integrated architecture.

Chapter 3

Multi-scale keypoints in V1 and beyond

Abstract: End-stopped cells in cortical area V1, which combine outputs of complex cells tuned to different orientations, serve to detect line and edge crossings, singularities and points with large curvature. These cells can be used to construct retinotopic keypoint maps at different spatial scales (Level-of-Detail). The importance of the multi-scale keypoint representation is studied in this chapter. It is shown that this representation provides very important information for object recognition and face detection. Different grouping operators can be used for object segregation and automatic scale selection. Saliency maps for Focus-of-Attention can be constructed. Such maps can be employed for face detection by grouping facial landmarks at eyes, nose and mouth. Although a face detector can be based on processing within area V1, it is argued that such an operator must be embedded into dorsal and ventral data streams, to and from higher cortical areas, for obtaining translation-, rotation- and scale-invariant detection.

3.1 Introduction

Our visual system can still be seen as a huge puzzle with a lot of missing pieces. Even in the first processing layers in area V1 of the visual cortex there remain many gaps, despite all knowledge already compiled [Hubel, 1995; Bruce et al., 2000; Rasche, 2005]. Nevertheless, some of the gaps are being filled by developing and studying computational models. Models of simple, complex and end-stopped cells have been developed more than ten years ago [Heitger et al., 1992]. Several inhibition models [Petkov et al., 1993b; Grigorescu et al., 2003], keypoint detection [Heitger et al., 1992; Würtz and Lourens, 2000; Rodrigues and du Buf, 2004b] and line/edge detection schemes [van Deemter and du Buf, 2000; Grigorescu et al., 2003; Elder and Sachs, 2004; Rodrigues and du Buf, 2004b], including disparity models [Fleet et al., 1991; Rodrigues and du Buf, 2004a], have become available. On the basis of such models and neural processing schemes, it is possible to create a cortical architecture for figure-ground segregation [Hupe et al., 2001; Rodrigues and du Buf, 2006a] and visual

attention or Focus-of-Attention (FoA) [Parkhurst et al., 2002; Rodrigues and du Buf, 2005b; Carmi and Itti, 2006]. In addition, object detection, categorization and recognition can be obtained by means of bottom-up and top-down data streams in the so-called "what" and "where" subsystems [Rensink, 2000; Deco and Rolls, 2004; Rodrigues and du Buf, 2006a].

We will focus exclusively on keypoints in this chapter. Heitger et al. [1992] developed a single-scale basis model that consists of single and double end-stopped cells in combination with complex inhibition schemes. Lourens and Würtz [1997] and Rodrigues and du Buf [2004b] presented a pseudo-multi-scale approach, in which detection stabilization at a fine scale is obtained by averaging keypoint positions over a few neighboring, coarser, micro-scales. A truly multi-scale analysis was introduced later [Rodrigues and du Buf, 2005b]. This idea was based on the fact that there are simple and complex cells tuned to different spatial frequencies, spanning multiple octaves; therefore, it can be expected that also end-stopped cells exist at all frequencies. We analyzed the multi-scale keypoint representation, from very fine to very coarse scales, in order to study its importance and possibilities for developing a cortical architecture, with an emphasis on FoA. Also, a new aspect was included, namely the application of non-classical receptive-field (NCRF) inhibition to keypoint detection. Before, NCRF inhibition had only been applied to contour detection [Grigorescu et al., 2003], in order to separate object structures from surface textures. Below, we will argue that NCRF inhibition can be applied to edges *and* to keypoints, for creating two data streams dedicated to object structures and surface textures, but *only* at the finest scales. Furthermore, we will show that the multi-scale keypoint representation can be combined with automatic scale selection, for obtaining keypoints which are most characteristic of objects, and that it can play a role in object segregation. The latter two processes are thought to be essential in the what and where subsystems, for a rapid detection of where an object may be and a first categorization to select most likely object templates in memory, after which all available features are used in object recognition.

A difficult and still challenging application, even in computer vision, is face detection. Despite the impressive number of methods devised for faces and facial landmarks [Yang et al., 2002], complicating factors are pose (frontal vs. profile), beards, moustaches and glasses, facial expression and image conditions (lighting, resolution). Despite these complications, we will study the multi-scale keypoint representation in the context of a plausible architecture for face detection. We add that we will not employ the multi-scale line/edge representation that also exists in area V1, in order to emphasize the importance of the information provided by keypoints. Also, we will not solve all complications referred to above, because we will argue, in the final Discussion, that low-level processing in area V1 needs to be embedded into a much wider context, including object templates stored in short- and long-term memory, and this context is expected to solve many problems.

There exists a vast literature concerning keypoints in computer vision, from basic feature extraction to object recognition, but much less in biological vision. Here we summarize a few approaches. Lourens and Würtz [1997] presented an object recognition system based on symbolic graphs, in which object corners are nodes and object contours are edges of the graphs. Their algorithm for corner detection is based on Heitger et al.'s model of cortical end-stopped cells, but they combined several scales and generalized to color channels [Würtz and Lourens, 2000]. Resulting corner detection was shown to be very stable in the presence of high-frequency textures, noise, varying contrast and rounded corners, see also Lourens et al. [2001] and Lourens and Würtz [2003]. In this processing, graph edges are constructed by following contours between corners, using local evidence from the multi-scale Gabor wavelet transform. Model matching is achieved by finding subgraph isomorphisms in global image

graphs.

Rosenthaler et al. [1992] also presented an integrated framework for extracting edges and keypoints. This detection scheme is based on analysis of oriented energy channels by using differential geometry. Barth et al. [1998] proposed end-stopped operators based on iterative, non-linear center-surround inhibition. Henricsson and Heitger [1994] showed that an independent representation of corner and junction features provides suitable stop conditions for an aggregation process which allows to divide contours into meaningful substrings. They demonstrated that the active role of corners and junctions in the linking of contours greatly reduces problems associated with purely edge-based methods.

Lindeberg [1999] presented a detailed study of the Gaussian-derivative scale-space representation that can be used for a variety of early visual tasks. Operations like feature detection, which includes keypoints, feature classification and shape computation can be directly expressed in terms of (non-linear) combinations of Gaussian derivatives at multiple scales. Hansen et al. [2001] developed a functional model of intra-cortical, recurrent, long-range interactions in V1 and proposed that long-range connections implement a multi-purpose preprocessing mechanism for main vision tasks, namely contour enhancement and corner detection. Later, Hansen and Neumann [2002] compared detected junctions based on the recurrent long-range interactions to junctions as obtained by a purely feed-forward model of complex cells. They also compared with two widely-used junction-detection schemes in computer vision, which are based on Gaussian curvature and the structure tensor. Ruzon and Tomasi [2001] used color distributions to detect edges, junctions and corners, whereas Kovesi [2003] described corner and edge detection on the basis of the phase-congruency model. Triggs [2004] demonstrated that keypoints detected by the Förstner-Harris method are very stable when changing the illumination.

In addition to all different views and ideas referred to above, we mention two special projects. The first has no biological background, whereas the second has some minor biological background. The SUSAN project [Smith and Brady, 1997] concerns an approach to edge and corner detection with structure-preserving noise reduction. Non-linear filtering is used to define which parts of the image are closely related to each individual pixel, where each pixel is associated to a local image region which has about the same intensity (pixel values). Feature detectors are based on the minimization of these local image regions, and the noise-reduction method uses the regions as smoothing neighborhoods. The SIFT project [Lowe, 2004] has seen many developments along the years, for instance the extraction of distinctive image features from scale-invariant keypoints. Distinctive, invariant image features can be used for a reliable matching of different views of an object or a scene.

Most methods presented above have no direct biological background, and those *with* a clear biological background [Heitger et al., 1992; Barth et al., 1998; Hansen et al., 2001] are limited to one, fine scale. The only exceptions are the papers by Lourens and Würtz referenced above, in which a few (fine) scales are used for keypoint stabilization. Furthermore, many methods are concerned with low-level feature extraction, for example for solving problems related to edge detection by employing keypoints. Extracted features are then used for high-level object detection in images, for example. In this chapter, we study keypoint scale space, from the finest to very coarse scales, and show that this space can be exploited in building biological—and computer—vision systems.

This chapter is organized as follows: Section 3.2 introduces basic cell models and NCRF inhibition, Section 3.3 presents keypoint detection with NCRF inhibition at fine scale, and multi-scale representation in section 3.4. Section 3.5 deals with object segregation and Section 3.6 with automatic scale selection. Section 3.7 is about Focus-of-Attention by saliency

maps, followed by face detection (facial landmarks) in Section 3.8. We conclude with a discussion in Section 3.9.

3.2 Basic cell models and NCRF inhibition

Gabor quadrature filters provide a model of cortical simple cells [Lee, 1996]. In the spatial domain (x, y) they consist of a real cosine and an imaginary sine, both with a Gaussian envelope. A receptive field (RF) is denoted by (see for example [Grigorescu et al., 2003])

$$G_{\lambda,\sigma,\theta,\varphi}(x,y) = \exp\left(-\frac{\tilde{x}^2+\gamma\tilde{y}^2}{2\sigma^2}\right)\cdot\cos(2\pi\frac{\tilde{x}}{\lambda}+\varphi), \tag{3.1}$$

with $\tilde{x} = x\cos\theta + y\sin\theta$ and $\tilde{y} = y\cos\theta - x\sin\theta$, the aspect ratio $\gamma = 0.5$ and σ determines the size of the RF. The spatial frequency is $1/\lambda$, λ being the wavelength. For the bandwidth σ/λ we use 0.56, which yields a half-response width of one octave. The angle θ determines the orientation (we use 8 orientations), and φ the symmetry (0 or $-\pi/2$). We can apply a linear scaling between $f_{\min}$ and $f_{\max}$ with hundreds of contiguous scales. Below, the scale of analysis will be given in terms of λ expressed in pixels, where $\lambda = 1$ corresponds to 1 pixel. Most images shown in this thesis have a size of 256×256 pixels.

Responses of even and odd simple cells, which correspond to real and imaginary parts of a Gabor filter, are obtained by convolving the input image with the RFs, and are denoted by $R^E_{s,i}(x,y)$ and $R^O_{s,i}(x,y)$, s being the scale, i the orientation ($\theta_i = i\pi/(N_\theta - 1)$) and N_θ the number of orientations (here 8). Responses of complex cells are then modelled by the modulus

$$C_{s,i}(x,y) = [\{R^E_{s,i}(x,y)\}^2 + \{R^O_{s,i}(x,y)\}^2]^{1/2}. \tag{3.2}$$

There are two types of end-stopped cells [Heitger et al., 1992], single (S) and double (D). If $[\cdot]^+$ denotes the suppression of negative values, and $\mathcal{C}_i = \cos\theta_i$ and $\mathcal{S}_i = \sin\theta_i$, then

$$S_{s,i}(x,y) = [C_{s,i}(x + d\mathcal{S}_{s,i}, y - d\mathcal{C}_{s,i}) - C_{s,i}(x - d\mathcal{S}_{s,i}, y + d\mathcal{C}_{s,i})]^+ \tag{3.3}$$

and

$$\begin{aligned} D_{s,i}(x,y) \;=\; & \Big[C_{s,i}(x,y) - \frac{1}{2}C_{s,i}(x + 2d\mathcal{S}_{s,i}, y - 2d\mathcal{C}_{s,i}) \\ & - \frac{1}{2}C_{s,i}(x - 2d\mathcal{S}_{s,i}, y + 2d\mathcal{C}_{s,i})\Big]^+ . \end{aligned} \tag{3.4}$$

The distance d is scaled linearly with the filter scale s (we use $d = 0.6s$). Figure 3.1 shows end-stopped responses at three scales in the case of the traffic-sign image shown in Fig. 3.2. These responses mark the triangle and arrow etc. at a fine scale, but at coarser scales they are very diffuse due to the size of the RFs. In the next step, all end-stopped responses along straight lines and edges are suppressed, for which tangential (T) and radial (R) inhibition are used:

$$I^T_s(x,y) = \sum_{i=0}^{2N_\theta-1} [-C_{s,i \bmod N_\theta}(x,y) + C_{s,i \bmod N_\theta}(x + d\mathcal{C}_{s,i}, y + d\mathcal{S}_{s,i})]^+ \tag{3.5}$$

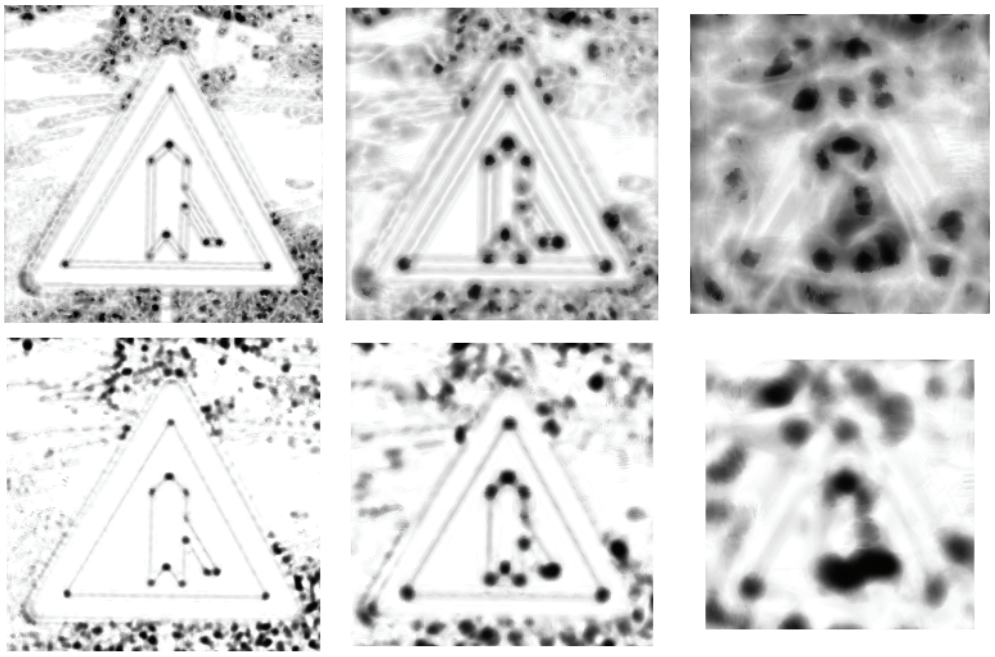

Figure 3.1: Single (top) and double (bottom) end-stopped responses at three scales ($\lambda = 4, 8, 16$).

and

$$I_s^R(x,y) = \sum_{i=0}^{2N_\theta-1} \Big[C_{s,i \bmod N_\theta}(x,y) - 4 \cdot C_{s,(i+N_\theta/2) \bmod N_\theta}(x + \frac{d}{2}\mathcal{C}_{s,i}, y + \frac{d}{2}\mathcal{S}_{s,i}) \Big]^+, \tag{3.6}$$

where $(i + N_\theta/2) \bmod N_\theta \perp i \bmod N_\theta$.

Non-classical receptive-field (NCRF) inhibition can be applied to suppress keypoints in textured regions. Models of NCRF inhibition are explained in more detail by Grigorescu et al. [2003]. There are two inhibition types: (a) anisotropic, in which only responses obtained for the same preferred RF orientation contribute to the suppression, and (b) isotropic, in which all responses over all orientations contribute equally to the suppression.

The anisotropic NCRF (A-NCRF) model is computed by an inhibition term $t^A_{s,\sigma,i}$ for each orientation i, as a convolution of the complex cell responses $C_{s,i}$ with the weighting function w_σ, with

$$w_\sigma(x,y) = [\mathrm{DoG}_\sigma(x,y)]^+ / \|[\mathrm{DoG}_\sigma]^+\|_1, \tag{3.7}$$

where $\|\cdot\|_1$ is the L_1 norm and

$$\mathrm{DoG}_\sigma(x,y) = \frac{1}{2\pi(4\sigma)^2} \exp(-\frac{x^2+y^2}{2(4\sigma)^2}) - \frac{1}{2\pi\sigma^2} \exp(-\frac{x^2+y^2}{2\sigma^2}). \tag{3.8}$$

The operator $b^A_{s,\sigma,i}$ corresponds to the inhibition of $C_{s,i}$, i.e. $b^A_{s,\sigma,i} = [C_{s,i} - \alpha t^A_{s,\sigma,i}]^+$, with α controlling the strength of the inhibition.

The isotropic NCRF (I-NCRF) model is obtained by computing the inhibition term $t^I_{s,\sigma}$ which does not depend on orientation i. For this the maximum response map of the complex

Figure 3.2: Keypoints detected at the finest scale, without (center) and with (right) NCRF inhibition.

cells is constructed: $\tilde{C}_s = \max\{C_{s,i}\}$, with $i = 0, ...N_\theta - 1$. The isotropic inhibition term $t^I_{s,\sigma}$ is computed by the convolution of the maximum response map $\tilde{C}_s$ with the weighting function w_σ, and the isotropic operator is $b^I_{s,\sigma} = [\tilde{C}_s - \alpha t^I_{s,\sigma}]^+$.

3.3 Keypoint detection with NCRF inhibition at fine scale

As already mentioned, NCRF inhibition permits to suppress keypoints which are due to texture, for example in textured parts of an object surface. We experimented with the two types of NCRF inhibition introduced above, but here we only present the best results which were obtained by I-NCRF at the finest scale.

All responses of the end-stopped cells $S_s(x,y) = \sum_{i=0}^{N_\theta-1} S_{s,i}(x,y)$ and $D_s(x,y) = \sum_{i=0}^{N_\theta-1} D_{s,i}(x,y)$ are inhibited by $b^I_{s,\sigma}$, where $\alpha = 1$ is used, and we obtain the responses $\tilde{S}$ and $\tilde{D}$ of S and D that are above a small threshold of $b^I_{s,\sigma}$. Then we apply $I_s = I^T_s + I^R_s$ for obtaining the keypoint maps $\tilde{K}^S_s(x,y) = \tilde{S}_s(x,y) - gI_s(x,y)$ and $\tilde{K}^D_s(x,y) = \tilde{D}_s(x,y) - gI_s(x,y)$, with $g \approx 1.0$, and the final keypoint map $\tilde{K}_s(x,y) = \max\{\tilde{K}^S_s(x,y), \tilde{K}^D_s(x,y)\}$. In the last step, local maxima of $\tilde{K}_s(x,y)$ in x and y are detected.

Figure 3.2 shows, left to right, input images and keypoints detected at the finest scale that will be used in this chapter, $\lambda = 4$, before and after I-NCRF inhibition. The face image ("face196") is part of the Psychological Image Collection at Stirling University (UK). As can be seen in Fig. 3.2, keypoint detection is very precise and mostly contour-related keypoints remain after inhibition. Although many texture-related keypoints have been suppressed, some may still appear because of strong, local contrast; see also Rodrigues and du Buf [2005b].

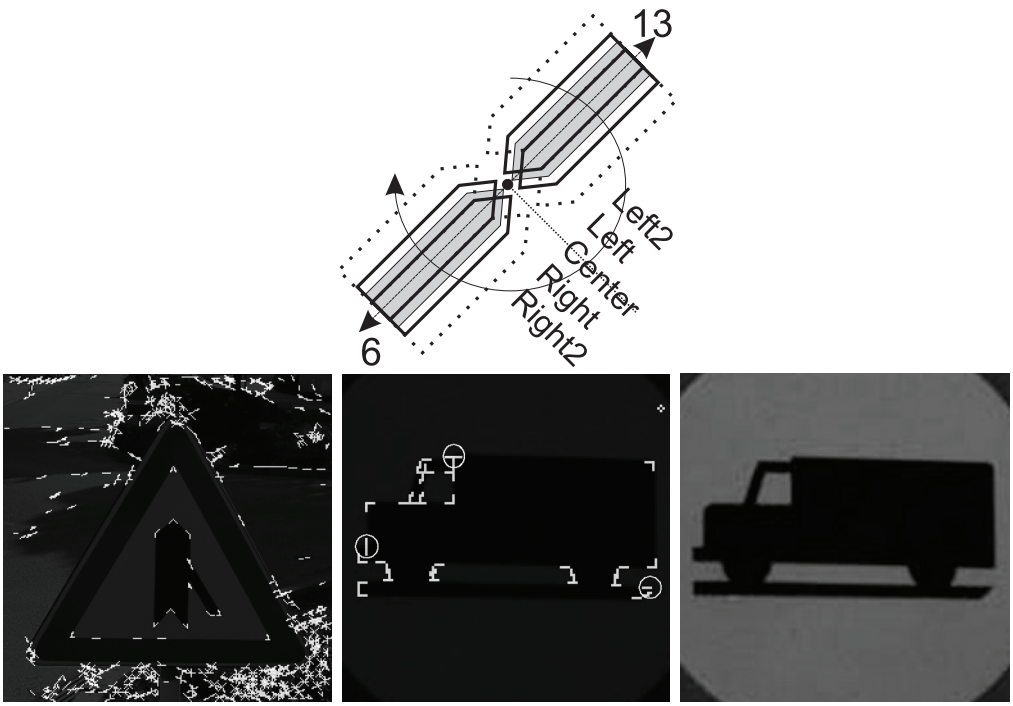

Figure 3.3: Keypoint classification. Top: pentagonal dendritic fields of grouping cells that probe simple and complex cells for (sub)dominant orientations and (a)symmetric directions. Bottom: van original image (right) and detected keypoints with vertex structures.

Detected keypoints provide important image information because they code local image complexity, for example for FoA (see below), but we can go one step further. Object detection and recognition is helped much if detected positions are complemented by the type of complexity. In other words, it is useful to classify keypoints according to the underlying vertex structure, such as K, L, T, +, etc. This is very difficult, because responses of simple and complex cells, which code the underlying lines and edges at the vertices, are unreliable due to response interference effects [du Buf, 1993]. This implies that responses must be analyzed in a larger neighborhood around each keypoint. This problem has been solved by processing simple- and complex-cell responses in four cell layers, each layer comprising various grouping and detection cells. This process is very close to basic line and edge detection, see Rodrigues and du Buf [2004b], which is beyond the scope of this chapter.

Figure 3.3 (top) shows two central, pentagonal, dendritic fields (shaded) and eight parallel ones around a keypoint, for directions 6 and 13. Grouping cells with such fields are necessary for probing simple and complex cells for dominant and sub-dominant *orientations* and then for symmetric or asymmetric *directions*; see Rodrigues and du Buf [2004b] for a detailed explanation. Figure 3.3 (bottom) illustrates the application of keypoint classification to two traffic signs, at scale $\lambda = 4$. All keypoints of the "van" image have been detected, but three directions are still missing (encircled). There, structures have a size of 2 to 4 pixels, and we are at the very limit of what can be achieved by using Gabor filters. Also present in the "van" image is a keypoint (small diamond) that was detected near the top-right corner, but due to the lack of structure in its neighborhood no direction has been attributed. In other words, this keypoint can be suppressed. It follows from Fig. 3.3 that detected and classified keypoints provide important information for object recognition, in this case the triangular sign with the arrow and the "van." This information must be complemented by lines and

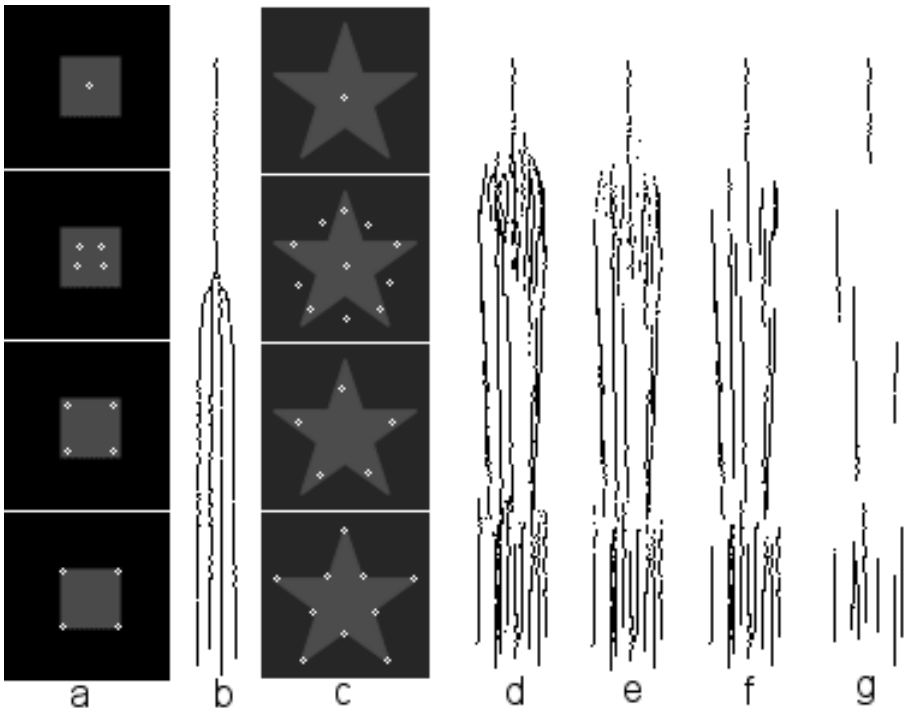

Figure 3.4: Keypoint scale space, with finest scale at the bottom: (a) square, (b) projected 3D keypoint trajectories of square, (c) and (d) star and projected trajectories, (e) micro-scale stability, (f) and (g) stability over at least 10 and 40 scales, respectively.

edges that are also extracted in area V1. Currently, the keypoint classification scheme is being implemented and optimized for application at arbitrary scale, but it is not yet clear whether vertex structure provides useful information in addition to detected lines and edges at coarse scales [Rodrigues and du Buf, 2004b].

3.4 Multi-scale keypoint representation

Although NCRF inhibition can be applied at any scale, we will not do this for two reasons: (a) we want to study keypoint behavior in scale space for applications like FoA and facial landmark detection, and (b) in many cases a coarser scale, or increased RF size, will automatically eliminate keypoints in fine textures. In the multi-scale case keypoints are detected the same way as done above, but now by using $K_s^S(x,y) = S_s(x,y) - gI_s(x,y)$, $K_s^D(x,y) = D_s(x,y) - gI_s(x,y)$ and the final map $K_s(x,y) = \max\{K_s^S(x,y),\, K_s^D(x,y)\}$.

For analyzing keypoint stability we can create an almost continuous, linear, scale space. In the case of Fig. 3.4, which shows projected trajectories of detected keypoints over scale in the case of a square and a star object, we applied 288 scales with $4 \leq \lambda \leq 40$. Figure 3.4 illustrates the general behavior: at fine scales contour keypoints are detected, at coarser scales their trajectories converge, and at very coarse scales there is only one keypoint left near the center of the object. However, it can also be seen (star object) that there are scale intervals where keypoints are unstable, even scales at which keypoints disappear and other scales at which they appear. (Dis)appearing keypoints are due to the size of the RFs in relation to the structure of the objects, analogous to Gaussian scale space [Koenderink, 1984; Lindeberg, 1994]. Unstable keypoints can be eliminated by (a) requiring stability over

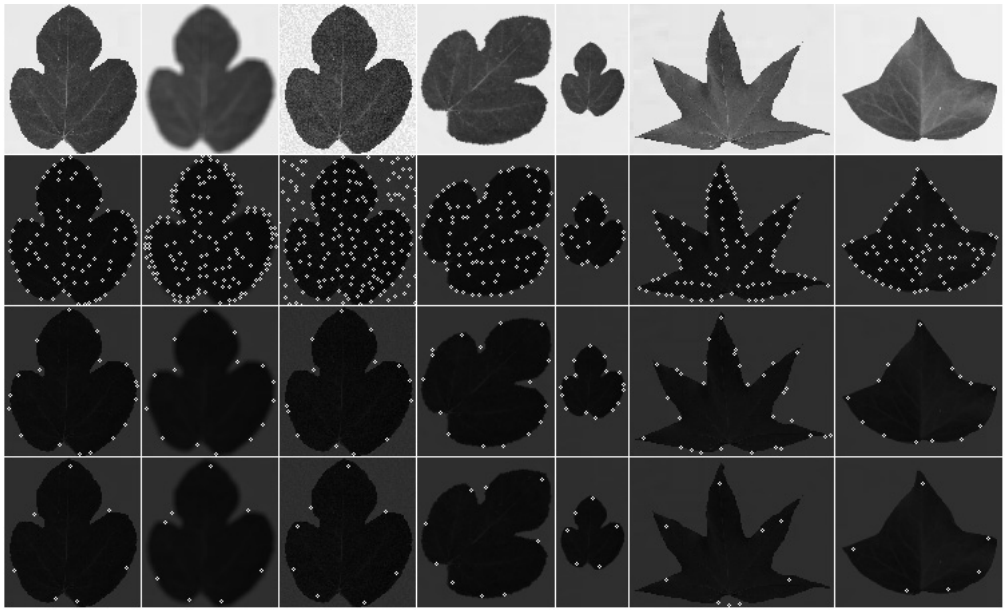

Figure 3.5: From left to right: ideal image, blurred, with added noise, rotated and re-scaled leaf, plus two other leaves. Keypoints detected without NCRF inhibition, at fine (2nd line) and medium scales (bottom two lines).

a few neighboring micro-scales [Rodrigues and du Buf, 2004b], by keeping keypoints that do not change position over 5 scales, the center one and two above plus two below (Fig. 3.4e), or (b) requiring stability over at least N_s neighboring scales (Figs 3.4f and g with $N_s = 10$ and 40, respectively). Such stabilizations are obtained by employing grouping cells with linear dendritic fields of different sizes over scale s. Assuming that keypoint cells are binary—they respond or they don't—grouping cells at all scales "sum" active keypoint cells, and if the sum (count) is below the necessary sum they can inhibit the keypoint cells. When keypoint cells may not be inhibited because of other processes, such as the ones described in the following sections, the grouping cells can inhibit gating cells which relay axons of keypoint cells.

The five leftmost columns in Fig. 3.5 illustrate that similar results are obtained after blurring, adding noise, rotation and rescaling of an object, a tree leaf, whereas the last two columns show results for other leaf shapes. In all cases, important contour keypoints remain at medium scales, and texture keypoints disappear without applying NCRF inhibition. In other words, NCRF inhibition is only useful for suppressing texture keypoints at the finest scales.

3.5 Object segregation

The "bandwidth" of the what and where subsystems is very limited, because only one object can be attended at any time, which explains for example change blindness [Rensink, 2000]. Both subsystems are "fed," bottom-up, by representations in area V1, and are "steered," top-down, from prefrontal (PF) cortex with templates of expected objects and expected positions [Deco and Rolls, 2004]. The faster the bottom-up and top-down data streams converge, the faster an object will be detected and recognised. Typically, objects are recognized within

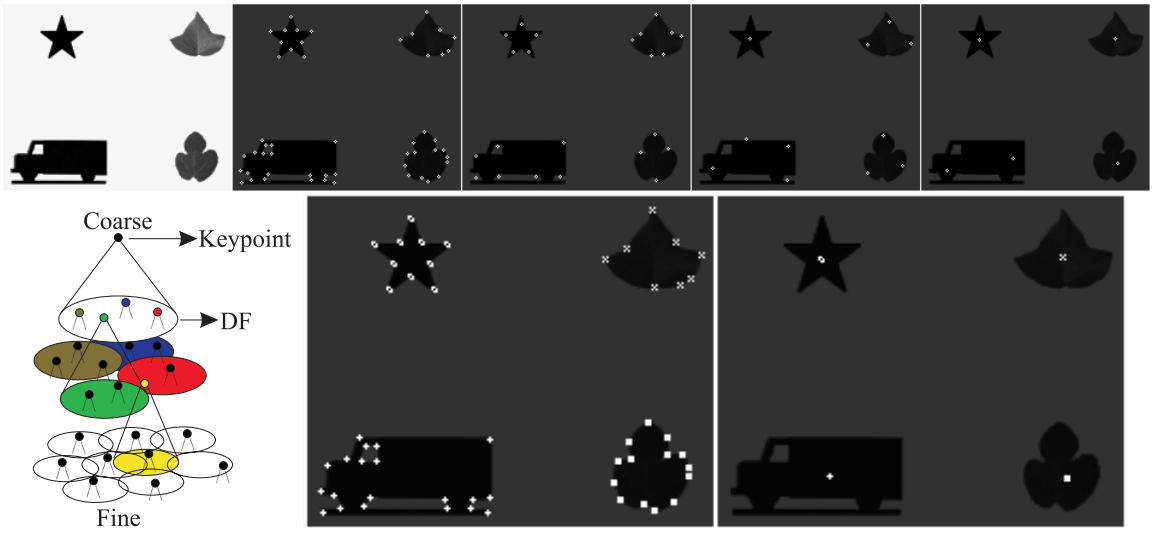

Figure 3.6: Object segregation. Top: input image with four objects and detected keypoints at four scales ($4 \leq \lambda \leq 50$). Bottom: linking keypoints at a very coarse scale (right) to the finest scale (center), with the principle (left; DF means dendritic field).

150–200 ms, and first category-specific activation of PF cortex starts after about 100 ms [Bar, 2003]. This implies that some information propagates very rapidly from V1 to PF, such that the where system can select possible positions, after which the what system can test hypotheses. An important aspect in this is segregation, i.e., the separation of objects and the grouping of object features. Keypoints may play an important role in this process.

We have seen (Fig. 3.4) that keypoint trajectories converge from the contours at fine scales to the centers of objects at coarse scales. This implies that object segregation by means of a coarse-to-fine-scale strategy is feasible. Figure 3.6 (top) shows an image with four objects, two tree leaves, a star and the van from the traffic sign. Again, at very coarse scales the keypoints are located near the centers of the objects. In the case of the elongated van, an even coarser scale is required in order to obtain only one keypoint in the center. Going from coarse to fine scales, keypoints will indicate more and more detail, until the finest scale is reached at which essential landmarks on contours remain.

At the coarsest level, each keypoint corresponds to one object. Each keypoint at a coarse scale is related to one or more keypoints at one finer scale, which can be slightly displaced. This relation is modelled by down-projection using grouping cells with a circular dendritic field, the size of which defines the region of influence. A responding keypoint cell activates a grouping cell. Only if the grouping cell is also excitated by responding keypoint cells one level lower, a grouping cell at the lower level is activated. This is repeated until the finest scale. Figure 3.6 illustrates the principle (bottom-left) with cones and the result (bottom-center). The labels of the four keypoints at the coarsest scale, represented by different symbols, have been attributed to the keypoints at the finest scale. This coarse-to-fine-scale process permits to link all keypoints belonging to the same object. Results shown were obtained with $\lambda \in [4, 50]$ and $\Delta\lambda = 4$.

A process as described above is supposed to occur completely in areas V1 and V2, although information—including keypoints—at coarse scales propagates faster than information at fine scales to inferior-temporal (IT) cortex [Bar, 2003]. This could imply that segregation is a dynamic effect or that it contributes dynamically to high-level object cate-

gorization, which starts with coarse scales and is refined by adding finer scales. In any case, this process must be complemented by the what subsystem, because if two or more objects are very close, detected keypoints at very coarse scales will group the objects together and they can only be separated by probing specific object templates at finer scales. How this is done is not yet clear, because of feedback from higher areas like MT [Hupe et al., 2001], but it is done within 80 ms after image onset, which is late enough to allow contributions from higher visual areas [Zhaoping, 2003].

3.6 Automatic scale selection

Apart from object segregation, other processes may play an important role in the fast where and slower what subsystems. Concentrating on keypoints—ignoring other features extracted in V1—there may be many scales and the tremendous amount of information may not propagate in parallel and at once to IT and PF cortex. It might be useful that keypoints which are most characteristic for an object are extracted and that these propagate first, for example for rapid object categorization. Above (Fig. 3.4) we have seen that different criteria for spatial stability over scales lead to different keypoint selections. One possibility is to select only one scale with the most characteristic keypoints. In computer vision, a similar approach has been applied by Lindeberg [1999], who selected the scale at which responses of Gaussian-derivative operators were strongest.

Here we propose that the scale is the one at which the maximum number of *stable* keypoints is detected. This can be achieved with a few, simple processes, in which we assume again that outputs of keypoint cells are binary. First, a retinotopic map by means of grouping cells is created; see also below, i.e., saliency maps for FoA. A diagram of keypoint, grouping and gating cells is shown in Fig. 3.7. The grouping cells marked A have linear dendritic fields (solid black lines) that connect to keypoint cells (solid dots; active cells are big dots). These grouping cells sum all active keypoint cells at their position, over scale, which yields a sort of histogram. Second, at each scale, active keypoint cells activate gating cells (triangular synapses next to open circles). These cells gate the outputs of grouping cells A (black dash-dotted axons) in the "histogram map" at the same position. Third, at each scale, other grouping cells (marked B) sum outputs of all gating cells. In other words, the latter grouping cells "count" stable keypoints at all individual scales. Fourth, the grouping cell with maximum activity is selected (winner takes all) and its axon activates other gating cells that gate outputs of keypoint cells at its scale. The outputs of the latter gating cells (Fig. 3.7, at top) provide the map which has the maximum number of stable keypoints. In the first step of this process, a scale-stability criterion as illustrated in Figs 3.4e, f or g can be included.

Figure 3.8 (top-left) shows the traffic-sign image with keypoints selected without applying a scale-stability criterion, which resembles detection at a fine scale after NCRF inhibition (Fig. 3.2, bottom-right). If stability over at least 20 scales is applied, many keypoints will disappear but the most important ones will remain (Fig. 3.8, top-center and -right). In the case of the face image, important keypoints at eyes, nose, mouth and contour remain, even those at the marks on the forehead and cheekbone; compare Fig. 3.8 (top-right) with Fig. 3.2 (top-right). Also shown in Fig. 3.8 (bottom) are keypoints obtained with the SUSAN algorithm [Smith and Brady, 1997], the state-of-the-art, though limited to fine scale, in computer vision. Comparing the SUSAN results with ours, we may conclude that advanced models of cortical processing can achieve similar, if not better, results.

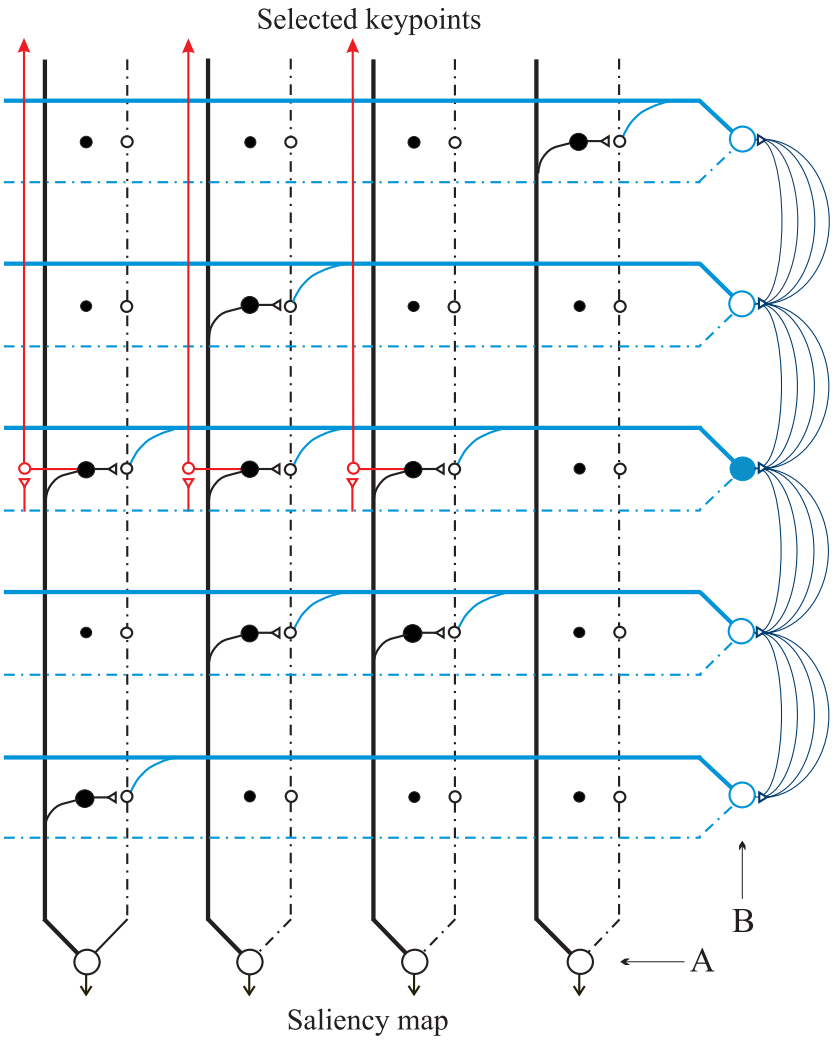

Figure 3.7: Schematic diagram for automatic scale selection, with horizontally the position and vertically the scale. Keypoint cells are represented by solid dots (active keypoint cells by big dots), grouping cells by big, open circles, and gating cells by small open circles. Dendrites are shown by solid lines and axons by dash-dotted lines. See text.

3.7 Focus-of-Attention by saliency maps

As mentioned above, the what and where subsystems are steered, top-down, on the basis of expected objects and positions in PF cortex. However, there is one complication that has not yet been mentioned: our eyes are constantly moving in order to suppress static projections of blood vessels etc. in our retinas. During a fixation, stable information propagates from the retinas via the LGN to V1, where first features are extracted, and then, also during the next saccade, to higher areas. Fixation points in regions where complex—and therefore important—information can be found are much more important than points in homogeneous regions. Focus-of-Attention, for guiding the where system in parallel with steering our eyes, is thought to be driven by an attention component in PF cortex because of overt attention: while strongly fixating our eyes at one point, we can direct mental attention to points in the neighborhood [Parkhurst et al., 2002]. For modelling FoA we need a map, called saliency map, which indicates the most important points to be analyzed (fixated). We propose a

Figure 3.8: Top: results of automatic scale selection, without scale stability (left) and with stability over 20 scales (center and right). Bottom: results obtained with the SUSAN algorithm [Smith and Brady, 1997].

simple scheme based on the multi-scale keypoint representation, because keypoints code local image complexity.

As done in the previous section, activities of all keypoint cells at position (x, y) are summed over scale s by grouping cells. These cells are the ones marked A in Fig. 3.7. At positions where keypoints are stable over many scales, this summation map will show distinct peaks at centers of objects, important sub-structures and contour landmarks. The height of the peaks provides information about their relative importance. In addition, such a summation map, with some simple processing of the projected trajectories of unstable keypoints, like low-pass filtering and non-maximum suppression, might also contribute to solving the segregation problem: the object center is linked to important structures, and these are linked to contour landmarks. Such a data stream is data-driven and bottom-up, and could be combined with top-down processing from inferior-temporal cortex in order to actively probe the presence of objects in the visual field [Deco and Rolls, 2004]. The summation map with links between the peaks might be available at higher cortical levels, where serial processing occurs for visual search, for example in the case when no object "pops out" and all objects must be screened sequentially.

Figure 3.9 (bottom; face196) shows keypoints at three different scales: (a) $\lambda = 4$, (b) $\lambda = 20$ and (c) $\lambda = 40$. We noticed that most if not all faces show a distinct keypoint on the middle of the line that connects the two eyes, like in Fig. 3.9b. Figure 3.9d shows the saliency map obtained on the basis of the entire scale space ($\lambda \in [4, 40]$) with 288 scales. Important peaks are found at the eyes, nose and mouth, but also at the hairline and even the chin and neck. The regions around the peaks were created by a very simple process: each keypoint has a Region-of-Interest (RoI) that can be used to process—during a fixation—other information inside the RoI, such as lines, edges, textures and disparity. The RoI is small at fine scales

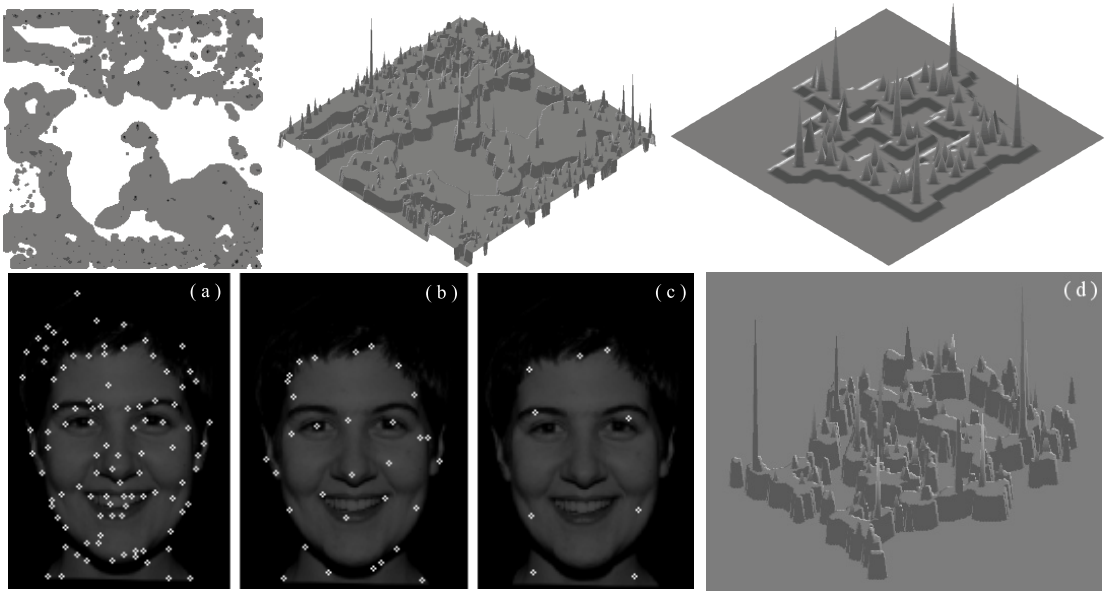

Figure 3.9: Top, left to right: saliency maps in 2D and 3D of the traffic sign and star object. Bottom: keypoints at fine, medium and coarse scales, plus saliency map.

and big at coarse scales. This is modelled by assuming circular axonal fields of keypoint cells, of size 3×3 at the finest scale ($\lambda = 4$) with linear scaling towards coarser scales. This means that the grouping cells marked A in Fig. 3.7 receive more input, and that the saliency map becomes a more diffuse "landscape" but still with high peaks. The maps shown in Fig. 3.9 have been thresholded, but this was only done for better displaying the structure of the maps, such that 3D projected views are not cluttered. The top row of Fig. 3.9 shows saliency maps in the case of the traffic-sign image (Fig. 3.2) and the star object (Fig. 3.4). In the former we can see the asymmetric region created by the keypoints at the bottom of and at the thin bar right to the arrow, in the latter the pentagonal structure of the star with peaks at the convex and concave vertices of the contour, in the triangles and in the center.

In Fig. 3.9d we can see the regions where important features are located, but it is quite difficult to see which peaks correspond to important facial landmarks. On the other hand, looking at Fig. 3.9b it is easy to see that some keypoints correspond to landmarks that we pretend to find in the next section, in this study limited to eyes, nose and mouth, but there are many more keypoints and at other scales (Fig. 3.9c) they are detected at other structures. Presumably, the visual system can use one "global" saliency map in combination with "partial" ones obtained by summing keypoints over smaller scale intervals, or even keypoints at individual scales, in order to optimize detection. This process can be steered by higher brain areas, which may contain prototype object maps with expected patterns, with approximate distances of eyes, nose and mouth. This can be part of the fast "where" data stream. Actual steering may consist of excitation and inhibition by pre-wired connections in keypoint scale space. This can be modelled by assuming grouping and gating cells which combine keypoint cells in approximate areas and at certain scales.

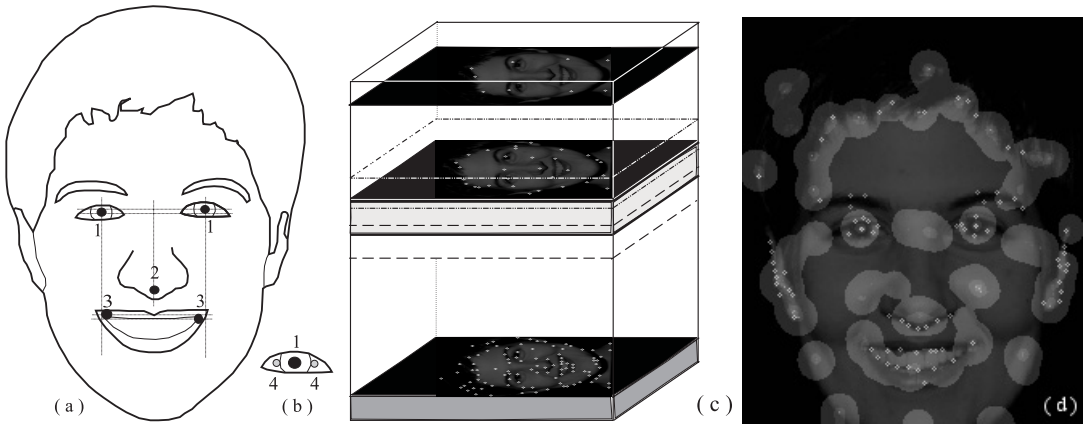

Figure 3.10: Left to right: (a) facial landmarks, (b) eye landmarks, (c) impression of keypoint scale space, and (d) partial saliency map at fine scales ($\lambda \in [4, 9]$) after NCRF inhibition.

3.8 Application: face detection

In our simulations we explored one possible scenario, see also Rodrigues and du Buf [2005c]. We assume the existence of very few layers of grouping cells, with dendritic fields in partial saliency maps that combine keypoints in specific scale intervals. The top layer with "face" cells groups axons of "eyes" (plural!), "nose" and "mouth" grouping cells. The "eyes" cells group axons of pairs of "eye" cells. Only the "eye," "nose" and "mouth" cells connect to the saliency maps, the "face" and "eyes" cells do not. This scenario consists of detecting possible positions of eyes, linking two eyes, then two eyes plus nose, and finally two eyes plus nose plus mouth. This is done dynamically by activating synaptic connections in the partial saliency maps, going from coarse to fine scales. We note that we did not yet include characteristic keypoints at other positions, like the one on the middle of the line that connects the two eyes (Fig. 3.9b).

We experimented with 30 faces—with different sizes and expressions—of the Stirling set (Fig. 3.13), and we used 7 partial saliency maps, each covering 40 scales distributed over $\Delta\lambda = 5$, but the scale intervals were overlapping 20 scales. The finest scale was at $\lambda = 4$. Examples of partial saliency maps are shown in Figs 3.10d and 3.11 (left). The search process starts at the coarsest scale interval, because there are much fewer candidate eye positions than there are at the finest scale interval, especially when a face is seen against a complex background. This is simulated by a feedback loop that activates connections to finer scale intervals, until at least one eye candidate is detected.

First, "eye" cells respond to significant peaks (non-maximum suppression and thresholding) in the selected saliency map. In the case of "face196" this was the map at $\lambda \in [13, 18]$, see Fig. 3.11 (left). This saliency map was the first one selected, because a peak at the center of an eye, as indicated by Fig. 3.10b-1, may only be accepted if there also are two stable symmetric keypoints (eye corners) at the 40 finest scales ($\lambda \in [4, 9]$), see Fig. 3.10b-4 and Fig. 3.10d. In order to reduce false positives, the latter is done after NCRF inhibition. If no single eye cell responds, the scale interval of the saliency map is not appropriate and the feedback loop will step through all saliency maps (Fig. 3.10c), until at least one eye cell responds.

Second, an "eyes" cell responds if two "eye" cells are active on an approximately horizontal line (Fig. 3.10a-1). An "eyes" cell is a grouping cell with two, symmetric, dendritic subfields.

Figure 3.11: Left: partial saliency map of face196 ($\lambda \in [13, 18]$). Right: keypoints used by eye, nose and mouth detection cells.

If no eye pair is detected, a new saliency map is selected (feedback loop).
Third, when two eyes can be grouped, a "nose" cell is activated, its dendritic field covering an area below the "eyes" and "eye" cells in the saliency map (Fig. 3.10a-2). If no peak is detected, a new saliency map is selected (feedback loop).
Fourth, if both "eyes" and "nose" cells respond, a "mouth" cell with two dendritic subfields at approximate positions of the two mouth corners (Fig. 3.10a-3) is activated. If keypoints are found, one "face" cell will be excitated. If not, a new saliency map is selected (feedback loop).

The process stops when one or no face has been detected, but in reality it might continue at finer scale intervals because there may be more faces with different sizes in the visual field (image). The result obtained in the case of "face196" is shown in Fig. 3.11, where +, □ and × symbols indicate detected and used keypoints at eyes, nose and mouth corners (actual positions of face and eyes cells are less important). More results are shown in Fig. 3.13. Of all 30 face images that we tested, one was problematic because of a very extreme expression, such that keypoints at mouth corners could not be grouped. In two cases, only the central part of a face was within the image border, which hampers detection of keypoints at eyes at coarse scales because of large filter sizes. In many applications such problems can be avoided. Nevertheless, a detection rate of 90% is encouraging in view of the extreme simplicity of the method, and compares well to other methods, which can be very complex and which must deal with the same problems [Yang et al., 2002].

Figure 3.13 also shows a correctly-detected face that was constructed by combining different fruits. The reason that this "fake" face was detected is that positions of facial landmarks as used in our model correspond to positions in real faces—our own visual impression, at first, also tells that it is a face. This effect is exploited in cartoons, even by the famous Italian painter Giuseppe Arcimboldo in the 16th century. Obviously, more features must be used, including the multi-scale line/edge representation in V1, but we must distinguish between face *detection* and *recognition*. Detection is thought to take place by means of keypoints and in the fast where system, after which additional features are available in the slower what system for recognition, including objects like fruits, in order to be able to distinguish between real and fake faces. In any case, we explored only one possible scenario in which grouping cells receive input at expected positions of eyes, nose and mouth. Such grouping

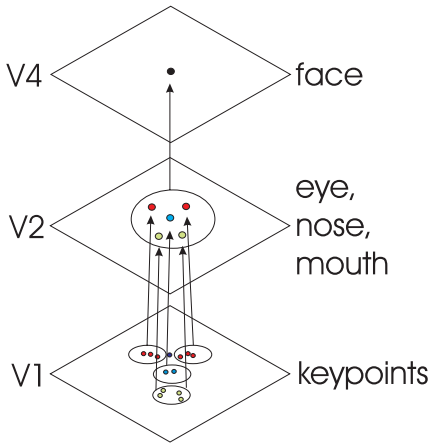

Figure 3.12: Instead of grouping keypoints at facial landmarks in area V1, such groupings may actually be done in higher areas V2 and V4.

cells might be located in V1, but also in V2 and V4, see Fig. 3.12 and the Deco and Rolls [2004] multi-area cortical architecture. As a consequence, only one "face cell" in V4 may be translation invariant, and therefore it may have a very large receptive field at the lowest (input) level.

Figure 3.14 shows the result of applying our coarse-to-fine-scale scenario to an image with a complex background. This background leads to a huge number of keypoints, especially at the finest scales (Fig. 3.14 center), with the possibility that random and unrelated keypoints can excitate "eye" and even "eyes" cells etc. However, this did not occur because of the coarse-to-fine strategy, in which a peak in the saliency map at a coarse scale (center pupil) must be grouped with two keypoints at the finest scales (eye corners). The additional groupings of keypoints at nose and mouth corners, at the coarser scales, increase selectivity. However, the result shown in Fig. 3.14 concerns a first experiment to test the detection scenario. Many more tests are required in order to validate and/or improve the method, including the detection of multiple faces—with different positions and sizes, eventually with partial occlusions—in images. This, and faces with different pose (frontal, 3/4 view, profile), requires the use of various templates in memory to steer the detection by activating different grouping cells with different spatial relations at different scales.

3.9 Discussion

As Rensink [2000] pointed out, the detailed and rich impression of our visual surround may not be caused by a rich representation in our visual memory, because the stable, physical surround already "acts" like memory. In addition, focussed attention is likely to deal with only one object at a time. His triadic architecture therefore separates focussed attention to coherent objects (System II) from non-attentional scene interpretation (Layout and Gist subsystems in System III), but both systems are fed by low-level feature detectors in System I.

In this chapter we showed that keypoints detected on the basis of end-stopped operators, and in particular a few partial saliency maps that cover overlapping scale intervals, provide very important information for object detection. Exploring a very simple processing scheme, faces can be detected by grouping together axons of keypoint cells at approximate retino-

Figure 3.13: Results obtained with different faces and expressions.

topic positions, and this leads to robust detection in the case of different facial expressions. However, the simple scheme explored only works if the eyes are open, if the view is frontal, and if the faces are approximately vertical. For pose-, rotation- and occlusion-invariant detection, the scheme must be fed by Rensink's short-term Layout and Gist subsystems, but also the long-term Scene Schema system that is supposed to build and store collections of object representations, for example of non-frontal faces.

We also showed that keypoints may play an important role in other cortical processes. A global saliency map provides ideal information for Focus-of-Attention, because distinct peaks are found at structures with a high complexity. This global saliency map can also be used for automatic scale selection, such that stable keypoints which are most characteristic for an object can be prepared for a first—but very fast—categorization. Furthermore, it was shown that linking keypoints from coarse to fine scales can contribute to object segregation.

We focussed on the keypoint scale space in this chapter. However, keypoint detection can be complemented with multi-scale line and edge detection, which is also supposed to occur in V1. It has already been shown that object segregation and categorization—for

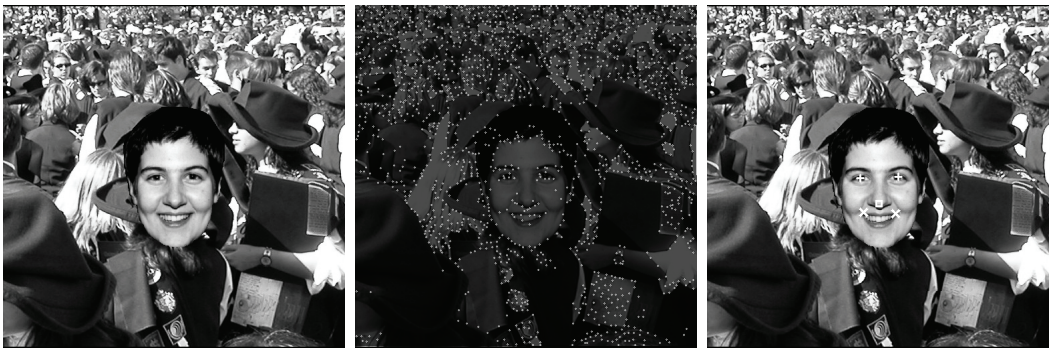

Figure 3.14: Result with a complex background, which yields a huge number of keypoints especially at fine scales (center).

example for distinguishing dogs, horses and cows—can also be achieved by only considering the line/edge scale space [Rodrigues and du Buf, 2006a]. This implies that the combination of detected keypoints and detected lines and edges will lead to improved performance, also enabling face *recognition*, but *how* all information can be combined in the best way remains an open question.

Owing to the impressive performance of current computers, it is now possible to test Rensink's triadic model [Rensink, 2000] in terms of Deco and Rolls' cortical architecture [Deco and Rolls, 2004]. The ventral what data stream (V1, V2, V4, IT) is supposed to be involved in object recognition, independently of position and scaling. The dorsal where stream (V1, V2, MT, PP) is responsible for maintaining a spatial map of an object's location, the spatial relationship of an object's parts, as well as moving the spatial allocation of attention. Both data streams are bottom-up and top-down. Apart from input via V1, both streams receive top-down input from a postulated short-term memory for shape features or templates in prefrontal cortical area 46, i.e., the more ventral area PF46v generates an object-based attentional component, whereas the more dorsal area PF46d specifies the location. As for now, we do not know *how* PF46 works. It might be the neurophysiological equivalent of the cognitive Scene Schema system mentioned above, but apparently the what and where data streams are necessary for obtaining view-independent object detection through cells with receptive fields of 50 degrees or more [Deco and Rolls, 2004]. However, instead of receiving input directly from simple cells, the data streams should receive input from feature extraction engines in V1 and beyond, including keypoint cells!

Chapter 4

Multi-scale lines and edges in V1 and beyond

Abstract: In this chapter we present an improved scheme for line and edge detection in cortical area V1. This scheme is based on responses of simple and complex cells, and it is multi-scale with no free parameters. We illustrate the multi-scale line/edge representation in automatic scale selection, in visual reconstruction, and we show how object segregation can be achieved with coarse-to-fine-scale groupings. A two-level object categorization scenario is tested in which pre-categorization is based on coarse scales only and final categorization on coarse plus fine scales. We also present a multi-scale object and face recognition model. It is shown that a new disparity model based on the multi-scale line and edge coding can be used to directly attribute depth information to detected lines and edges. Processing schemes are discussed in the framework of a complete cortical architecture.

4.1 Introduction

The visual cortex detects and recognizes objects by means of the "what" and "where" subsystems. The "bandwidth" of these systems is limited: only one object can be attended at any time [Rensink, 2000]. In a current model by Deco and Rolls [2004], the ventral what system receives input from area V1 which proceeds through V2 and V4 to IT (inferior temporal cortex). The dorsal where system connects V1 and V2 through MT (medial temporal) to area PP (posterior parietal). Both systems are controlled, top-down, by attention and short-term memory with object representations in PF (prefrontal) cortex, i.e., a what component from PF46v to IT and a where component from PF46d to PP. The bottom-up (visual input code) and top-down (expected object and position) data streams are necessary for obtaining size, rotation and translation invariance.

Signal propagation from the retinas through the LGN (lateral geniculate nucleus) and areas V1, V2 etc., including feature extractions in V1 and groupings in higher areas, takes time. Object recognition is achieved in 150–200 msec, but category-specific activation of PF cortex starts after about 100 ms [Bar, 2004]. In addition, IT cortex first receives coarse-scale

information and later fine-scale information. Apparently, one very brief glance is sufficient for the system to develop a gist of the contents from an image [Oliva and Torralba, 2006]. This implies that some information propagates very rapidly and directly to "attention" in PF cortex in order to pre-select possible object templates and positions that then propagate down the what and where systems. This process we call object categorization, which cannot be obtained by the CBF model by Riesenhuber and Poggio [2000a] because categorization (e.g. a cat) is obtained by grouping outputs of identification cells (cat1, cat2, cat3). In other words, categorization would be obtained *after* recognition. In contrast, the LF (Low Frequency) model [Oliva et al., 2003; Bar, 2004] assumes that categorization is obtained *before* recognition: low-frequency information that passes directly from V1/V2 to PF cortex, although the LF information actually proposed consists of lowpass-filtered images but not of e.g. outputs of simple and complex cells in V1 tuned to low spatial frequencies. The latter option will be explored in this chapter.

After object categorization on the basis of coarse-scale information has narrowed the set of objects to be tested, the recognition process can start by applying also fine-scale information. We will focus on how such processes can be embedded in the architecture referred to above, with special focus on face recognition. Despite the impressive number and variety of computer-vision methods devised for faces and facial landmarks, see e.g. Yang et al. [2002], we show that very promising results with a cortical model can be obtained, even in the case of some classical complications involving changes of pose (frontal and 3/4), facial expression, some lighting and noise conditions, and the wearing of spectacles.

There exists a vast literature, from basic feature extraction to object segregation, categorization and recognition, and from image reconstruction, scale stabilization to stereo, in computer vision, but much less in biological vision. We therefore continue with a very brief summary of approaches related to this chapter, with spatial focus on the biological methods.

In addition to a few general overviews, see e.g. [Hubel, 1995; Bruce et al., 2000; Rasche, 2005; Miikkulainen et al., 2005], there also are detailed and quantitative models of simple, complex, end-stopped, bar and grating cells [Heitger et al., 1992; Petkov and Kruizinga, 1997], plus various models for inhibitions [Heitger et al., 1992; Petkov et al., 1993b; Barth et al., 1998; Rodrigues and du Buf, 2006a], edge detection [Smith and Brady, 1997; Elder and Zucker, 1998; Kovesi., 1999; Grigorescu et al., 2003], combined line and edge detection [Verbeek and van Vliet, 1992; van Deemter and du Buf, 2000; Rodrigues and du Buf, 2004b, 2006a], and keypoint detection [Würtz and Lourens, 2000; Hansen et al., 2001; Lowe, 2004; Rodrigues and du Buf, 2005b]. Other models address saliency maps and Focus-of-Attention [Itti and Koch, 2001; Parkhurst et al., 2002; Deco and Rolls, 2004; Rodrigues and du Buf, 2006d], figure-ground segregation [Heitger and von der Heydt, 1993; Hupe et al., 2001; Zhaoping, 2003; Rodrigues and du Buf, 2006a], and object categorization [Riesenhuber and Poggio, 2000a; Leibe and Schiele, 2003; Csurka et al., 2004; Rodrigues and du Buf, 2006a]. Concerning faces, various approaches have been proposed, from detecting faces and facial landmarks to the influence of different factors such as race and age [Delorme and Thorpe, 2001; Yang et al., 2002; Ban et al., 2003; Rodrigues and du Buf, 2005c], including final face recognition [Kruizinga and Petkov, 1995; Zhao et al., 2003; Rodrigues and du Buf, 2006c,d]. Yet other models have been devised for disparity [Fleet et al., 1991; Ohzawa et al., 1997; Qian, 1997; Rodrigues and du Buf, 2004b], automatic scale selection [Lindeberg, 1994], visual reconstruction [Rodrigues and du Buf, 2006b], brightness perception [du Buf, 2001] and visual pattern detection at very low contrast [du Buf, 2005].

In this chapter we show that one basic process, namely line and edge detection in V1 (and V2), can be linked to most if not all the topics mentioned above. We present an improved

scheme for multi-scale line/edge (MS-L/E) extraction in V1, which is truly multi-scale with no free parameters. We illustrate the MS-L/E interpretation (coding and representation) for automatic scale selection and explore the importance of this interpretation in object reconstruction, segregation, categorization and recognition. Since experiments with possible Low-Frequency models based on lowpass-filtered images, following Bar [2004], gave rather disappointing results, which is due to smeared blobs of objects that lack any structure, we propose that categorization is based on coarse-scale L/E coding, and that recognition involves all scales. Processing schemes are discussed in the framework of a complete cortical architecture. We emphasize that the multi-scale keypoint information also extracted in V1, which was shown to be very important for facial landmark detection [Rodrigues and du Buf, 2006d] (see Section 3.8), and other important features such as texture and continuity information that can be retrieved from bar and grating cells [du Buf, 2007], will not be employed here because we want to focus completely on the multi-scale line/edge information in V1 and beyond. Therefore, this chapter complements the previous one dedicated to keypoints, and the following chapter is about how the multi-scale line/edge and keypoint representations can be integrated.

In Section 4.2 we present line/edge detection and classification in single- and multi-scale contexts, plus the application of NCRF inhibition. Section 4.3 illustrates the visual reconstruction model. Section 4.4 deals with object segregation, Section 4.5 with automatic scale selection, and Section 4.6 with object categorization. This is followed by face recognition in Section 4.7 and disparity in Section 4.8. We conclude with the discussion in Section 4.9.

4.2 Line and edge detection and classification

In the previous chapters it was explained that Gabor quadrature filters provide a good model of receptive fields (RFs) of cortical simple cells. In the spatial domain (x, y) they consist of a real cosine and an imaginary sine, both with a Gaussian envelope, see Section 3.2 or [Lee, 1996; Grigorescu et al., 2003; Rodrigues and du Buf, 2006d]. As in Chapter 3, an RF is given by Eq. 3.1, where $1/\lambda$ is the spatial frequency, λ being the wavelength. Here we apply exactly the same parameter values. For the bandwidth σ/λ we use 0.56, which yields a half-response width of one octave. The angle θ determines the orientation (we use 8 orientations), and φ the phase symmetry (0 or $-\pi/2$). We apply filters with an aspect ratio of 0.5. Below, the scale s of analysis will be given in terms of λ expressed in pixels, where $\lambda = 1$ corresponds to 1 pixel. Most images shown in this paper have a size of 256×256 pixels. We can apply a linear scaling between $f_{\min}$ and $f_{\max}$ with either a few discrete scales or with hundreds of almost contiguous scales. Responses of even and odd simple cells, which correspond to the real and imaginary parts of a Gabor filter, are denoted by $R^E_{s,i}(x, y)$ and $R^O_{s,i}(x, y)$, s being the scale, i the orientation ($\theta_i = i\pi/(N_\theta - 1)$) and N_θ the number of orientations (we use $N_\theta = 8$). Responses of complex cells are modeled by the modulus following Eq. 3.2.

Figure 4.1 shows responses of even and odd simple cells and complex cells in the dominant orientation (i.e., the orientation with the maximum response of the eight oriented complex cells) at three scales in the case of the Fiona image shown in Fig. 4.4 (top-left). This information is used for the extraction of lines and edges.

A basic scheme for single-scale line and edge detection based on responses of simple cells works as follows [van Deemter and du Buf, 2000]: a positive (negative) line is detected where R^E shows a local maximum (minimum) and R^O shows a zero crossing. In the case of edges the even and odd responses are swapped. This gives 4 possibilities for positive and negative

Figure 4.1: Responses of even simple cells (top), odd simple cells (middle) and complex cells (bottom) at three scales ($\lambda = 4, 8, 16$).

events. An improved scheme [Rodrigues and du Buf, 2004b] consists of combining responses of simple and complex cells, i.e., simple cells serve to detect positions and event types, whereas complex cells are used to increase the confidence. Since the use of Gabor modulus (complex cells) implies a loss of precision at vertices [du Buf, 1993], increased precision was obtained by considering multiple scales (i.e., a few neighboring micro-scales).

The algorithms described above work reasonably well but there remain a few problems: (a) either one scale is used or only a very few scales for increasing confidence, (b) some parameters must be optimized for specific input images or even as a function of scale, (c) detection precision can still be improved, and (d) detection continuity at curved lines/edges must be guaranteed. Therefore we present an improved algorithm with no free parameters, truly multi-scale and with new solutions for problems (c) and (d).

With respect to precision, simple and complex cells respond beyond line and edge terminations, for example beyond the corners of a rectangle. In addition, at line or edge crossings, detection leads to continuity of the dominant events but to gaps in the sub-dominant events. These gaps must be reduced in order to reconstruct continuity. Both problems can be solved by introducing new inhibition schemes, like the radial and tangential ones used in the case of keypoint operators [Rodrigues and du Buf, 2004b]; see Section 3.2. Here we use lateral

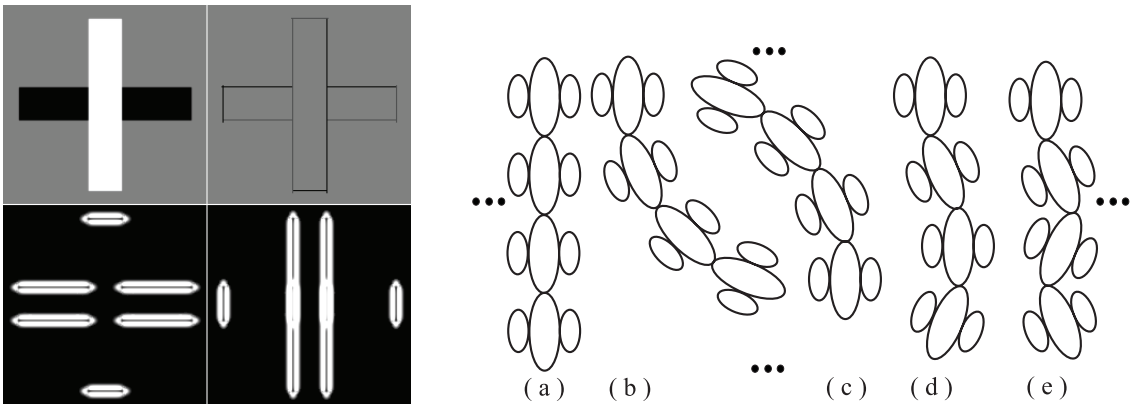

Figure 4.2: Left: input pattern (a cross, at top-left), the summation of lateral and cross-orientation inhibition for $\theta = \{0, \pi/2\}$ (bottom) and the detection result (top-right) with no spurious events beyond the corners and no gaps at the junctions. Right: curvature continuity.

(L) and cross-orientation (C) inhibition, defined as

$$I^L_{s,i}(x,y) = [C_{s,i}(x + d\mathcal{C}_{s,i}, y + d\mathcal{S}_{s,i}) - C_{s,i}(x - d\mathcal{C}_{s,i}, y - d\mathcal{S}_{s,i})]^+ + [C_{s,i}(x - d\mathcal{C}_{s,i}, y - d\mathcal{S}_{s,i}) - C_{s,i}(x + d\mathcal{C}_{s,i}, y + d\mathcal{S}_{s,i})]^+ ; \quad (4.1)$$

$$I^C_{s,i}(x,y) = \left[C_{s,(i+N_\theta/2)}(x + 2d\mathcal{C}_{s,i}, y + 2d\mathcal{S}_{s,i}) - 2.C_{s,i}(x,y) + C_{s,(i+N_\theta/2)}(x - 2d\mathcal{C}_{s,i}, y - 2d\mathcal{S}_{s,i})\right]^+ , \quad (4.2)$$

where $(i + N_\theta/2) \perp i$, with $\mathcal{C}_{s,i} = \cos\theta_i$, $\mathcal{S}_{s,i} = \sin\theta_i$, and $d = 0.6s$. Inhibition is applied to the responses of complex cells, where β controls the strength of the inhibition (we use $\beta = 1.0$), i.e. $\hat{C}_{s,i} = \left[C_{s,i}(x,y) - \beta(I^L_{s,i}(x,y) + I^C_{s,i}(x,y))\right]$.

Figure 4.2 (at left) shows a cross formed by two bars (top-left), the summation of L and C inhibition for $\theta = \{0, \pi/2\}$ (bottom) and the detection result (top-right) with no spurious events beyond the corners and with no gaps at the junctions.

Line and edge detection is achieved by constructing a few cell layers on top of simple and complex cells; see Fig. 4.3 for a wiring diagram. The first layer serves to select active regions and dominant orientations. At each position, responses of complex cells are summed ($\hat{C}_s = \sum_{i=0}^{N_\theta - 1} \hat{C}_{s,i}$), and at positions where $\hat{C}_s > 0$ an output cell is activated. At active output cells, the dominant orientation is selected by gating one complex cell on the basis of non-maximum suppression of $\hat{C}_{s,i}$. The gating is confirmed or corrected by an excitation and inhibition process of dominant orientations in a local neighborhood.

In the second layer, event type and position are determined on the basis of active output cells (1st layer) and gated simple and complex cells. A first cell complex checks responses of simple cells $R^E_{s,i}$ and $R^O_{s,i}$ for a local maximum (or minimum by rectification) using a dendritic field size of $\pm\lambda/4$, λ being the wavelength of the simple cells (Gabor filter). The active output cell is inhibited if there is no maximum or minimum. A second cell complex does exactly the same on the basis of responses of complex cells. A third cell complex gates four types of zero-crossing cells on the basis of simple cells, again on $\pm\lambda/4$. If there is no zero-crossing, the output cell is inhibited. If there is a zero-crossing, the active output cell at the position of the zero-crossing cell determines event position and the active zero-crossing cell determines event type.

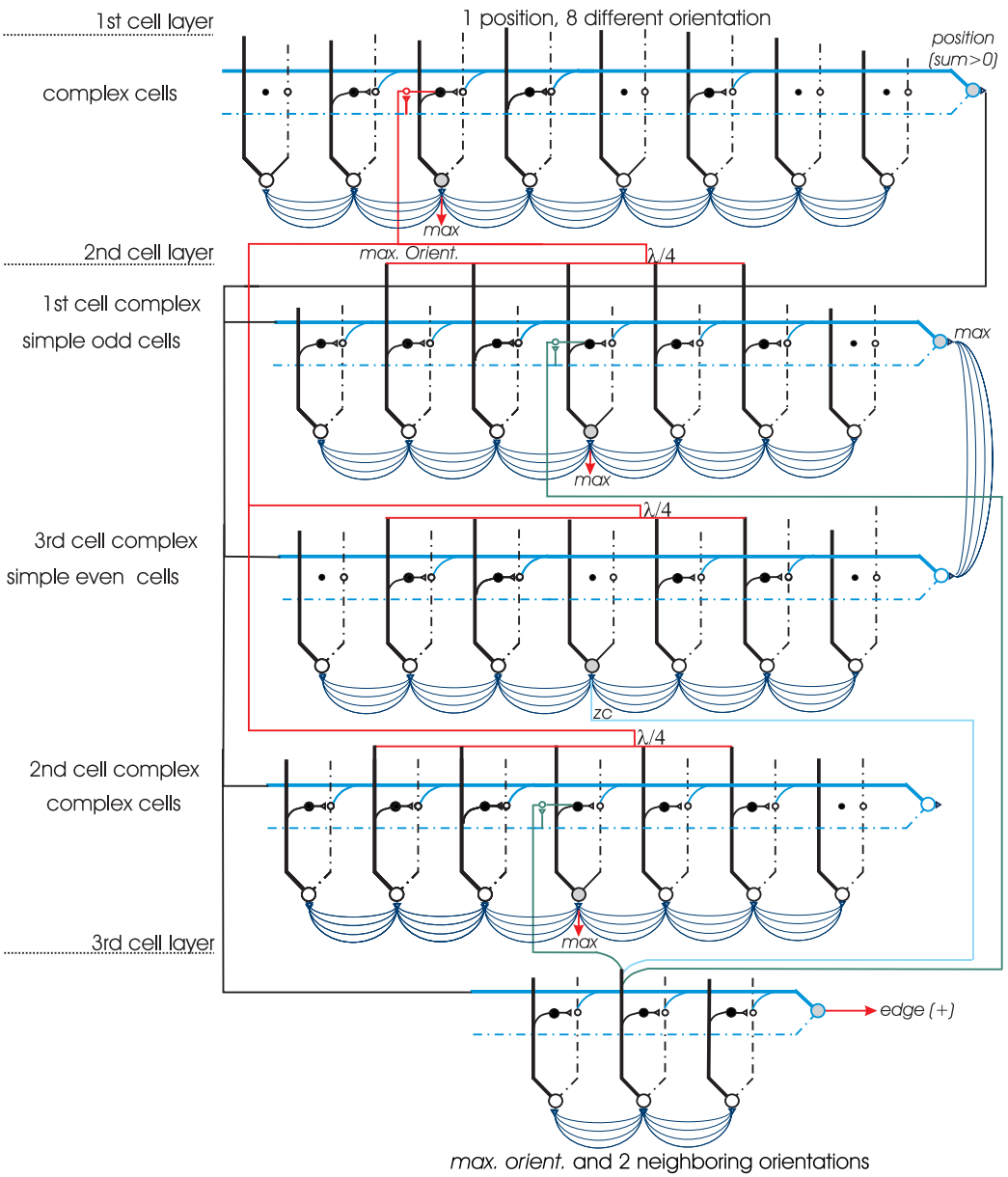

Figure 4.3: Schematic diagram for line/edge detection (single event and single scale) using 8 orientations. Cells are represented by solid dots (active cells by big dots), grouping cells by big open circles, and gating cells by small open circles. Dendrites are shown by solid lines and axons by dash-dotted lines (see text).

In the third layer, the small loss of accuracy due to the use of responses of complex cells in the second layer is compensated. This is done by correcting local event continuity, considering the information available in the second layer, but by using excitation of output cells by means of grouping cells that combine simple and complex cells tuned to the same and two neighboring orientations, see Fig. 4.2 (right). The latter process is an extension of linear grouping [van Deemter and du Buf, 2000] and a simplification of using banana wavelets [Krüger and Peters, 1997]. In the same layer, also event type is corrected in small neighborhoods, restoring type continuity since cell responses may be distorted by interference effects when two events are very close [du Buf, 1993].

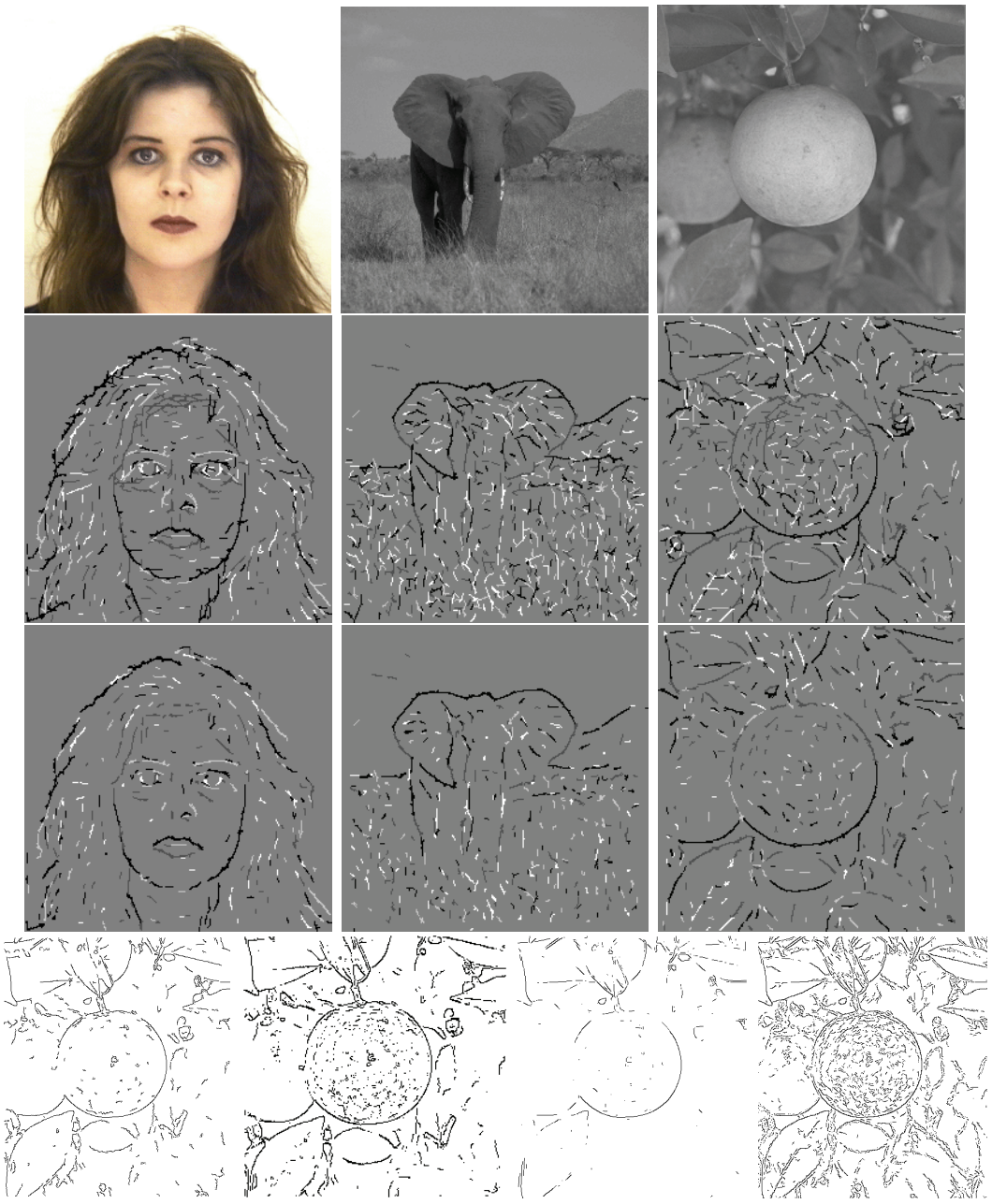

Figure 4.4: Line and edge detection at the finest scale. The second row shows positive and negative lines and edges, coded by gray level, without applying NCRF inhibition. The third row shows the same with NCRF inhibition. The bottom row shows edges detected by the (from left to right) Bergholm, Canny, Iverson and Nalwa algorithms in the case of the orange image.

The second row in Fig. 4.4 shows detection results with positive and negative lines and edges coded by different levels of gray (white, light gray, dark gray and black, respectively). Detection accuracy is very good and there remain many small events due to low-contrast textures and the fact that no threshold value has been applied (event amplitudes, for example the responses of complex cells at positions where events were detected, are not shown in Fig. 4.4).

In the previous chapter (see Section 3.2) it was shown that non-classical receptive field

(NCRF) inhibition can be used to suppress information in textured regions [Grigorescu et al., 2003; Rodrigues and du Buf, 2005a]. Instead of applying such inhibition only to keypoint detection at fine scales, it can also be applied to line and edge detection. The third row in Fig. 4.4 shows detection results with (I-)NCRF inhibition applied to the responses of the complex cells (above a small threshold $b^I_{s,\sigma}$). As a result, many small events in the face and hair (Fiona), ears and grass (elephant), and orange and tree have been suppressed and the most important events remain. For comparing our results obtained with NCRF inhibition in the case of the elephant image we refer to Grigorescu et al. [2003], but we note that they developed contour (edge) detection algorithms, whereas we can distinguish between edges and lines with different polarities, which is necessary for visual reconstruction; see below. The bottom row in Fig. 4.4 allows to compare our results (orange image) with state-of-the-art (but edge only) algorithms in computer vision, i.e. Bergholm, Canny, Iverson and Nalwa, see Heath et al. [2000] and also *http://marathon.csee.usf/edge/edge_detection.html*.

4.2.1 Multiple scales

We now focus on the multi-scale line/edge representation. Although NCRF inhibition can be applied at each scale, we will not do this for two reasons: (a) we want to illustrate line and edge behavior in scale space for applications like categorization, recognition and visual reconstruction, and (b) in many cases a coarser scale, i.e., increased RF size, will automatically eliminate texture detail. For illustrating scale space we can create an almost continuous, linear scaling with hundreds of scales between $\lambda \in [4, 52]$, but here we will present only a few scales in order to show a few properties and complications.

The top two rows in Fig. 4.5 show events detected at five scales in the case of ideal, solid square and star objects. At fine scales (to the left) the edges of the square are detected, as are most parts of the star, but not at the very tips of the star. This illustrates an important difference between normal computer vision and developing cortical models. The latter must be able to construct brightness maps (see Chapter 6), and at the tips of the star, where two edges converge, there are very fine lines. The same effect occurs at coarser scales, until entire triangles are detected as lines and even ten pairs of opposite triangles (to the right). In the case of the square, lines will be detected at diagonals, which vanish, with small lengths and amplitudes, at very coarse scales. The third row in Fig. 4.5 shows a mug, one of the objects that will be used in object categorization, and the bottom two rows show the Fiona and Kirsty images, two of the images that will be used in face recognition. This figure shows that detail disappears at coarser scales; there the result is more "sketchy" and abstract, a generalization property that will be exploited in object categorization.

Figure 4.6 illustrates the concept of stabilization over multiple scales, which will be applied in the object recognition model, applying different criteria for scale stability. In the case of the leaf image, from top to fourth row: single-scale detection without stability criterium, micro-scale stability over a few neighboring scales [Rodrigues and du Buf, 2004b], and stability over 10 and 40 scales ($\Delta\lambda = 5$). Bottom row: Kirsty face image with stabilization over 10 scales. This figure shows that many important detected events are rather stable over many scales, which is very important for tasks like visual reconstruction and object recognition.

Figure 4.5: Top two rows: multi-scale line/edge representations of a square and a star at, from left to right, $\lambda = \{4, 12, 18, 24, 40\}$. Bottom three rows: a mug and two faces at $\lambda = \{4, 8, 12, 24, 28\}$.

4.3 Visual reconstruction

Image reconstruction can be obtained by assuming one lowpass filter plus a complete set of (Gabor) bandpass filters that cover the entire frequency domain—this concept is exploited in image coding. The goal of our visual system is to detect objects, with no need, nor capacity, to reconstruct a complete image of our visual environment, see change blindness and the limited "bandwidth" of the what and where subsystems [Rensink, 2000]. Yet, the image that we perceive in terms of brightness must somehow be created. A normal image coding scheme, for example by summing responses of simple cells, requires accumulation in one cell layer which contains a brightness map, but this would require "yet another observer" of this map in our brain. A solution to this dilemma is to assume that detected lines and edges are interpreted symbolically: an active "line cell" is interpreted as having a Gaussian intensity profile with a certain orientation, amplitude and scale, the size of the profile being coupled

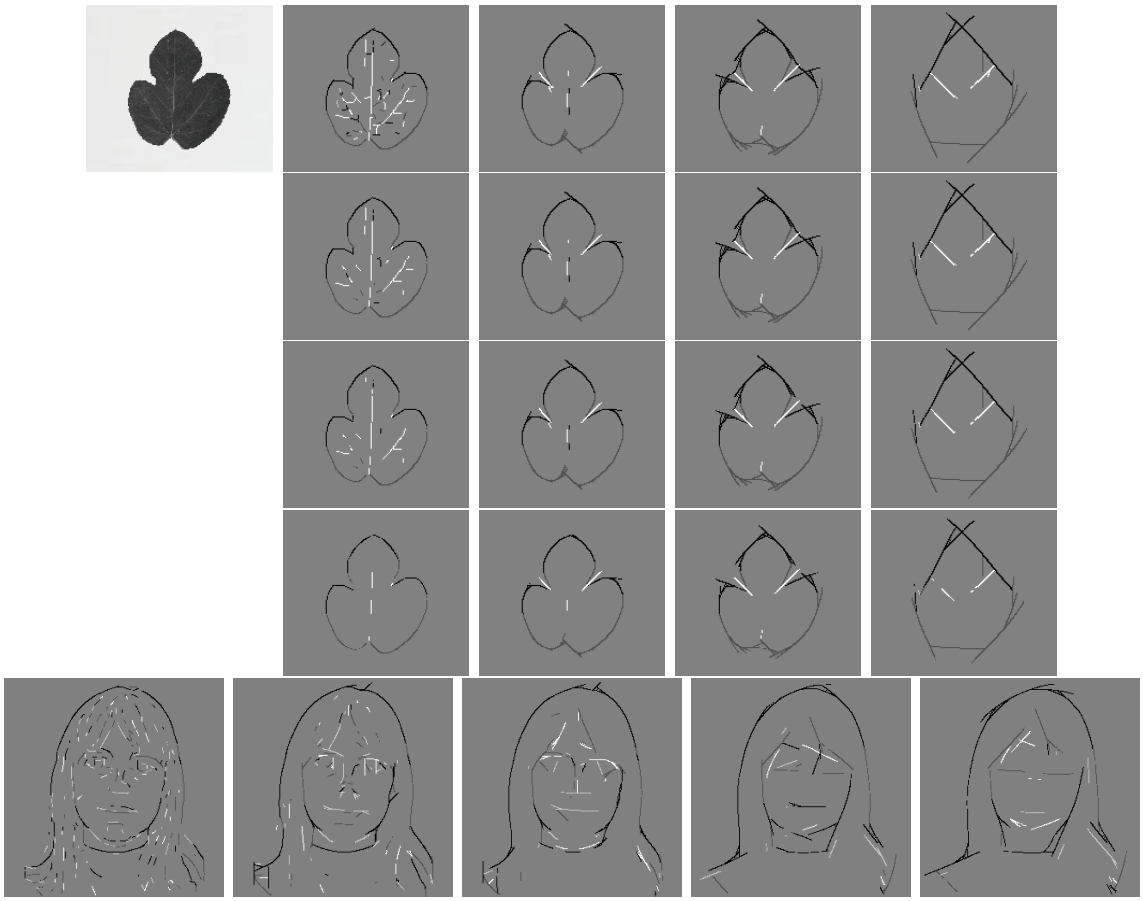

Figure 4.6: The top four rows show, left to right, input image (tree leaf) and multi-scale event detection at $\lambda = \{4, 9, 16, 36\}$. Top to bottom: single-scale detection, micro-scale stability, and stability over 10 and 40 scales. The bottom row shows results with stabilization over 10 scales at $\lambda = \{4, 8, 12, 24, 28\}$ in the case of the Kirsty image.

to the scale of the underlying simple and complex cells. An active "edge cell" is interpreted the same way, but with a bipolar, Gaussian-truncated, errorfunction profile; for details and illustrations see Chapter 6. As for image coding, this representation must be complemented with a lowpass filter, a process that happens to exist by means of retinal ganglion cells with photoreceptive dendritic fields *not* (in)directly connected to rods and cones, the main photoreceptors [Berson, 2003].

One brightness model [du Buf, 1994; du Buf and Fischer, 1995] is based on the symbolic line and edge interpretation, it explains Mach bands [Pessoa, 1996b] by the fact that responses of simple cells do not allow to distinguish between lines and ramp edges, and it was shown to be able to predict many brightness illusions such as simultaneous brightness contrast and assimilation, which are two opposite induction effects (the model referred to above was only tested in 1D and has now been extended to 2D).

Here we will not go into more detail because a detailed explanation of the 2D extension is presented in Chapter 6. We only illustrate the symbolic (re)construction process in 2D that will be exploited in face recognition. We note that we write (re)construction because there is no simple and straightforward reconstruction. The left part of Fig. 4.7 shows, top to bottom, symbolic interpretations of positive and negative edges and lines at fine (left) and

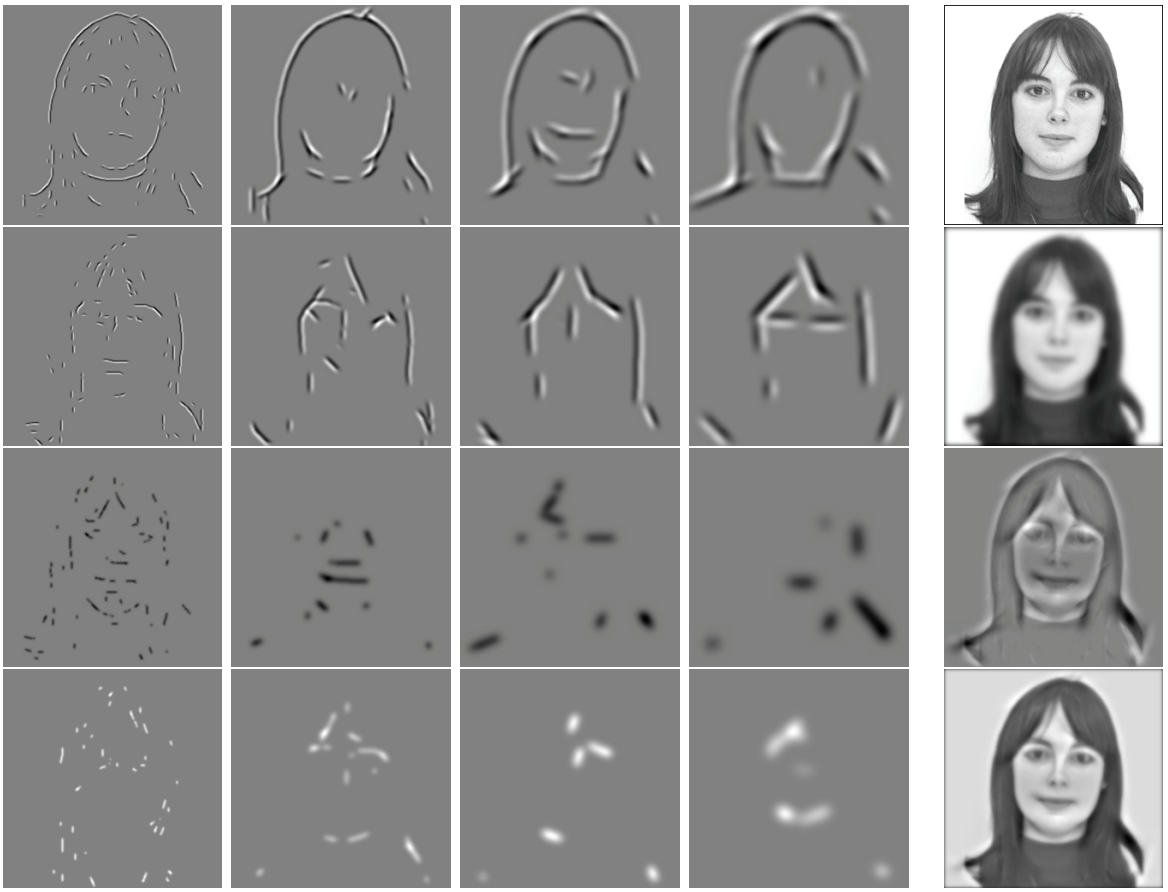

Figure 4.7: Left part: multi-scale symbolic line and edge interpretations with, top to bottom, negative and positive edges and lines. Rightmost column: reconstruction of the Kirsty image with, top to bottom: input image, lowpass-filtered image, summation of symbolic line and edge interpretations shown in the left part, and the final reconstruction.

coarse (right) scales. The rightmost column illustrates visual (re)construction of the Kirsty image, from top to bottom: input image, lowpass-filtered image (LP_σ), the summation of symbolic line (L_s) and edge (E_s) interpretations (the sum of all images in the left part), and the final reconstruction (R), i.e.

$$R = \gamma \cdot LP_\sigma + (1 - \gamma) \cdot \sum_{s=1}^{N_s} [\frac{1}{N_s} . (L_s + E_s)], \tag{4.3}$$

with $\gamma = 0.5$. Obviously, the use of more than four scales leads to better (re)constructions, but the relative weighting of the lowpass and all the scale components is still under investigation. In principle one can use the same number of scales as used later in the object categorization and recognition processes, for example $N_s = 8$ scales. Summarizing, the multi-scale line and edge interpretation allows to (re)construct the input image, and this representation will be used for e.g. face recognition but it is also the basic concept of the new 2D brightness model.

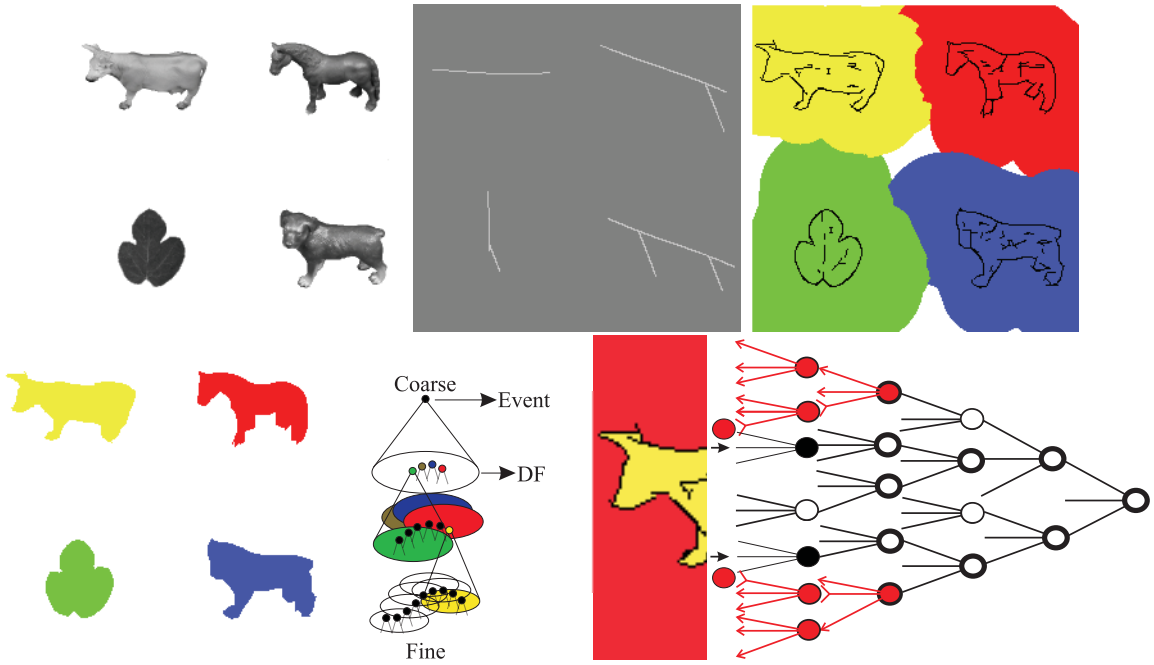

Figure 4.8: Object segregation. Top row, left to right: input image with four objects, the representation at $\lambda = 40$, and regions-of-influence with I marking the interior. Bottom row, left to right: result of figure-ground segregation, coarse-to-fine projection (DF denotes dendritic field), and activation and inhibition of grouping cells (right).

4.4 Object segregation

Until here we have illustrated multi-scale line and edge detection in area V1 and the symbolic interpretation for visual (re)construction in brightness perception, but one of the other goals of the visual cortex is to detect and recognize objects by means of the what and where systems. Rensink [2000] argued that these systems can attend only one object at any time. In the model by Deco and Rolls [2004] (see Introduction), the ventral what system receives input from V1 which proceeds through V2 and V4 to IT. The dorsal where system connects V1 and V2 through MT to area PP. Both systems are "controlled," top-down, by attention and short-term memory with object representations in PF cortex, i.e., a what component from PF46v to IT and a where component from PF46d to PP. The bottom-up (visual input code) and top-down (expected object and position) streams are necessary for obtaining size, rotation and position invariance, which means that object templates in memory may be normalized. Here we will not go into more detail, because our goal is not to (re)implement the model. Our goal is to show how the line and edge code can be used in the what and where systems, focusing on multi-scale processing.

Figures 4.5 and 4.6 show typical event maps of different objects, with detail at fine scales and more abstract, "sketchy" information at coarse scales. At a very coarse level, each individual event (group of responding line/edge cells) or connected group of events corresponds to one entire object, see Fig. 4.8 top-center. Each event at such a coarse scale is related to events at one finer scale, which can be slightly displaced or rotated, and this continuity continues to fine scales. This relation is modeled by downprojection using grouping cells with a dendritic field (Fig. 4.8 bottom-center, see also Section 3.5), the size of which defines the region-of-influence. Responding event cells at all scales activate grouping cells, which

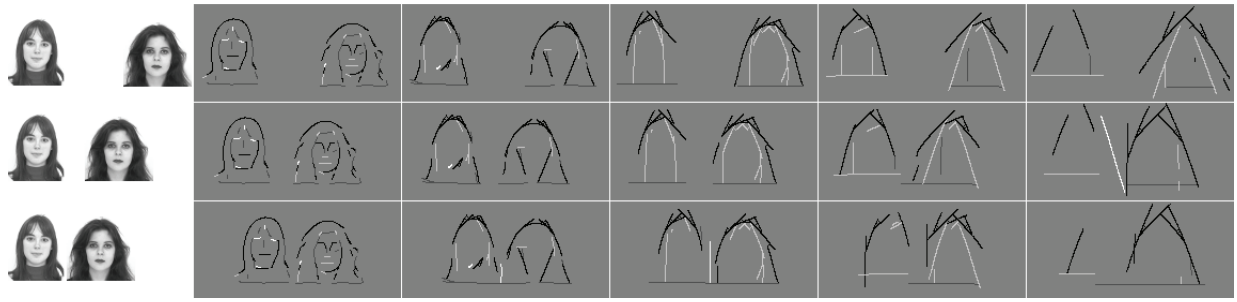

Figure 4.9: Object interference at coarse scales ($\lambda = \{5, 15, 25, 35, 45\}$).

yields big regions-of-influence (Fig. 4.8 top-right). This coarse-to-fine-scale process is complemented by inhibition: other grouping cells at the finest scale are activated by responding event cells at that scale and these grouping cells excitate the grouping cells at the one coarser scale but inhibit active grouping cells outwards, as shown in red in Fig. 4.8 (bottom-right). This results in a figure-ground map at the first coarser scale "above" the finest scale (Fig. 4.8 bottom-left). Results shown were obtained with $\lambda \in [4, 52]$ and $\Delta\lambda = 4$.

A process in V1 as described above can be part of the where system, but it needs to be embedded into a complete architecture. In addition, when two objects are very close, they will become connected at coarse scales, see Fig. 4.9, and separation is only possible by the what system that checks features (lines, edges and keypoints) of individual objects. In other words, object segregation is likely to be driven by "attention" in PF cortex, for example by means of templates that consist of coarse-scale line/edge representations, and this process is related to object categorization.

4.5 Automatic scale selection

Apart from object segregation, other processes may play an important role in the fast where and slower what systems. Concentrating on lines and edges (events)—ignoring other features extracted in V1—there may be many scales and the tremendous amount of information may not propagate in parallel and at once to IT and PF cortex. It might be useful that lines and edges which are most characteristic for an object are extracted and that these propagate first, for example for a rapid object categorization. In Fig. 4.6 we have seen that different criteria for spatial stability over scales lead to different line/edge selections. One possibility is to select only one scale with the most stable lines and edges.

As was done in the case of keypoints, our proposed scheme consists of selecting the scale which counts the maximum number of *stable* events. This can be achieved with a few, simple processes, in which we assume that outputs of event cells are binary.

First, a retinotopic map of grouping cells is assumed. A diagram of event, grouping and gating cells is shown in Fig. 4.10. This diagram is sub-divided into four parts, with the top-left part for positive edges, the top-right part for negative edges, and similarly the bottom parts for positive and negative lines. In a neural layer the four parts can be mixed if their retinotopic mapping is preserved. All four maps show the same positions and scales, with horizontally the position and vertically the scale. The grouping cells marked A have linear dendritic fields (solid black lines) that connect to event cells (solid dots; active cells are big dots). These grouping cells sum all active event cells at their position, over scale,

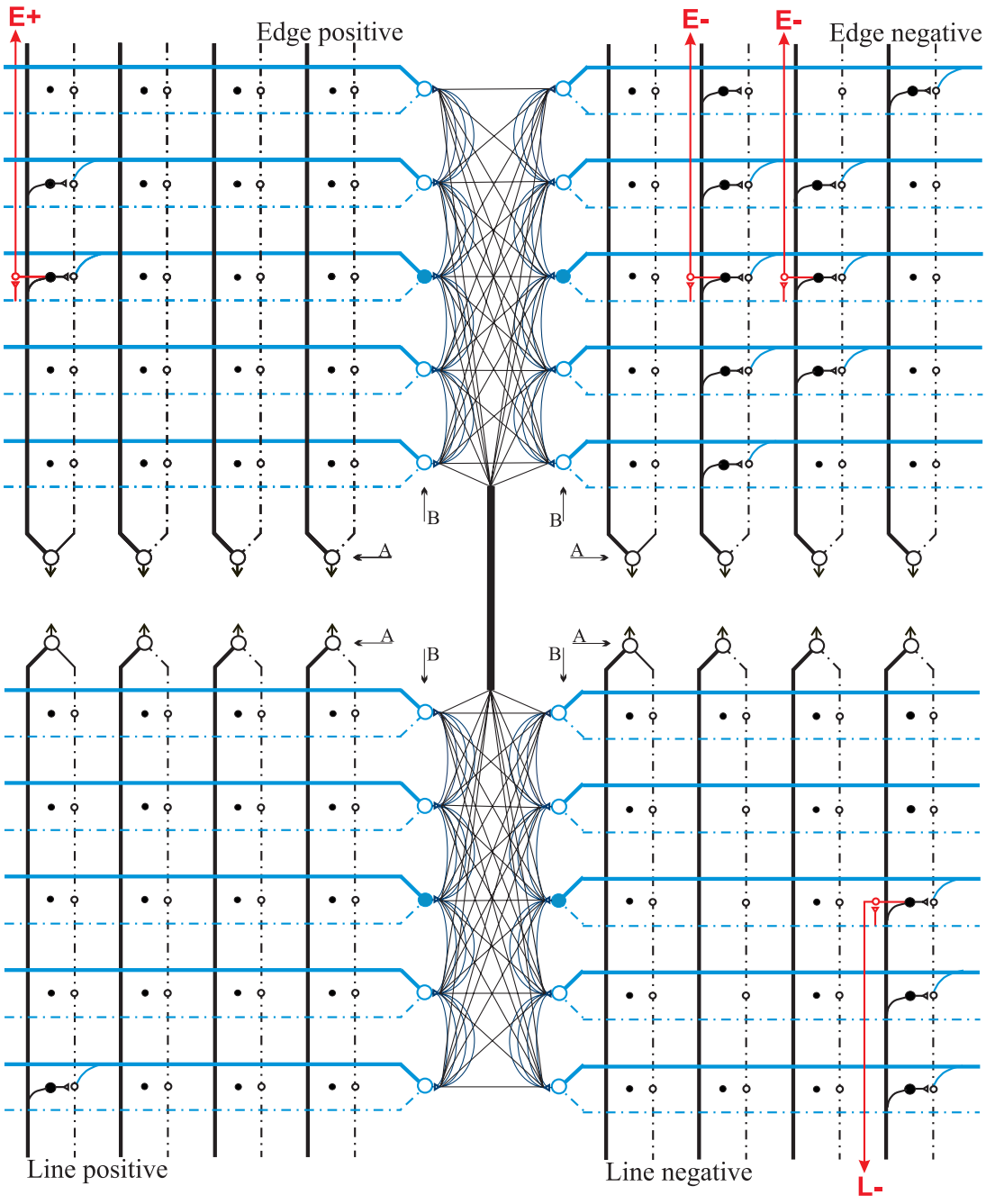

Figure 4.10: Schematic diagram for automatic scale selection, with horizontally the position and vertically the scale. Four event maps are used for positive edges (top-left), negative edges (top-right), positive lines (bottom-left) and negative ones (bottom-right). Events cells are represented by solid dots (active event cells by big dots), grouping cells by big open circles, and gating cells by small open circles. Dendrites are shown by solid lines and axons by dash-dotted lines. The positions and scales in the four maps are the same.

which yields a sort of histogram. Second, at each scale, active event cells activate gating cells (triangular synapses next to open circles); these gate the outputs of grouping cells A (black dash-dotted axons) in the "histogram map" at the same position. Third, at each scale, other grouping cells (marked B) sum outputs of all gating cells. In other words, the

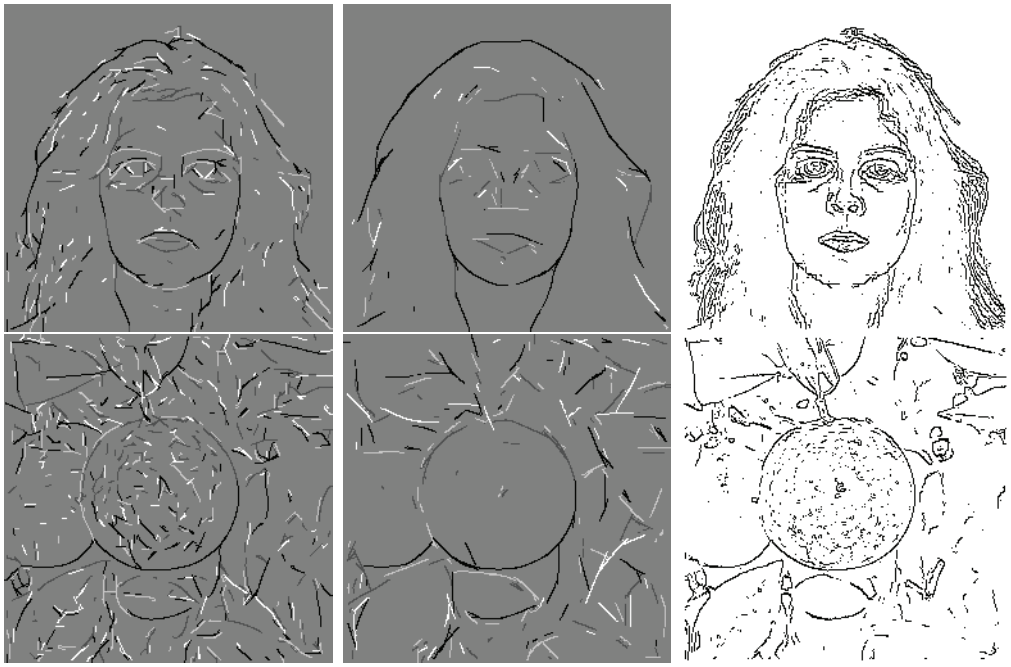

Figure 4.11: Automatic scale selection applied to Fiona and orange images. Left: automatic scale selection without a stability criterion. Center: with stability over 20 scales. Right: for comparison the results obtained with the Canny edge algorithm.

latter grouping cells "count" stable events at all individual scales. Fourth, the grouping cell with maximum activity is selected (winner takes all) and its axon activates other gating cells that gate outputs of event cells at its scale. The outputs of the latter gating cells provide the map which has the maximum number of stable events; see also Section 3.6.

Figure 4.11 (at left) shows results of automatic scale selection without an additional stability criterion in the case of the Fiona and orange images. The center image show results when stability over at least 20 scales is required. Many events have disappeared but the most important ones remain. The right images show, for comparison, results of Canny's edge detector. Results obtained with other edge detectors can be found in Heath et al. [2000].

4.6 Object categorization

Object recognition is a clearly defined task: a certain cat, like the neighbors' red tabby called Toby, is recognized or not. Categorization is more difficult to define because there are different levels, for example (a) an animal, (b) one with four legs, (c) a cat, and (d) a red tabby, before deciding between our own red tabby called Tom and his brother Toby living next door. It is as if we were developing categorization by very young children: once they are familiar with the family's cat, every moving object with four legs will be a cat. With age, more features will be added. Here we explain our experiments with a two-level approach; three types of objects (horses, cows, dogs) are first grouped (animal), which we call *pre-categorization*, after which *categorization* determines the type of animal. Instead of creating group templates in memory on the basis of lowpass-filtered images as proposed by the LF model [Oliva et al., 2003; Bar, 2004], we will exploit coarse-scale line and edge templates.

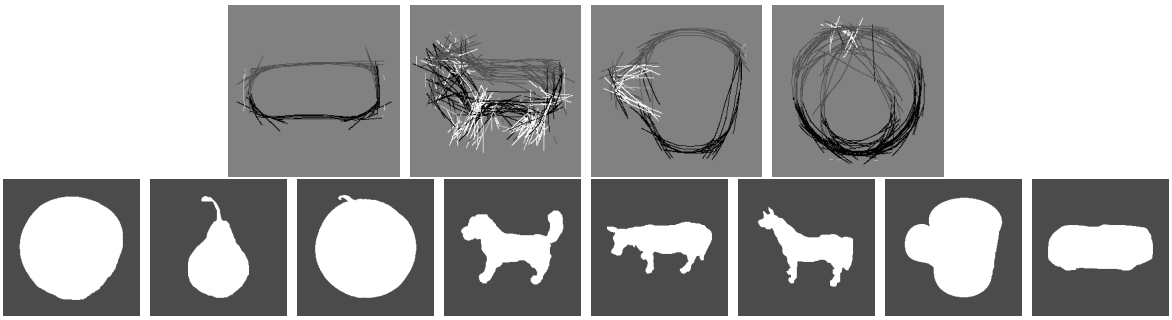

Figure 4.12: Top: templates for pre-categorization based on 15 and 5 images at $\lambda = 32$. Bottom: examples of segregated objects.

In addition, pre-categorization will be based on line and edge templates of contours, i.e., solid objects, available through segregation (Fig. 4.12), to generalize shape and to eliminate surface detail.

We used the ETH-80 database [Leibe and Schiele, 2003], in which all images are cropped such that they contain only one object, centered in the image, plus a 20% border area. Images were rescaled to a size of 256×256 pixels. We selected 10 different images of 8 groups (dogs, horses, cows, apples, pears, tomatoes, cups/mugs and cars), in total 80 images. Figure 4.13 shows examples. Because views of objects are also normalized (e.g. all animals with the head to the left), and because different objects within each group are characterized by about the same line/edge representations at coarser scales, group templates can be constructed by combining randomly-selected images. The multi-scale line/edge representation was computed at 8 scales equally spaced on $\lambda \in [4, 32]$.

4.6.1 Pre-categorization

Here the goal is to select one of the groups: animal, fruit, cup or car. We used the three coarsest scales with λ equal to 24, 28 and 32 pixels. Group templates were created by combining all images (30 animals, 30 fruits, 10 cups, 10 cars), and by random selections of half (15 and 5) and one third (10 and 3) of all images. By using more images, a better generalization can be obtained, for example the legs of animals can be straight down or more to the front (left). Figure 4.12 shows examples of segregated objects and line/edge templates when using half of all images. For each group template, at each of the three scales, a positional relaxation area was created around each responding event cell, by assuming grouping cells with a dentritic field size coupled to the size of underlying complex cells [Bar, 2003]. These grouping cells sum the occurrence of events in the *input images* around event positions in the *templates*, a sort of local correlation, and activities of all grouping cells were then grouped together (global correlation). The final groupings were compared over the 4 templates, scale by scale, and the template with maximum response was selected. Finally, the template with the maximum number of correspondences over the 3 scales was selected. Table 4.1 summarizes results (misclassified images) in the form of mean(st. deviation).

Obviously, positional relaxation leads to better results when not all images are used in building the templates, and using more images is always better. Using relaxation *and* more images increases shape generalization, however with the risk of running into over-generalization, which did not occur in our tests. On the average, different random selections gave very similar results when the three sub-groups (horses/cows/dogs and ap-

ples/pears/tomatos) were about equally represented. Most errors occurred, with and without relaxation, between car/animal and cup/fruit. These errors can be explained by the global correlations between the elongated (car/animal) and round (cup/fruit) shapes, see Fig. 4.12.

4.6.2 Categorization

After pre-categorization, assuming zero errors, there remains one problem in our test scenario: the animal group must be separated into horse, cow and dog, and the fruit group into apple, pear and tomato. We could have used 6 templates (cups and cars have already been categorized), but we experimented with 8 templates and all 80 images, and applied the multi-scale line/edge representations at all 8 scales (λ equal to 4, 8, 12, 16, 20, 24, 28 and 32) of the real input images (not of the solid, segregated objects). We did this because categorization is supposed to be done *after* pre-categorization, i.e., when also fine-scale information has propagated to IT cortex (see Introduction).

Templates were constructed as above with random selections. Final groupings (global correlations) were compared over the 8 scales and the one with most coherent (maximum) correspondences was selected (in the case of 4–4 we simply took the last one). Table 4.1 presents results (misclassifications) obtained with positional relaxation.

	all	half	third
pre-categorization template construction	30/10	15/5	10/3
error without relaxation	0.0%	5.7%(0.6)	8.0%(1.7)
error with relaxation	0.0%	3.0%(1.0)	4.3%(0.6)
categorization template construction	10	5	3
error with relaxation	0.0%	9.3%(2.1)	12.7%(4.0)

Table 4.1: Results obtained with pre-categorization and categorization.

Again, by using more images in building the templates, generalization is improved and the number of miscategorized images decreases. When using half (5) or even one third (3) of all images, all car and cup images were correctly categorized, and no fruits were categorized as animals and vice versa. Typical miscategorizations were dog/cow, horse/dog, horse/cow and apple/tomato. Figure 4.13 shows, apart from examples of images and group templates created by combining 5 images (top), the more difficult images with a white triangle in the bottom-right corner. It should be stressed that this is an extremely difficult test, because no color information has been used and apples and tomatos have the same, round shape. By contrast, all pear images, with a tapered shape, have been correctly categorized. The fact that most problems occurred with the animals was expected, given the small differences of heads, necks and tails (Fig. 4.13). Categorization is the last step before recognition in which attention shifts to finer scales that reflect minute differences. Nevertheless, only about 9 errors in 80 images (the "50/50 training and testing" scenario) is a very promising starting point for refining the algorithms, for example by using a more hierarchical scenario with more categorization steps, in which attention is systematically steered from coarse to fine scales.

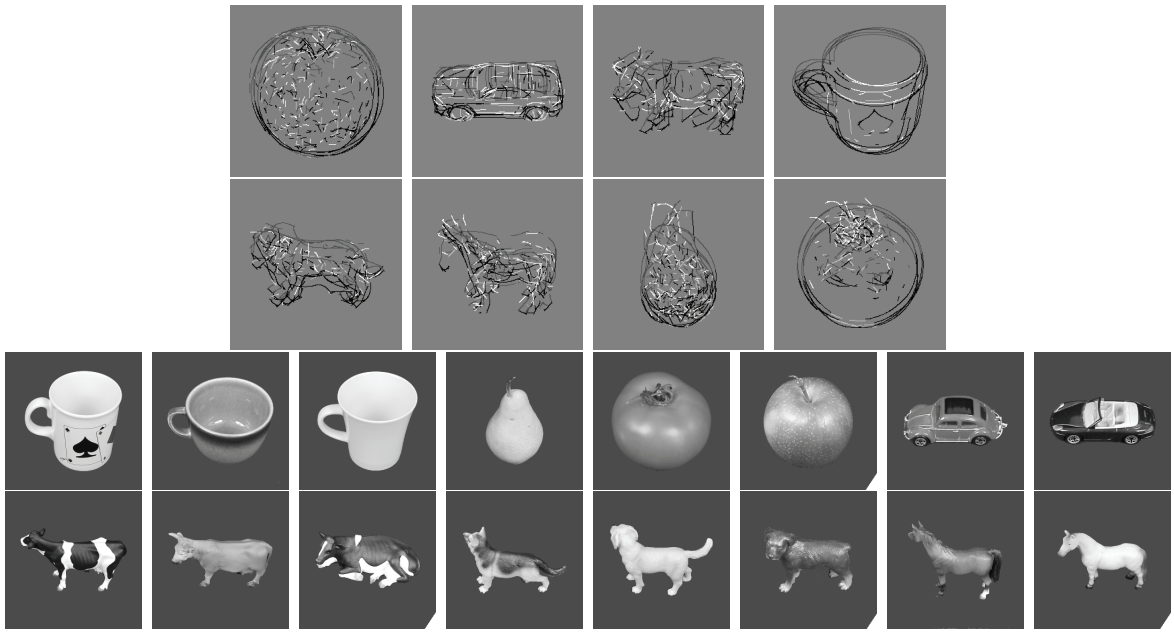

Figure 4.13: Top: templates for final categorization based on 5 images at $\lambda = 8$. Bottom: examples of object images, the more difficult ones are marked by a white triangle in the bottom-right corner.

4.7 Face recognition

The final goal in vision is object recognition, but here we focus on face recognition by the multi-scale line and edge representations. This completes face detection as presented in the previous chapter (based on [Rodrigues and du Buf, 2006d]), in which saliency maps and the multi-scale keypoint representation have been used for detecting facial landmarks and thus faces. In addition, it was also shown that keypoints can be used for Focus-of-Attention, i.e., to "gate" detected keypoints in associated Regions-of-Interest. The same process can be used to gate detected lines and edges in the Regions-of-Interest. The idea of combining keypoints with lines and edges resembles the bottom-up data streams in the where (FoA) and what (lines/edges) subsystems; for more details see Rodrigues and du Buf [2006b]. Of course, this is a simplification because processing is limited to cortical area V1, whereas in reality the two subsystems contain higher-level feature extractions in areas V2, V4 etc. [Hamker, 2005]. The same way, top-down data streams are simplified by assuming that stored face templates in memory, that have been built through experience, are limited to lines and edges, and that a few canonical views (frontal, 3/4) are normalized in terms of position, size and rotation: faces are expected to be vertical; for translation, size and rotation invariance see e.g. Deco and Rolls [2004] and Chapter 5. An additional simplification is the strict attributions of keypoints and lines/edges to the two subsystems: keypoints can also be used in the what system and line and edges also in the where system.

In our experiments we use 8 primary scales $\lambda_1 = \{4, 8, 12, 16, 20, 24, 28, 32\}$ with $\Delta\lambda_1 = 4$. Each primary scale is supplemented by 8 secondary scales with $\Delta\lambda_2 = 0.125$, such that, for example, $\lambda_{2,\lambda_1=4} = \{4.125, 4.250, ..., 5.000\}$. These secondary scales are used for stabilization. The model consists of the following steps:

(A) Multi-scale line/edge detection and stabilization. To select the most relevant facial features, detected events must be stable over at least 5 scales in a group of 9 (1 primary

Figure 4.14: Examples of images of eleven persons seen against a dark or bright background with different size normalizations.

plus the 8 secondary scales).

(B) Construction of four symbolic representation maps. At each primary scale, stable events (positions) are expanded by Gaussian cross-profiles (lines) and bipolar, Gaussian-truncated errorfunction profiles (edges), the sizes of which being coupled to the scale of the underlying simple and complex cells; see Fig. 4.7 (the four leftmost columns). Responses of complex cells are used to determine the amplitudes of the profiles. As a result, each face image is represented by 4 maps at each of the 8 primary scales.

(C) The recognition process. We assume that templates (views) of faces are stored in memory and that these have been built through experience. Template images of all persons are randomly selected from all available images: either one frontal view or two views, i.e., one frontal plus one 3/4 view; see also Valentin et al. [1997]. Each template in memory is thus represented by 32 line/edge maps (point B above). Two recognition schemes have been tested:

Scheme 1: At each scale, events in the 4 representation maps (the 4 leftmost columns in Fig. 4.7) of an input image are compared with those in the corresponding maps of a template. Co-occurrences are summed by grouping cells, which yields a sort of event-type and scale-specific correlation. Then, the outputs of the 4 event-type grouping cells are summed by another grouping cell (correlation over all event types). This results in 8 correlation factors. These factors are compared, scale by scale, over all templates in memory, and the template with the maximum number of co-occurrences over the 8 scales will be selected (in the case of equal co-occurrences we simply select the second template).

Scheme 2: Instead of comparing representations scale by scale, only one global co-occurrence is determined by more levels of grouping cells, i.e., first over maps of specific event types, then over event types, and finally over scales. The template with the maximum is selected by non-maximum suppression.

From the Psychological Image Collection at Stirling University (UK) we selected 100 face images of 26 persons in frontal or frontal-to-3/4 view, with different facial expressions. From those, 13 persons are seen against a dark background, with a total of 53 images, of which 40

images are in frontal view, 11 images are in (very near) 3/4 view (4 persons), and 2 frontal images with added Gaussian and speckle noise (1 person). The other 13 persons (47 images) are seen against a light background, in frontal or near-frontal view. For typical examples see Fig. 4.14. All persons are represented with at least 3 different facial expressions. In view of the tremendous amount of data already involved in our simple experiments, huge databases cannot (yet) be processed.[1]

All recognition tests involved the entire set of 100 images, although results will also be specified in terms of the subsets of 53 and 47 images in order to analyze the influence of the two different backgrounds and size normalizations. For each person we used two different types of templates: (1) only one frontal view, and (2) two views, frontal and 3/4, but only in the case of 4 persons represented by images in frontal and 3/4 views. In all cases, template images were created by random selection of input images. In order to study robustness with respect to occlusions, a second set of tests was conducted in which partially occluded representations of input images were matched against complete representations of templates.

Table 4.2 presents the results by testing all images ("all") and by specifying (splitting) these in the case of a dark ("black)" or light ("white") background. The penultimate column "scales" lists the percentage of correct scales that lead to correct recognition in the case of "all" and Scheme 1, where 100% corresponds to 800 because of 8 scales and 100 images. The last column "base line" lists the number of all 100 images that have been recognized with absolute certainty, i.e., when Scheme 1 and 2 and all scales point at the same person.

recogn. scheme	2	2	2	1	1	1	1	base
images	all	black	white	all	black	white	scales	line
frontal view	91.0	90.6	91.5	89.0	86.8	91.5	85.5	71
frontal plus 3/4	96.0	100.0	91.5	96.0	100.0	91.5	91.8	81

Table 4.2: Results of face recognition, without partial occlusions.

Comparing columns "all," "black" and "white," there are significant differences because dark and blond hair against dark and light backgrounds cause different events, or even no events, at the outline of the hair. Although the "all" results are reasonably close to the best results, separation of different backgrounds can lead to better but also to worse results. This aspect certainly requires more research. Best results were obtained when using two templates with frontal and 3/4 views. Using all events, both recognition schemes yielded a recognition rate of 96%, whereas 81 was the base line with absolute certainty. The difference of 15% is due to relative ranking with some uncertainty. In future research it will make sense to increase the base line, especially when larger databases with more variations will be considered. It should be mentioned that small changes in the hairstyle, or in the face like spectacles (Fig. 4.14 3rd row, second from left), or even small pose changes (Fig. 4.14 4th row, two leftmost) did not much affect classification, as expected, due to the generalization at coarse scales. However, dramatic changes like the one shown in Fig. 4.14 4th row, the five rightmost images, which show Kirsty before and after a change of hairstyle, lead to incorrect results if we consider only one group, but to correct results if we consider two groups, before and after.

[1]One hundred images of 256×256 pixels, with 72 scales and at each scale 4 representation maps, plus the necessary storage capacity for responses of simple and complex cells necessary for line/edge detection, most in floating-point precision.

Figure 4.15: Occlusion types 1 to 5 from left to right.

Our best result of 96.0% is a little bit better than the 94.0% obtained by Petkov et al. [1993a] and the 93.8% by Hotta et al. [2000], and it is very close to the 96.3% reported by Ekenel and Sankur [2005], despite the fact that in all studies the number of tested faces and the databases are different.

In the last experiments we tested the influence of the 5 types of occlusion as shown in Fig. 4.15, using all 100 images and applying recognition Scheme 2 with templates that combine frontal and 3/4 views. Because of the tremendous amount of storage space (and CPU time) involved, all representations were not re-computed (500 images!) but the occlusions were directly applied to the already computed representations, thereby suppressing event information in the recognition process. This is an approximation of real occlusions, but it indicates the relative importance of different facial regions in the recognition scheme. Table 4.3 presents results in terms of "rate (base line)," which must be compared with the bottom part of Table 4.2, i.e., the first and last columns.

In the "all events" case and occlusion type 4, instead of 81% only 64 was obtained. But this is the base line: 64 of all 100 images are correctly classified with absolute certainty. The maximum rate for this occlusion (93%) is very close to the maximum without occlusion (Tab. 4.2, 96%), and slightly worse if compared to the other occlusions. This shows that the multi-scale representation, in particular the shape of the head and hair at the coarser scales, is very robust and contributes most to the recognition. The reason for this can be seen in Fig. 4.7: the stable and "sketchy" information without too much detail at coarse scales. Nevertheless, some contradiction seems to appear when we exclude the eyes (occlusion type 5). In this case we expected a small decrease in performance relative to occlusion types 1 to 3, but it resulted in the best performance of 97%. An analysis learned that this is due to only one image that failed recognition in occlusion types 1 to 4 but *not* in type 5. In contrast, the base line is lower, as expected (75 instead of 81). The main conclusion is therefore that face and hair contribute about equally to face recognition.

occlusion type	1	2	3	4	5
scheme 2	96.0 (80)	95.0 (74)	96.0 (67)	93.0 (64)	97.0 (75)

Table 4.3: Results obtained with partial occlusions for the frontal plus 3/4 views.

4.8 Disparity estimation

An additional source of information that will improve categorization and recognition rates is disparity or 3D depth. Here we present a disparity model which is not based on explicit phase extraction [Fleet et al., 1991] nor amplitude summations [Ohzawa et al., 1997]. The

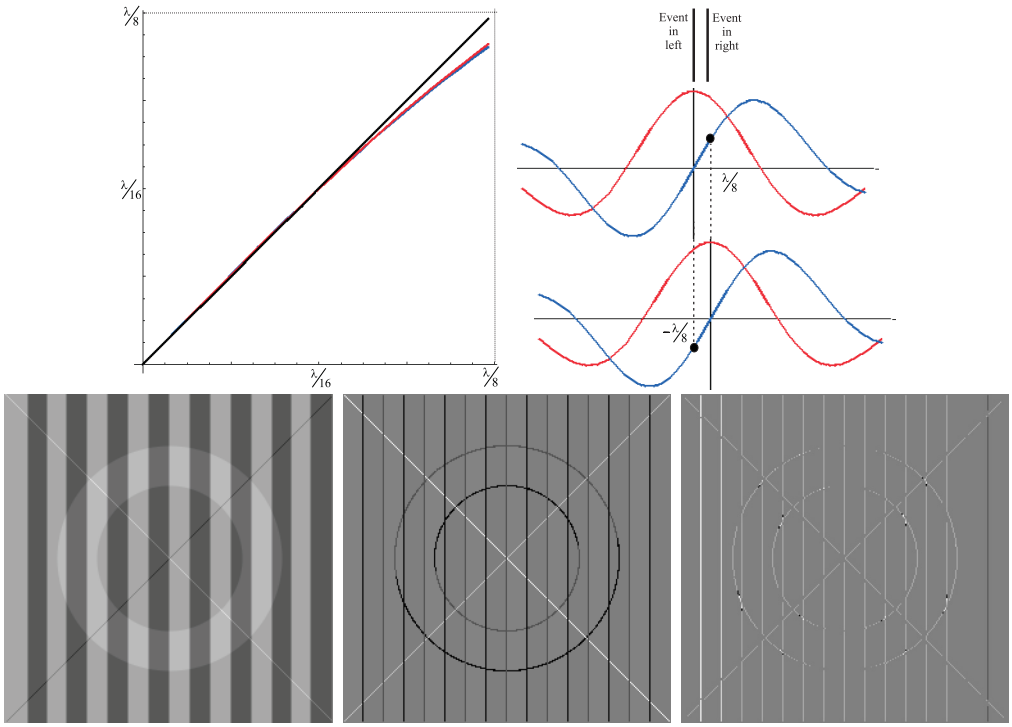

Figure 4.16: Top-left: linear Gabor responses on $[0, \lambda/8]$, at a line (red) and at an edge (blue). Top-right: disparity extraction (see text). Bottom, left to right: the "ledge" test image, line and edge coding, and 2D representation of depth coded by gray level.

new disparity model is only based on the multi-scale line and edge coding and uses responses of simple cells which are already available in the models.

The model is based on the central linear part of the Gabor responses, i.e., the sinusoidal part with $\sin x \approx x, |x| < \pi/4$. Assuming ideal events, i.e., lines with a Dirac profile and edges with a Heaviside step profile, or nonideal ones obtained by Gaussian filtering, plus complex Gabor filters with the same orientation, the responses are (re-scaled) Gabor functions and complex errorfunctions. It has been shown that the latter can be approximated by scaled Gabor functions [du Buf, 1993]. In other words, both line and edge responses are essentially scaled Gabor functions with the sinusoidal part, real or imaginary, being linear on $\pm\lambda/8$; see Fig. 4.16. One step in line/edge detection consists of checking the Gabor response $R_i^O(x, y)$ (the odd, imaginary part in the case of a line), or $R_i^E(x, y)$ (the odd, real (!) part in the case of an edge) for a zero crossing on $\pm\lambda/4$ (Fig. 4.16 top-right). Here, for disparity, we apply the same event detection steps to two images, left and right. In the case of the *left* image, we (1) check the existence of an event of the same type in the *right* image on $\pm\lambda/8$, and (2) if so, we take $\pm R^O$ or $\pm R^E$ of the *right* image at the event (zero crossing) position in the *left* image. The sign depends on the event polarity and, in order to obtain values which do not depend on the event amplitude, $\pm R^O$ or $\pm R^E$ is divided by the modulus (complex cell response) of the *left* image, which is maximum at the event position. After this normalization yet another one is applied: the response is divided by the scale s of the filter. Hence, the slope of the linear response part will not depend on the event amplitude nor on the filter scale, i.e., disparity estimates obtained at different scales will be the same.

The same processing can be done in the case of the right image, by exchanging *left* and *right*. Of course, the disparity estimates need to be calibrated once using real data, like the

way babies need to learn distances in the first months. One problem we encountered were small fluctuations of the disparity estimates, especially at the finest scales. These are due to the fact that we need to work at discrete pixel positions, and the maximum of the modulus used in the first normalization is therefore not the theoretical maximum. We solved this by averaging disparity estimates over neighboring micro-scales.

An example of disparity estimation (for now only tested with a synthetic image) is shown in Fig. 4.16 bottom-right. The stereo images were obtained by shifting left, in one copy of the "ledge" test image (Fig. 4.16 bottom-left), the first vertical edge 3 pixels, the following edge 2, and the next edges 1 pixel. The second-last edge was not changed, whereas the last one was shifted right. The diagonal lines and the ring were shifted left 1 pixel. In Fig. 4.16 (bottom-right) the disparity is coded by gray level. Figure 4.17 shows in color the types of events and disparity in 3D, and at the bottom the keypoint vertex structure (see Section 3.3 or Rodrigues and du Buf [2006d]). There are still some problems with disparity detection around keypoints, and experiments with more synthetic and real images are now being carried out to study and solve the problems.

4.9 Discussion

Computer vision for realtime applications requires tremendous computational power because all images must be processed from the first to the last pixel. Probing specific objects on the basis of already acquired context may lead to a significant reduction of processing. This idea is based on a few concepts from our visual cortex [Rensink, 2000]: (1) our physical surround can be seen as memory, i.e., there is no need to construct detailed and complete maps, (2) the bandwidth of the what and where systems is limited, i.e., only one object can be probed at any time, and (3) bottom-up, low-level feature extraction is complemented by top-down hypothesis testing, i.e., there is a rapid convergence of activities in dendritic/axonal cell connections from V1 to PF cortex.

In previous papers and in Chapter 3 we have shown that keypoint scale-space is ideal for constructing saliency maps for Focus-of-Attention (FoA) [Rodrigues and du Buf, 2005b], and that faces can be detected by grouping facial landmarks defined by keypoints at eyes, nose and mouth [Rodrigues and du Buf, 2005c]. On the other hand, line and edge scale-space may be ideal for object and face recognition. Obviously, these two representations in V1 complement each other and both can be used for object detection, categorization and recognition. Our impression is that keypoints provide better information for the fast where system (FoA), whereas lines and edges are better suited for the slower what system. However, this still needs to be tested in the context of a complete cortical architecture with ventral and dorsal data streams that link V1 to attention in PF cortex [Deco and Rolls, 2004].

In this chapter we presented an improved scheme for line and edge detection in V1, and illustrated the multi-scale representation for visual reconstruction. This representation, in combination with a lowpass filter, yields a (re)construction that is suitable for extending our brightness model [du Buf and Fischer, 1995] from 1D to 2D, for example for modeling brightness illusions (see Chapter 6).

We also presented a plausible scheme for object segregation, which results in binary, solid objects that can be used to obtain a rapid pre-categorization on the basis of coarse-scale information only. This approach works much better if compared to using lowpass-filtered images, i.e., smeared blobs that lack object-specific characteristics [Oliva et al., 2003; Bar,

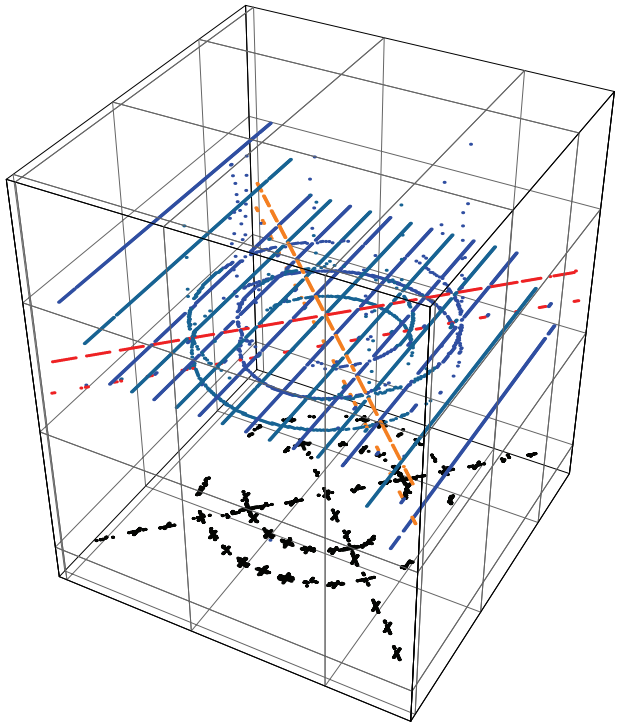

Figure 4.17: 3D wireframe representation of lines, edges, depth and vertex structures.

2004]. Final categorization was tested by using the real objects and more scales, coarse and fine. The results obtained are very promising, taking into account that the tested schemes are extremely simple. Only a fraction of available information, i.e., the line/edge code without amplitude and color information, and without a linking of scales as explored in the segregation model, has been used so far. More extensive tests are being conducted, with more images and objects, concentrating on a linking of scales and a steering of attention from coarse to fine scales. Such improved schemes are expected to yield better results, from very fast detection (where) to slower categorizations (where/what) to recognition (what). The balance between keypoint and line/edge representations in these processes is an important aspect.

The line and edge interpretations at coarser scales lead to stable abstractions of image features (Figs 4.5 and 4.9). This explains, at least partly, the generalization that allows to classify faces with noise, spectacles, and relatively normal expressions and views (Fig. 4.14). It should be stressed that the recognition scheme is not yet complete, because a hierarchical linking from coarse to fine scales, as already applied in the detection/segregation process, has not been applied. Such an extension can lead to better recognition rates, especially when multiple views (frontal, 3/4 and lateral) of all persons are included as templates in memory. In addition, the multi-scale keypoint representation [Rodrigues and du Buf, 2005c] (see Chapter 3), which has been ignored here, will contribute very important information.

Finally, we presented a new disparity model which, although still being in an initial development stage, allows to directly attribute depth to lines and edges and thereby create a 3D "wireframe" representation (Fig. 4.17). Such a wireframe representation is used in modeling solid objects in computer graphics. The fact that many simple and complex cells are disparity tuned suggests that our visual system processes 3D objects in the same way, probably simplifying 3D object recognition.

All multi-scale processing and the representations, including keypoints, are restricted to areas V1 and V2. On the other hand, the Deco and Rolls scheme [Deco and Rolls, 2004], with ventral and dorsal data streams, necessary for obtaining position and size invariance through projections via areas V2, V4 etc., is solely based on responses of simple cells. In the future (see Chapter 5), this scheme must be based on features extracted in V1, and further multi-scale processing can be added in the higher areas V2 to PF. We expect that such extensions in adaptive up and down projections will lead to much better results. Afterall, our visual system does not have any difficulty in categorizing objects or telling apart persons!

Chapter 5

Integrated architecture

Abstract: Object categorization and recognition require that templates with canonical views are stored in memory. Such templates must somehow be normalized, yet representing an object at all retinotopic positions with many rotations and sizes. Partial invariance can be obtained by dynamic routing from cortical area V1 via V2 and V4 to higher areas. Our recent work (see Chapters 3 and 4) on feature extractions in V1, which yield multi-scale representations of lines, edges and keypoints, allows to develop a functional model with coarse-to-fine-scale processing, from object segregation via different categorization levels until final recognition is achieved. In this chapter we present a novel method for obtaining 2D translation, rotation and size invariance. Dynamic routing of major peaks in saliency maps allows feature maps of input objects and stored templates to converge. We illustrate the construction of group templates and the invariance method in the context of an integrated cortical architecture.

5.1 Introduction

Object detection, segregation, categorization and recognition are linked processes which cannot be completely sequential; they must be done in parallel, at least partially, and therefore they are overlapping significantly [Rensink, 2000]. These processes are achieved in the ventral "what" and dorsal "where" pathways [Deco and Rolls, 2004], with bottom-up feature extractions in areas V1, V2, V4 and IT (what) in parallel with top-down attention from PP via MT to V2 and V1 (where). The latter is steered by possible object templates in memory, i.e., in prefrontal cortex with a what component in PF46v and a where component in PF46d. The Deco and Rolls model can explain invariance and attention, also the facts that cells at higher cortical areas have bigger receptive fields and are coding more complex patterns. However, their model is based on simple cells in V1, whereas we are aiming at functional feature extractions in V1 and beyond, for example face detection at high level by grouping outputs of eye and mouth detectors at medium level, the latter detectors combining keypoints at low level [Rodrigues and du Buf, 2006d]. The ultimate goal is to integrate feature extractions into a cortical architecture, although even relatively simple computational

models require tremendous storage capacity. This is not a surprise, since our brain counts about 10^{12} cells—roughly 150 times the earth's population count in 2005—with 10^{14} to 10^{15} interconnections [Hubel, 1995], a significant part of which being devoted to vision.

We are studying three related problems: when, where and how does categorization take place [Nunes et al., 2006b]. The "when" problem allows for two hypotheses. The easy one is to assume that categorization occurs after recognition [Riesenhuber and Poggio, 2000a]: if specific neurons respond in the case of recognizing dog-1, dog-2 and dog-3, a grouping cell can combine all responses: a dog. This view is too simplistic, because the system must collect evidence for a specific object or object group in order to select possible templates in memory. For example, when we glance a portrait made by Arcimbaldo, the famous, 16th-century Italian painter, our first reaction is "a face!" but then follows "fruits?" and finally "ah, the cheek is an apple!"

When categorization occurs before recognition [Grill-Spector and Kanwisher, 2005], the "where" problem is, at least partly, solved: it must take place at a very high level, with access to object templates, and just before recognition. In fact, recognition can be seen as a last categorization step. Therefore, the "how" problem can be solved by taking into account feature extractions in V1 and beyond and the propagation of features to higher cortical areas. In the previous two chapters, we concentrated on the extraction of low-level primitives: lines, edges and keypoints, all multi-scale; see also [Rodrigues and du Buf, 2004b, 2006a,d,b]. Keypoint scale space provides ideal information for constructing saliency maps for Focus-of-Attention (FoA), and the grouping of keypoints at different scales is robust for e.g. face detection; see [Rodrigues and du Buf, 2006d] or Section 3.8. Therefore, keypoints and FoA are thought to be major cornerstones of the where system. In parallel, it was shown that the multi-scale line/edge representation provides ideal information for object and face recognition; see [Rodrigues and du Buf, 2006b] or Section 5.3 below). The latter may be done in the what system. However, detection in the fast where pathway (a face!) must be linked with categorization and recognition in the slower what pathway (whose face?). The balance between the use of lines/edges and keypoints in the two pathways is still an open question.

A less open question concerns the use of features detected at different scales: information at coarse scales propagates first to higher areas, after which information at progressively finer scales arrives there [Bar, 2004]. This probably implies that coarse-scale information is used for a first, fast but rough categorization, after which categorization is refined using information at progressively finer scales, until the object is recognized. It has been proposed that a first categorization is based on a lowpass-filtered image of the object [Bar, 2003], but a smeared blob lacks structure. In our own experiments ([Rodrigues and du Buf, 2006a] or Section 4.6) we therefore applied a different approach: after segregation, the coarse-scale line/edge representation of the solid object (outline) is used for pre-categorization, after which all information is used for final categorization and then recognition.

Any 3D object can lead to an infinite number of different projected images on the retinae, due to variations in position, distance, lighting and other factors including rotation and deformation. Nonetheless, we recognize familiar objects in a manner which is largely invariant to such transformations. The ability to identify objects despite all possible transformations is central to visual object recognition. However, this still is a poorly understood mechanism [Cox et al., 2005] and transform-tolerant recognition remains a major problem in the development of artificial vision systems. In our brain, transform-invariant object recognition is automatic and robust, but it ultimately depends on experience [Tarr, 1995; Wallis and Bülthoff, 2001; Cox et al., 2005]. Recent findings (e.g. [Cox et al., 2005]) even support the

idea that visual representations in the brain are plastic and largely a product of our visual environment, and that invariant object representations are not rigid nor finalized—they are continually evolving entities, ready to adapt to changes in the environment. This idea complicates the classical idea of static representations in which only two but related problems need to be solved: (1) partial invariance to reasonable transformations like 2D rotation in the case of any canonical object view, which is addressed in this chapter, and (2) the total number of (3D) canonical object views that must be stored in memory. However, also plasticity can be explored at the two levels, in this chapter in the form of dynamic routing for obtaining partial invariance to reasonable transformations.

One goal of this chapter is to show that low-level processing in terms of multi-scale feature extractions can be extended to higher-level processing: invariance in object categorization and recognition. As a consequence, extended models can cover more cognitive aspects in the near future. For example, processes like the learning of new objects or new, unexpected views of known objects become subject to explicit modeling. In the next section we introduce existing models of object recognition. Section 5.3 deals with partial and global saliency maps and face recognition. Invariance by dynamic routing, the construction of group templates and experimental results are presented in Section 5.4. Final sections concern an integrated architecture and discussion with lines for future research.

5.2 Recognition models

There are several approaches to biological object recognition; see [Riesenhuber and Poggio, 2000b] for a review. In this section we focus on approaches which, to some degree, are related to our own approach and architecture, or because of their importance in terms of results.

SIFT or Scale Invariant Feature Transform [Lowe, 2004] has no profound biological background. Local invariant features allow to efficiently match small parts of cluttered images with arbitrary rotations, sizes, changes of brightness and contrast, and other transformations. The basic idea is to partition the image into many small but overlapping pieces, each of which is described in a way invariant to the possible transformations. Then each piece can be matched to known objects in a database. In the matching process, keypoint descriptors are used which are highly distinctive; this allows that even a single feature can be linked to its correct match, with significant probability, in a large database of features. It should be emphasized that Lowe's keypoint descriptors are completely different from our own keypoints (see Chapter 3), and that the resulting characteristics are also different. Despite all differences, both Lowe's and our keypoints represent highly distinctive points in an image or scene.

The Laterally Interconnected Synergetically Self-Organizing Map or LISSOM model [Miikkulainen et al., 2005] consists of a "family" of computational models which aim to replicate the detailed development of the visual cortex. Its rationale is that the cortex organizes itself, using Hebbian learning to adapt feedforward and lateral interconnections between neurons, in order to capture correlations in both visual inputs and internally generated sources of activation. Originally implemented at the V1 level, but with extensions down to the LGN and retina as well as up beyond area V1, the model can explain invariant (only size and viewpoint) detection of objects, e.g. faces. SpikeNet, proposed by Thorpe et al. [2004], is also a bottom-up process. It is based on a novel coding scheme that uses the order in which cells fire spikes, rather than firing rates, to encode information. It was shown that such coding can yield recognition of objects when as few as 1% of the neurons in the model have fired

a spike. Hamker [2005] presented a feature based computational model for invariant (only translation) object detection in complex backgrounds (natural scenes) driven by attention in V4 and IT. Petkov and colleagues (see e.g. [Petkov and Kruizinga, 1997]) focus on the development of biologically-motivated image processing and computer vision algorithms; see also [Ghosh and Petkov, 2005]. They use functional descriptions of different types of neurons in the cortex to develop computational models, for example of bar and grating cells (see also [du Buf, 2007] for improved models of bar and grating cells). Petkov and his colleagues also demonstrated the effect of the neural mechanism known as non-classical receptive field inhibition or surround suppression [Grigorescu et al., 2003], for suppressing edge detection in textured regions. The latter implies two "pathways," one for object edges for object detection and the other for texture processing and therefore object surfaces.

The collaboration called "Detection and Recognition of Objects in the Visual Cortex" integrates effort at several laboratories in the USA: Poggio, DiCarlo and Miller at MIT, and Riesenhuber and colleagues at Georgetown University. The aim is a quantitative hierarchical model of recognition, probing the relations between identification and categorization and the properties of selectivity and invariance of neural mechanisms in IT cortex (see e.g. [Walther et al., 2002; Riesenhuber, 2005]). An integrated architecture, much like our own, reflects the general organization of the visual cortex in a stack of layers from V1 to IT to PF cortex ([Riesenhuber and Poggio, 2000c; Serre and Riesenhuber, 2004]). With respect to invariance properties, it consists of a sequence of two main modules based on two key ideas: (1) a "maximum operator" that provides invariance at several levels of the hierarchy, and (2) a neural network, that learns a specific task, is based on a set of cells tuned to example views. Ferster and colleagues at Northwestern University are concentrating on the pooling operation—the maximum vs. linear sum of inputs—performed by complex cells in V1 [Lampl et al., 2004]. Koch and colleagues at Caltech are extending the basic recognition model by integrating a saliency-based and essentially bottom-up attentional model [Walther et al., 2005].

In Rensink's [Rensink, 2000] triadic architecture, early preattentive processes feed into both an attentional system concerned with coherent objects, and a non-attentional system concerned with scene gist and layout. Instead of operating sequentially, the latter two subsystems operate concurrently for providing a context that can guide the allocation of attention. In this view, attention is no longer a central gateway through which all information must pass, but just one system that operates concurrently with several other (sub)systems. Furthermore, a scene is experienced via a "virtual representation" in which object representations are formed in a "just-in-time" fashion, only existing as long as they are needed.

Deco and Rolls [2004] presented an invariant model that incorporates feedback-biasing effects of top-down attentional mechanisms in a hierarchically-organized set of cortical areas with convergent feedforward connectivity, reciprocal feedback connections and local area competition. The model displays space-based and object-based covert visual search by using attentional top-down feedback from either the PP or the IT cortical modules, with interactions between the ventral and dorsal data streams occurring in V1 and V2. The same authors in [Deco and Rolls, 2005] described a computational framework and showed how an attentional state held in short-term memory in PF cortex can, by top-down processing, influence the ventral and dorsal data streams in different cortical areas. Biased competition can account for many aspects of visual attention. They even showed how an attentional bias within PF cortex can influence the mapping of sensory inputs to motor outputs. Stringer et al. [2006] showed that invariant object recognition can be based on spatio-temporal continuity (during object translation and rotation) with "continuous transformation (CT) learn-

ing," which operates by mapping spatially similar input patterns to the same postsynaptic neurons in a competitive neural network system.

Olshausen et al. [1993] described a model that relies on a set of control neurons, which dynamically modify the synaptic strengths of intracortical connections such that information from a windowed region of the primary cortex is selectively routed to higher cortical areas. Local spatial relationships (i.e. topography) within the attentional window are preserved as information is routed through the cortex. This enables attended objects to be represented in higher areas within an object-centered reference frame that is position and size invariant. Olshausen et al. hypothesize that the pulvinar (at the posterior part of the thalamus) may provide the control signals for routing information through the cortex. In preattentive mode, the control neurons receive their input from a low-level "saliency map" representing potentially interesting regions of a scene. During the pattern-recognition phase, control neurons are driven by the interaction between top-down (memory) and bottom-up (retinal input) sources. Of all models, this one is the most similar to our own model; see also Discussion.

Oliva and Torralba [2006] proposed that different aspects should also be considered for scene recognition, such as our possibility to understand the meaning (gist) of a complex and new scene very quickly, even when the image is blurred and/or presented for a very short time. Evidence from behavioral, imaging and even computational studies on fast scene perception suggest an alternative view on the role of objects as being the most important elements for constructing gist. Oliva and Torralba argued that we do not need to perceive the objects in a scene in order to establish its semantic meaning, at least not in early stages of processing. Mechanisms involved in natural scene recognition may be independent from those involved in recognizing objects, and fast scene recognition does not need to be built on top of the processing of objects, but can be analyzed in parallel by scene-centered mechanisms. A scene image is initially processed as a single entity and local information about objects and parts comes into play at a later processing stage. In addition, in contrast to the traditional view that cortical analysis related to object recognition involves serial information propagation along a bottom-up hierarchy in ventral and/or dorsal areas [Bar et al., 2006], recent findings support the idea that top-down mechanisms play an important role. But it remains puzzling how such processing would be initiated.

Indeed, the existence of top-down processes that facilitate or steer recognition implies that high-level cortical areas, such as the left orbitofrontal cortex, are activated at a very early stage. Bar et al. [2006] proposed that a partially analyzed version of the input image (i.e. a very blurred image, comprised of the low spatial frequency components) is projected rapidly from early visual areas directly to PF cortex, possibly in the dorsal pathway. This coarse representation is subsequently used to activate predictions about the most likely interpretations of the input image (gist) in recognition-related regions within the temporal cortex. Combining this top-down "initial guess" with bottom-up systematic analysis facilitates recognition by substantially limiting the number of object representations that need to be considered. This idea is strongly linked to the limited "bandwidth" of the system as described by Rensink [2000], i.e., only one object can be attended at any time.

5.3 Partial and global saliency maps and face recognition

As mentioned before, an important part of the model is based on responses of end-stopped cells in V1. In this section we show how FoA (keypoint-based saliency maps) influence object

recognition (in this case faces).

Again, in building saliency maps we assume that detected keypoints are summed over all scales, which is a retinotopic (neighborhood-preserving) projection by grouping cells. Keypoints which are stable over many scales will result in large and distinct peaks; see Section 3.7. In other words, since keypoints are related to local image complexity, such a saliency map codes local complexity. In addition, different saliency maps can be created at different scale intervals, from fine to coarse scales, indicating interesting points at those scales with associated Regions-of-Interest (RoIs). Such information is very important in steering our eyes, because fixation points in complex regions (eyes, nose, mouth) are much more important than those in more homogeneous regions (forehead, cheeks). Figure 5.1 (top row) shows detected keypoints at fine (left), medium and coarse (right) scales.

Regions surrounding the peaks can be created by assuming that each keypoint has a certain RoI, the size of which is coupled to the scale (size) of the underlying simple and complex cells [Rodrigues and du Buf, 2005c]. A global saliency map obtained by summing over all scales codes image complexity at all scales. Likewise, partial saliency maps can be constructed that code complexity at specific scale intervals. Figure 5.1 (middle row) shows, left to right, four partial saliency maps from fine to coarse scales, obtained by assuming 8 neighboring scales around the center scales (shown in the top row), plus the global saliency map, for $g = 1.0$; see Section 3.3 for the g inhibition parameter. The bottom row shows the same in the case that $g = 0.25$. Summarizing, less tangential and radial inhibition leads to the more "complete" saliency maps shown in the bottom row, and in the global map we can see, more or less, the structure of the input image (Fig. 5.1 bottom-right), in particular the regions around the eyes, nose, mouth etc. Actually, these regions correspond to the regions that contain many fixation points as measured by tracking the eyes of a person who is looking at a face [Pomplun et al., 1997]. Below, in face recognition, these regions will be used to "gate" detected lines and edges, and we will experiment with different options.

As was explained in Chapter 4, the multi-scale line/edge representation will be exploited, because this characterizes facial features, and saliency maps will be used for Focus-of-Attention, i.e., to "gate" detected lines and edges in associated Regions-of-Interest. This resembles the bottom-up data streams in the where (FoA) and what (lines/edges) subsystems.

As in Section 4.7, and in order to be able to compare results, we use 8 primary scales $\lambda_1 = \{4, 8, 12, 16, 20, 24, 28, 32\}$ with $\Delta\lambda_1 = 4$. Each primary scale is supplemented by 8 secondary scales with $\Delta\lambda_2 = 0.125$, such that, for example, $\lambda_{2,\lambda_1=4} = \{4.125, 4.250, ..., 5.000\}$. These secondary scales are used for stabilization and the construction of partial saliency maps. In our recognition model (Section 4.7) step (B) is going to be subdivided into steps (B.1) and (B.2):

(B.1) Construction of four symbolic representation maps. At each primary scale, stable events (positions) are expanded by Gaussian profiles (lines) and bipolar, Gaussian-truncated errorfunction profiles (edges), the sizes of which being coupled to the scale of the underlying simple and complex cells; see Fig. 4.7. Responses of complex cells are used to determine the amplitudes of the profiles. As a result, each face image is represented by 4 maps at each of the 8 primary scales; the same as in Chapter 4.

(B.2) Construction of saliency maps. Two types of maps are created: (1) one global saliency map (GSM), combining all keypoints at all 72 scales, and (2) eight partial saliency maps (PSM) combining the primary and their secondary scales; see Fig. 5.1. The GSM can be used to gate all representation maps (step B.1) at all scales, whereas PSMs are used to gate the maps at the same primary scales.

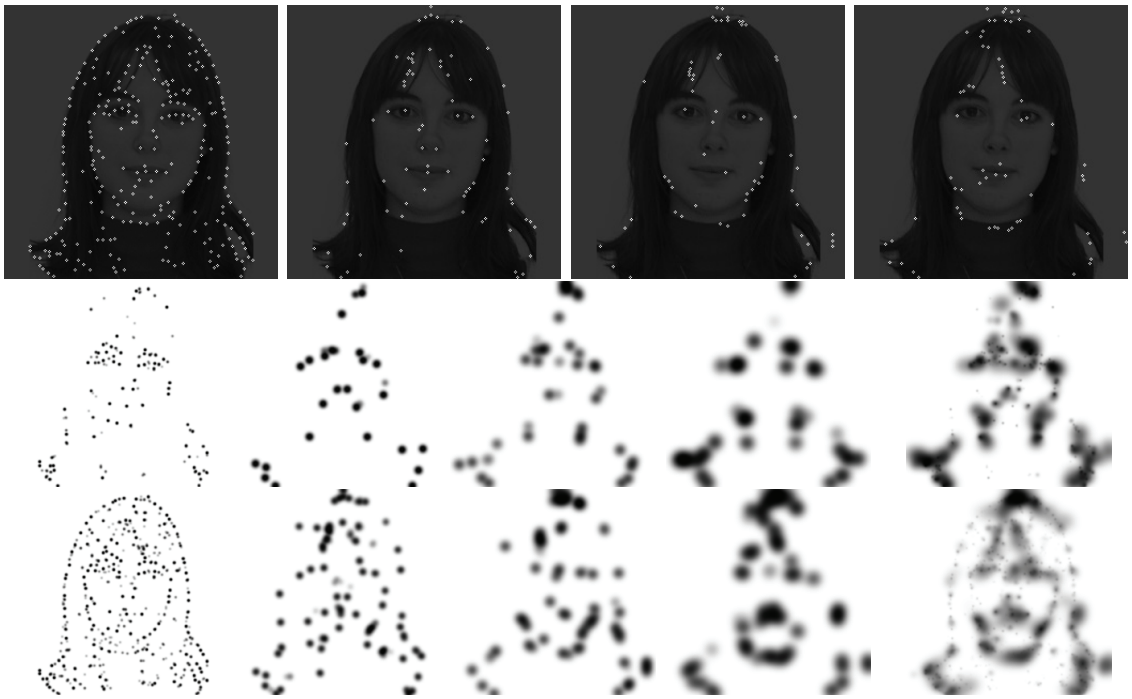

Figure 5.1: Top row: keypoints detected at four scales. Middle row: four partial saliency maps and the global (rightmost) map using $g = 1$. Bottom row: the same using $g = 0.25$. For explanation see text.

We only simulate event selections by employing global and partial saliency maps obtained with different inhibition parameters g in keypoint detection: the higher g, the more precise detection will be and, consequently, less line/edge events will be available for categorization and/or recognition. The reason for this choice is the fact that coarse-scale information from area V1 propagates to IT (inferior temporal) cortex first, for a first but very coarse categorization, after which information at increasingly finer scales arrive at IT [Bar, 2003]. This means that the system starts with coarse line/edge representations and partial saliency maps at those scales, to be simulated with $g = 1.0$, then refines the search with partial maps with $g = 0.25$, and can finish recognition with a fine-scale and/or global saliency map, also with $g = 0.25$. Figure 5.2 illustrates information available for recognition.

5.3.1 Results

All recognition tests involved the entire set of 100 images from the Psychological Image Collection at Stirling University (UK); the same data set presented in Section 4.7. Results will also be specified in terms of the white/black subsets in order to analyze the influence of the different backgrounds, and with two different types of templates: (1) only one frontal view, and (2) two views, frontal and 3/4. Robustness with respect to occlusions was also tested. In all cases, template images were selected randomly. Results presented in Chapter 4 and here were obtained by using the same templates. To simplify comparison we repeat in the tables the results of Section 4.7 on lines marked "all events."

Table 5.1 presents the results obtained by using partial saliency maps ("PSM") with $g = 1.0$ and 0.25, global saliency maps ("GSM") with $g = 0.25$, and all detected events, i.e., without applying saliency maps ("all events"). Results concern a mix of all images ("all")

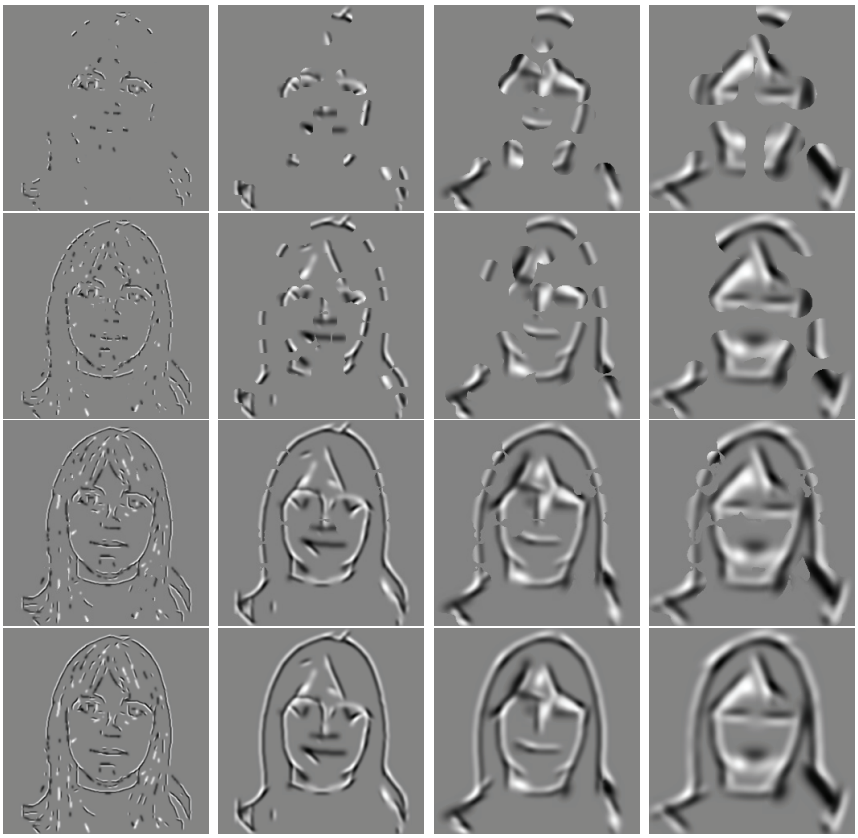

Figure 5.2: Combined event representations gated by different saliency maps, from top to bottom: PSM with $g = 1.0$, PSM with $g = 0.25$, GSM with $g = 0.25$, and all detected events.

and separated dark ("black") and light ("white") backgrounds. The column "scales" lists the percentage of correct scales that lead to correct recognition in the case of "all" and scheme 1. The last column ("base line") lists the number of all 100 images that have been recognized with absolute certainty.

Using no saliency maps, i.e., using all detected events, yields best results, which was expected. Also expected was the increasing rates in the four lines, because the use of different saliency maps implies more or less information available for recognition, see Fig. 5.2. Best results were obtained when using two templates with frontal and 3/4 views, using all events resulted in 96%, whereas 81 was the base line with absolute certainty. The increasing base line (67, 69, 79, 81) implies that a system simulating dynamic processing may have an easier task: after the first step, already 67 of 100 images have been identified using early and therefore limited information, and only the remaining 33 must be scrutinized in a second step. The timeline of the results in the table would correspond to an arrow in Fig. 5.2 which points from top to bottom, but in reality it would be one which points from top-right to bottom-left.

The partial occlusions tested were the same as in Section 4.7. Table 5.2 presents results in terms of "rate (base line)," which must be compared with the bottom part of Table 5.1, i.e., the first and last columns.

In the case of PSM with $g = 0.25$, the base line of 69 (Tab. 5.1) drops to 54 in the case of the most severe occlusion type 4 (eyes, nose and mouth). As mentioned in Section 4.7,

templates	only frontal view							
recogn. scheme	2	2	2	1	1	1	1	base
images	all	black	white	all	black	white	scales	line
PSM $g = 1.0$	86.0	83.0	89.4	87.0	83.0	91.5	82.8	61
PSM $g = 0.25$	89.0	86.8	91.5	88.0	86.8	89.4	83.0	63
GSM $g = 0.25$	90.0	90.6	89.4	89.0	90.6	89.4	85.9	70
all events	91.0	90.6	91.5	89.0	86.8	91.5	85.5	71
templates	frontal plus 3/4 view							
PSM $g = 1.0$	94.0	98.1	89.4	94.0	96.2	91.5	88.6	67
PSM $g = 0.25$	95.0	98.1	91.5	93.0	96.2	89.4	89.1	69
GSM $g = 0.25$	95.0	100.0	89.4	95.0	100.0	89.4	92.5	79
all events	96.0	100.0	91.5	96.0	100.0	91.5	91.8	81

Table 5.1: Results obtained without occlusions.

	frontal plus 3/4 views; recogn. scheme 2				
occlusion type	1	2	3	4	5
PSM $g = 0.25$	95.0 (68)	96.0 (66)	92.0 (62)	83.0 (54)	93.0 (66)
all events	96.0 (80)	95.0 (74)	96.0 (67)	93.0 (64)	97.0 (75)

Table 5.2: Results obtained with partial occlusions.

in the "all events" case, instead of 81 only 64 was obtained. But this is the base line: still 54 or 64 of all 100 images are classified with absolute certainty. The maximum rate for this occlusion (all events, 93%) is very close to the maximum without occlusion (Tab. 5.1, 96%), and slightly worse if compared to the other occlusions. This shows that the multi-scale representation, in particular the shape of the head and hair at the coarser scales, is very robust and contributes most in the recognition. The reason for this can be seen in Fig. 4.5: the stable and "sketchy" information without too much detail at coarse scales.

The line/edge representation at coarser scales provides a stable abstraction of facial features (Figs 4.5 and 4.7). This explains, at least partly, the generalization that allows to classify faces with noise, glasses, relatively normal expressions and views (Fig. 4.14). The main problems were: (1) a change of hairstyle and extreme expression (Fig. 4.7 top-right with long hair was recognized, but not Fig. 4.14 bottom-right with short hair and big smile); and (2) insufficient image normalization; in Fig. 4.14, the fourth and fifth images on the 3rd row and the third image on the 4th row were problematic. These were three of only four images which were not recognized; hence, the overall recognition rate of 96 in 100). However, the first image on the 4th row was recognized!

The problem of insufficient normalization can be solved because faces can be detected by grouping keypoints at eyes, nose and mouth [Rodrigues and du Buf, 2005c]. By using detected keypoints at the two eyes and mouth corners, images can be morphed such that the central part of a face is normalized in terms of size and position. This procedure can also guarantee that templates in memory are really representative. However, similar solutions for hairlines and non-frontal views must be developed; see also Valentin et al. [1997]. As for now, correct face recognition in the case of a drastic change of hairstyle and expression remains a research topic. Keeping in mind that face normalization (invariance) is a special case, in the next section we present an invariant object recognition scheme.

5.4 Invariant object recognition

In our experiments we used the ETH-80 database [Leibe and Schiele, 2003] in which all images are cropped such that they contain only one object, centered in the image, plus a 20% border area. The views of all objects are also normalized, e.g. all animals with the head to the left (see Fig. 5.7). We selected 10 different images in each of 8 groups (dogs, horses, cows, apples, pears, tomatoes, cups and cars). The selected images were used at three levels: four types of objects (animals, fruits, cars, cups) for *pre-categorization*. Two of those were subdivided into three types (animals: horses, cows, dogs; fruits: tomatoes, pears, apples) for *categorization*; the same as in Section 4.6.2. Final *recognition* concerns the identification of each individual object (e.g. horse number 3) within the corresponding group (e.g. horses). A selection of all normalized objects was used to create a set of modified objects, by applying translations, rotations, re-scalings and even deformations; see the Results section. In what follows it is important to keep in mind that templates in memory are always based on original, normalized objects in the database, against which modified objects will be tested. Figure 5.7 shows in neighboring left-right columns the normalized and examples of modified objects, except for the 3rd image from left at the top row which is a modified image (the modified objects shown on the bottom row were not correctly categorized or recognized).

As explained above, an SM indicates the most important positions to be analyzed, because it is constructed on the basis of the multi-scale keypoint representation where keypoints code local image complexity on the basis of end-stopped cells. At positions where keypoints are stable over many scales, this summation map will show distinct peaks: at centers of objects (coarse scales), at important sub-structures (medium scales) and at contour landmarks (fine scales). The height of the peaks provides information about their relative importance. Such saliency maps are crucial for Focus-of-Attention and are part of the data stream which is data-driven and bottom-up. This data stream can be combined with top-down processing from IT cortex in order to actively probe the presence of objects in the visual field [Deco and Rolls, 2004]. In our own experiments we assume that SMs are also part of object and group templates in memory, and that these are used to project representations of input objects onto representations of templates by means of dynamic routing.

We explored the following scenario: each object template consists (partly) of significant peaks of the saliency map obtained by non-maximum suppression and thresholding. A grouping cell, with its dendritic field (DF) in the SM, is positioned at the central keypoint (CKP) that represents the entire object/template at very coarse scales (Fig. 5.3a); such central keypoints at coarse scales are always located at or close to the object's centroid; see Figs 3.4 and 3.6 or Rodrigues and du Buf [2006d]. The grouping cell triggers the object-template matching process, the invariant method consisting of steps a to f:

(a) Central keypoints at very coarse scales of an input object and a template are made to coincide (Fig. 5.3b; T stands for translation). This can be seen as a translation of all keypoints (SM peaks) of the object to the ones of the template (or vice versa), but in reality there is no translation: only a dynamic routing by a hierarchy of grouping cells with DFs in intermediate neural layers such that the response of the central grouping cell of the template is maximum.

(b) The same routing principle of step (a) is applied to the two most significant SM peaks (from all scales), one of the input object and one of the template. Again, grouping cells at those peaks and with DFs in the intermediate layers serve to link the peaks by dynamic routing, but this time for compensating rotation and size (Fig. 5.3b; R and S). The resulting routing (translation, rotation and size projection) is then applied to all significant

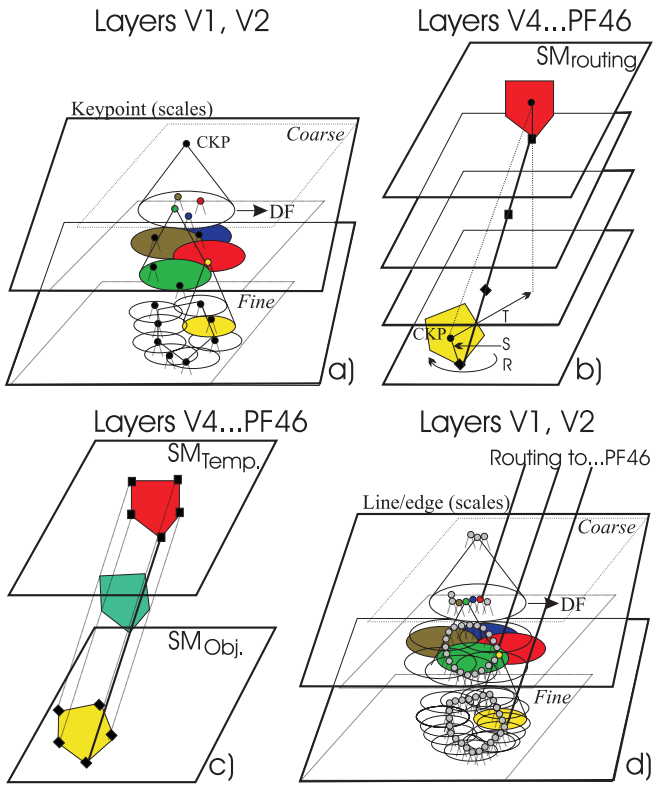

Figure 5.3: Dynamic routing scheme: the principle.

peaks (Fig. 5.3c) because they belong to a single object/template.

Figure 5.4 illustrates the above two steps. At top-left, central keypoints of template and input object excitate cells at intermediate levels through axonic fields, spreading activations in separate top-down (solid circle) and bottom-up (open circle) trees. This enables grouping cells at all levels to combine the top-down and bottom-up activations (shown in red). Once this first routing has been established, it can be propagated laterally to routing cells at all levels, but only one way, for example top-down. Using similar cell structures, most significant peaks in SMs are used to refine the routing (Fig. 5.4 top-right in green and bottom-left in blue). In the Discussion this process is also called "anchoring."

(c) All other significant SM peaks of the input object and of the template are tested in order to check whether sufficient coinciding pairs exist for a match. To this end another hierarchy of grouping cells is used: from many local ones with a relatively small DF to cover small differences in position due to object deformations etc., to one global one with a DF that covers the entire object/template. Instead of only summing activities in the DFs, these grouping cells can be inhibited if one input (peak amplitude of object, say) is less than half of the other input (in this case of the template).

(d) If the global grouping of corresponding pairs of significant peaks is above a threshold (e.g. half of the maximum peak in the SM), the invariant match is positive. If not, this does not automatically mean that input object and template are different: the dynamic routing established in step (b) may be wrong. Steps (b-c) are then repeated by inhibiting the most significant peak of the object and selecting the next biggest peak.

(e) If no global match can be achieved, this means that the input object does not correspond to the template or that the view of the object (deformation, rotation or size) is not represented by the template. In this case the same processing is applied using all

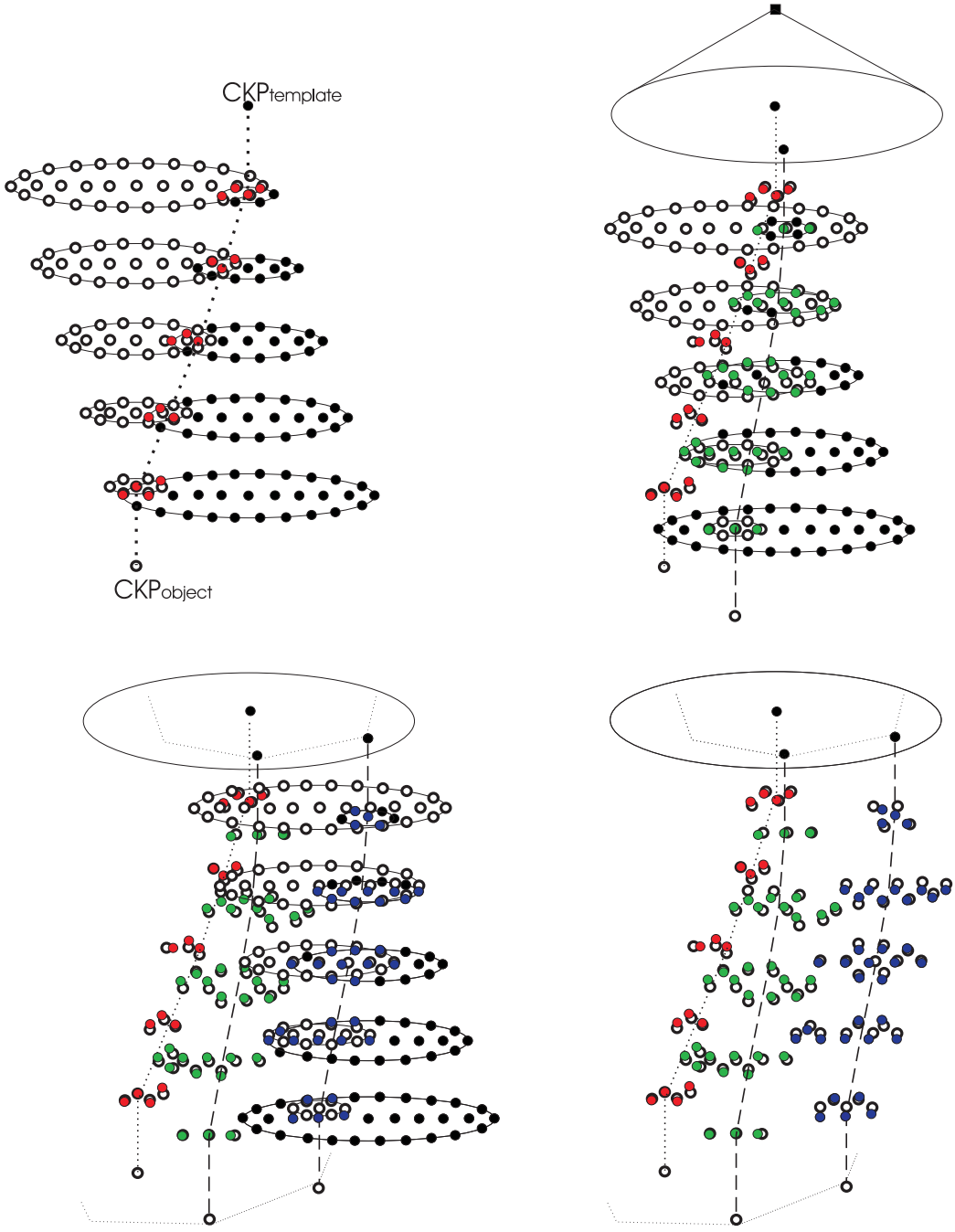

Figure 5.4: Dynamic routing scheme: spreading and grouping.

other templates in memory until the ones are found which could match. Although this process is simulated sequentially in our experiments, in reality this could be done in parallel by means of associative memory [Rehn and Sommer, 2006].

(f) Until here, only saliency maps were used to find possibly matching templates, but mainly for dynamic routing which virtually "superimposes" the input object and templates. In this step the dynamic routing of keypoints is also applied to the multi-scale line/edge representations in order to check whether object and a template really correspond (Fig. 5.3d). Again, this is done by many grouping cells with small DFs (local correlation of line/edge events) and one with a big DF (global object/template correlation); see Rodrigues and

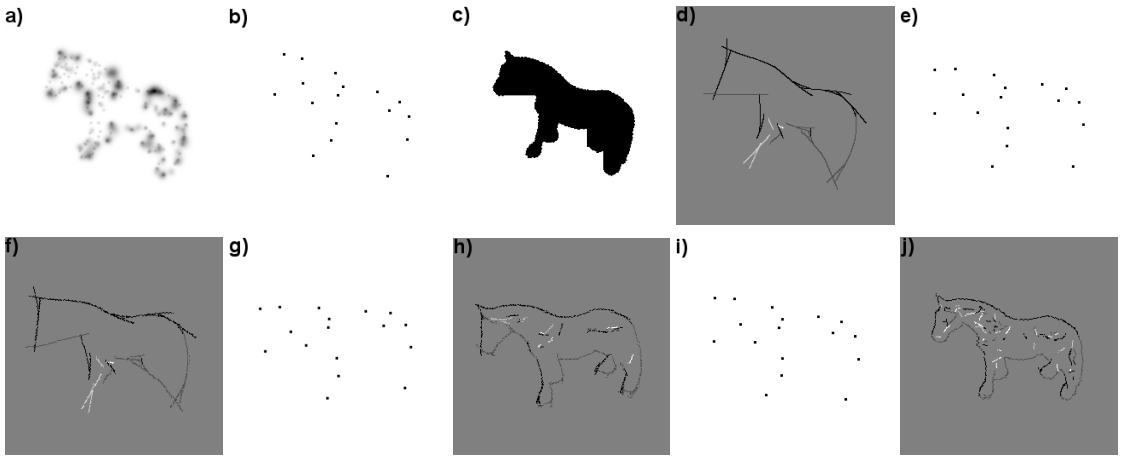

Figure 5.5: (a) Saliency map of modified horse8, (b) SM peaks, (c) segregated image and (d) line/edge coding of segregated image at $\lambda = 24$. (e-f) SM peaks and line/edge map of normalized horse8 (after the dynamic routing) in pre-categorization. (g-h) The same with line/edge map at $\lambda = 8$ in categorization. (i-j) The same with line/edge map at $\lambda = 4$ in final recognition. Input object and matching object (used only in recognition) are shown in Fig. 5.7 (marked by a black and white corner triangle).

du Buf [2006a]. The use of small DFs can be seen as a relaxation: two edges of object and template count for a match if they are at the same position but also if they are very close to each other. The size of the DFs is coupled to the size of underlying complex cells [Bar et al., 2006].

The template information used in step (f) depends on the level of categorization. In the case of the first, coarse, pre-categorization **(f.1)**, only line/edge events (Fig. 5.5d) at 3 coarse scales of the segregated, binary object (Fig. 5.5c) are used, because (a) segregation must be done before or at an early stage of categorization and (b) coarse-scale information propagates first from V1 to higher cortical areas. Global groupings of lines and edges are compared over all possibly matching templates, scale by scale, and then summed over the 3 scales, and the template with the maximum sum is selected (winner-takes-all; Fig. 5.5f shows a projected and matching line/edge map after dynamic routing. In the case of the subsequent finer categorization **(f.2)**, the process is similar, but now we use line/edge events at all 8 scales obtained from the object itself instead of from the binary segregation. Figure 5.5g and h show projected peaks and line/edge map used in categorization. Final recognition **(f.3)** differs from categorization (f.2) in that line and edge events are treated separately: object lines must match template lines and edges must match edges. This involves three additional layers of grouping cells, two for local co-occurrences of lines and edges and one global. Figure 5.5i and j show projected peaks and the line/edge map used in recognition. See Rodrigues and du Buf [2006a,b] or Sections 4.6.2 and 4.7 for complete explanations of the matching processes in the case of only using normalized object views.

5.4.1 The creation of group templates

Good object templates in memory—both line/edge maps and saliency maps—are fundamental for obtaining good recognition results, but at the same time group templates must be generic enough to represent only one category for (pre-)categorization. Different line/edge

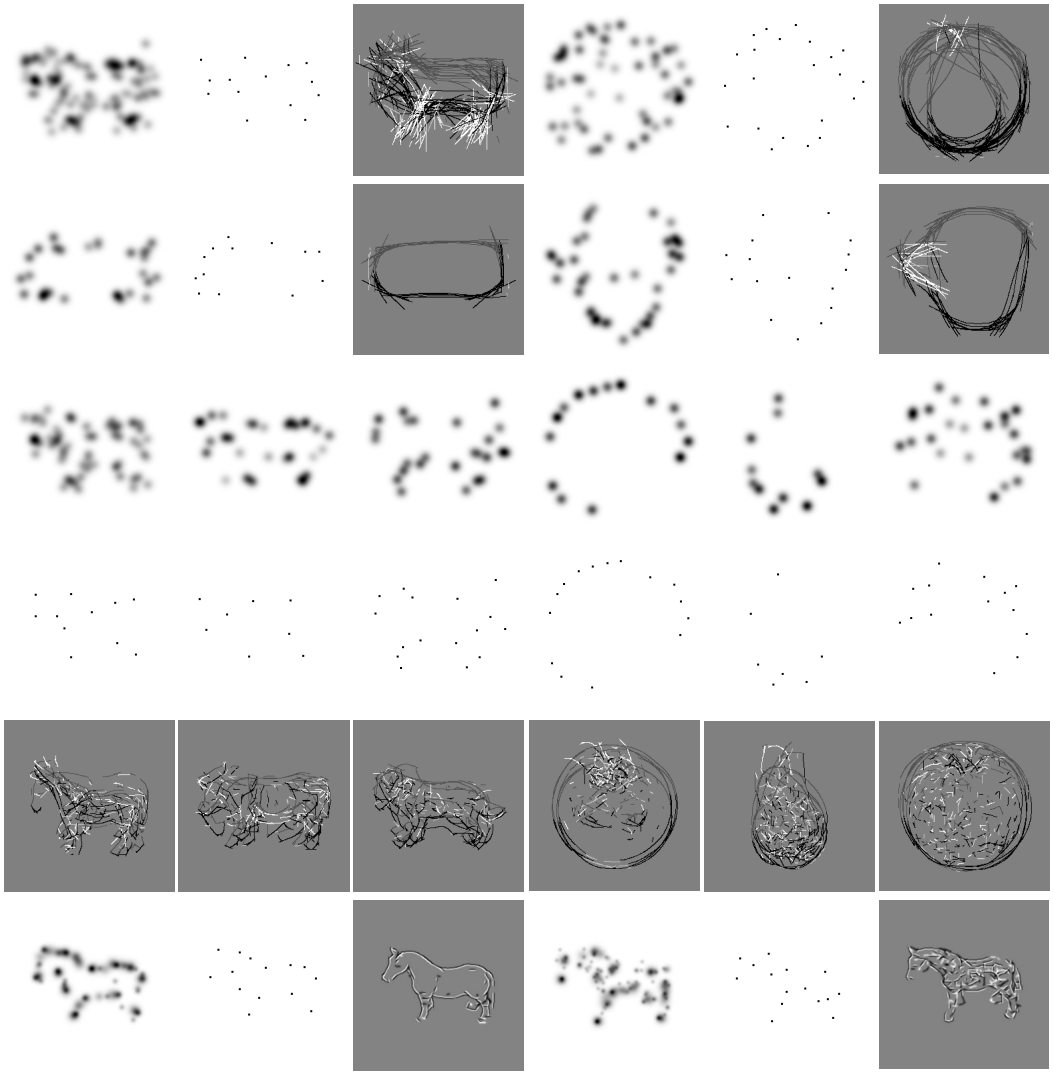

Figure 5.6: Top two rows: group templates for pre-categorization (animal, fruit, car and cup). Middle three rows: the same for categorization (horse, cow, dog, tomato, pear and apple). Bottom line: Templates for recognition, examples of two different horses.

templates with increasing detail are used in pre-categorization, categorization and final object recognition, but also different saliency maps in the dynamic routing for invariance. In order to create the group templates for pre-categorization (animal, fruit, car, cup), the saliency maps of the normalized objects in the database were selected randomly: for each group we summed half of the SMs, i.e., 5 SMs in the case of the 10 cups and cars, and 15 SMs in the case of animal (or fruit) with 10 images each of dogs, horses and cows (or apples, pears and tomatoes). The resulting peaks in the summed SMs were used for the dynamic routing of the SM-peaks of *modified* input objects. In the case of the second categorization of animals and fruits, the same procedure was followed: 5 randomly selected SMs of horses, dogs and cows, and of apples, pears and tomatoes. At present, only normalized objects are used for constructing SMs for the group templates. In contrast, different line/edge maps, as explained in step f in the previous section, are already being used in pre-categorization, categorization and final object recognition, essentially applying the same procedure: random selection of images and logical combination of event maps (for details see Rodrigues

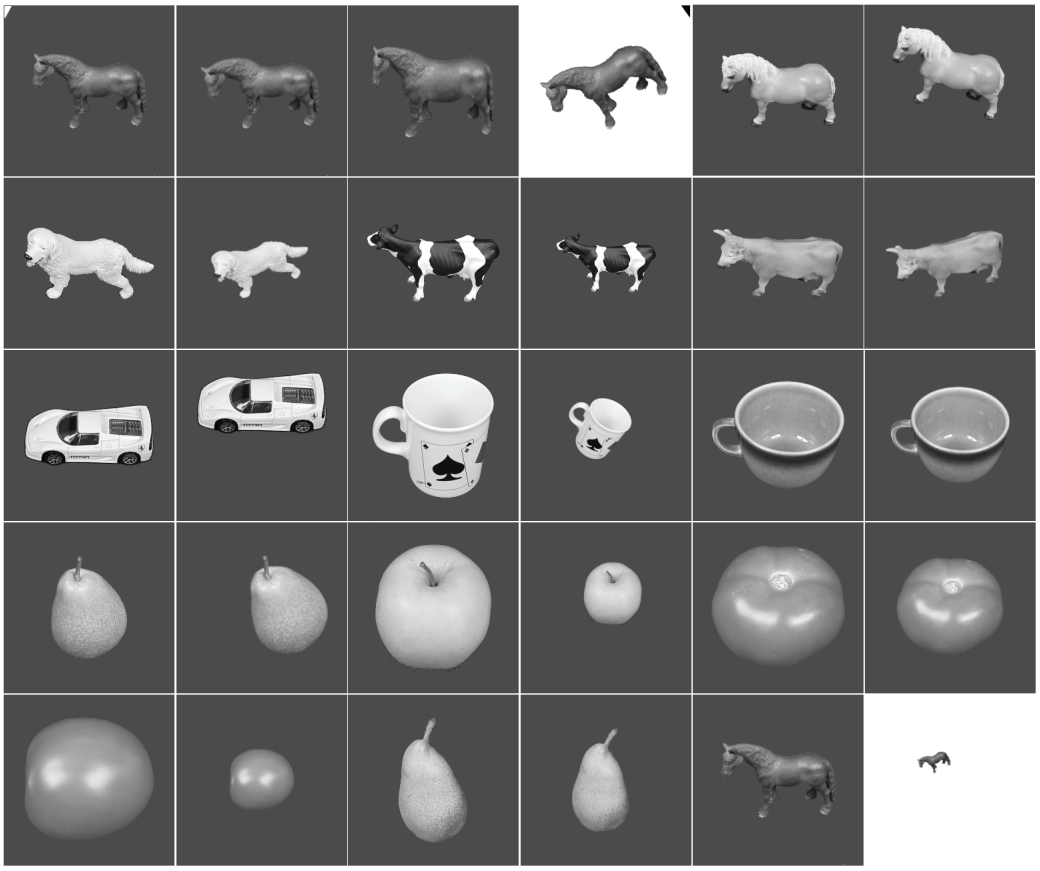

Figure 5.7: Examples of objects used for categorization and recognition.

and du Buf [2006a] or Sections 4.6 and 4.7). In the near future, the entire process will be implemented and tested in a completely dynamic way, including the use of increasingly more detail in stabilizing and refining the dynamic routing.

Rows 1 and 2 of Fig. 5.6 show the templates used in pre-categorization with, from left to right, saliency map, significant peaks and line/edge map at $\lambda = 32$ (one of three scales used) for the animal, fruit, car and cup groups. Rows 3 to 5 show the same for categorization ($\lambda = 8$ for the line/edge maps, one of eight scales used) with, from left to right: horse, cow, dog, tomato, pear and apple group templates. The bottom row shows two individual object templates used in recognition, i.e., two examples of the 10 different horses, with the line/edge map at $\lambda = 4$ (one of the eight scales used). Summarizing, Fig. 5.6 shows the template information in memory on the basis of *normalized* objects against which *modified* objects are matched.

5.4.2 Results

In order to test invariant processing, a set of modified input images was created by manipulations like translations, rotations and zooms, including deformations (e.g. the head of a horse moving up or down relative to the body). We created 64 additional input images of the most distinct objects: 20 manipulated horse images (horses were used as a special test case for recognition); 6 dogs, 6 cows, 4 tomatoes, 4 pears and 4 apples; plus 10 cars and 10 cups. Typical images are shown in Fig. 5.7: the top line shows the same horse normalized (marked

by white triangle) and with the head more down, bigger, and rotated and scaled against a white background. The use of this extended database allows to compare our results with invariant processing to previous results obtained with only normalized objects (Section 4.6): mean error (standard deviation) of 3.0(1.0)% in the case of pre-categorization and 9.3(2.1)% in the case of categorization. These results were also obtained by using 8 scales equally spaced on $\lambda \in [4, 32]$.

Results obtained with the 64 modified images were quite good: pre-categorization (animal, fruit, car, cup) failed in 12 cases. Of the remaining 52 images, categorization (animal: horse, cow, dog; fruit: tomato, pear, apple) failed in 8 cases. Recognition failed for 4 of the 44 remaining images. Analyzing each of the three levels separately and considering all objects available at each level, we may say that the error rate in pre-categorization is 18.7%, in categorization it is 21.9%, and in recognition 14.8% was achieved. This is an average of 18.5% at the three processing levels, but the overall error rate of the entire system, from 64 input images to 24 not-correctly recognized objects, is 37.5%. However, these numbers are not definitive because they concern a first test of the concept and many errors can be explained. For example, extreme size variations (see below) are expected to cause problems, in subsequent experiments the maximum variations can be determined, and if extreme variations are limited the numbers will improve.

As for our previous results obtained with normalized objects, Section 4.6, categorization errors occurred mainly for apples and tomatoes, which can be explained by the fact that the shapes are very similar and no color information has been used. In pre-categorization there appeared an increased error rate of fruits which were categorized as cups. This mainly concerned pears and can be explained by the tapered-elliptical shape in combination with size variations, such that keypoints and line/edge events of input pears can coincide with those of the cups-group template (see Fig. 5.6 top-right). As expected, especially in the case of recognition, problems occurred with extreme size variations. The scales used ($\lambda \in [4, 32]$) are related to the size of the objects and the level of detail that can be represented. Figure 5.7 (middle three images in the fourth column) shows the smallest objects that could be dealt with by using these scales. The image at bottom-right proved too extreme.

It should be emphasized that the method can be applied to images that contain multiple objects. Although our visual system has a limited "bandwidth" and can test only one object at any time [Rensink, 2000], this problem can be solved by sequential processing of all detected and segregated objects; see [Rodrigues and du Buf, 2006d]. However, if object segregation and recognition are coupled processes, we are left with a typical chicken-or-egg problem, unless the process is controlled by e.g. the gist system (see Discussion). Finally, it should be mentioned that dynamic routing of keypoints (significant peaks in saliency maps) and line/edge events in intermediate neural layers has consequences for the minimum number of canonical object views in memory, i.e., the number of templates. If a horse template has the head to the left and legs down, but an input horse has been rotated (2D) by 180 degrees such that the head is to the right and the legs up, dynamic routing will not be possible because there will be a crossing point in the routing at some level. In this case a separate template is necessary. In addition, recognition in the case of 3D rotation may require more templates because of asymmetrical patterns of a horse's fell. Extensive experiments with many more object views are required to determine the minimum number of templates.

5.5 Integrating the architecture

The processing scheme presented above is based on the combination of multi-scale features derived from five cell types: even and odd simple cells, complex cells, and single and double end-stopped cells, assuming additional line, edge, keypoint and saliency cells in the corresponding maps plus many grouping and gating cells. For a better understanding, the architecture shown in Fig. 5.8 is organized in a "features and blocks" fashion, where the different blocks and image features necessary to go from object detection to recognition are related by arrows.

The first task is to get the gist of the scene by a rapid but global classification [Oliva and Torralba, 2006]. After this all the objects can be analyzed, but sequentially, i.e., only one object at any time [Rensink, 2000]. Individual objects are analyzed in a multi-level recognition process [Grill-Spector and Kanwisher, 2005], and interesting positions to be analyzed after the gist stage are stored in a "waiting list" (normally, this is modeled by sequential processing of most-to-less-important peaks in a saliency map, simulating eye movements and fixation points, with inhibition of returns to already analyzed positions; see [Walther et al., 2002; Prime and Ward, 2006]).

Objects can be categorized or recognized at different levels, and some objects do need several processing levels before recognition is achieved. For example, in the case of a horse called Ted recognition can be achieved after three levels: animal, horse, Ted. However, this is a very rigid scheme in which all horses need to go through all levels. If Ted's fell is very characteristic, and no other known object, animal or car etc., displays a similar pattern, Ted could be recognized instantaneously by using other information channels, for example devoted to color and/or texture. But such channels are not yet implemented and our model is restricted to multi-scale line/edge and keypoint representations. Nevertheless, in our model an object can also be recognized at an early level, if a measure for correspondence—a match with one template in memory—is much bigger than a threshold level and correspondence measures of all other templates are much smaller than the threshold. This happens when learning new objects, for example in early childhood when seeing an object for the first time, then seeing it repeatedly, thereby confirming the object's template in memory, until seeing similar objects and constructing a group template. After this, the object is first categorized at the first level and may be recognized at the second level. This process can be seen as a decision tree—and the construction of the tree—in which finer and more specific object details are used by means of increasingly finer scale representations. However, instead of having a straightforward and bottom-up or data-driven decision tree, feedback loops at each level and between levels are necessary to adjust or correct the data flow when more or other information is necessary to characterize an object, and top-down feedback can be controlled by higher areas of the cortex. Such feedback may influence all lower layers: (a) with a spatial focus (FoA) by means of gating parts of saliency maps, and (b) controlling the number of required scales at a specific time (adaptive coarse-to-fine-scale processing instead of applying a rigid timing).

Figure 5.8 shows the generalized architecture, where each block represents the type of feature involved (and scales), as well as the processing done at the different stages. The blocks are displayed in a sequential way with early processing at the top and later processing toward the bottom. Only three levels are shown (1, 2 and n), but n is variable. At each level, three templates are shown (A, B and N), but N is variable and a function of the level. Features are indicated by SM (saliency map), LE (line-edge code) and LErepr. (symbolic line/edge representation), the latter two with an indication of the scales used (All scales or

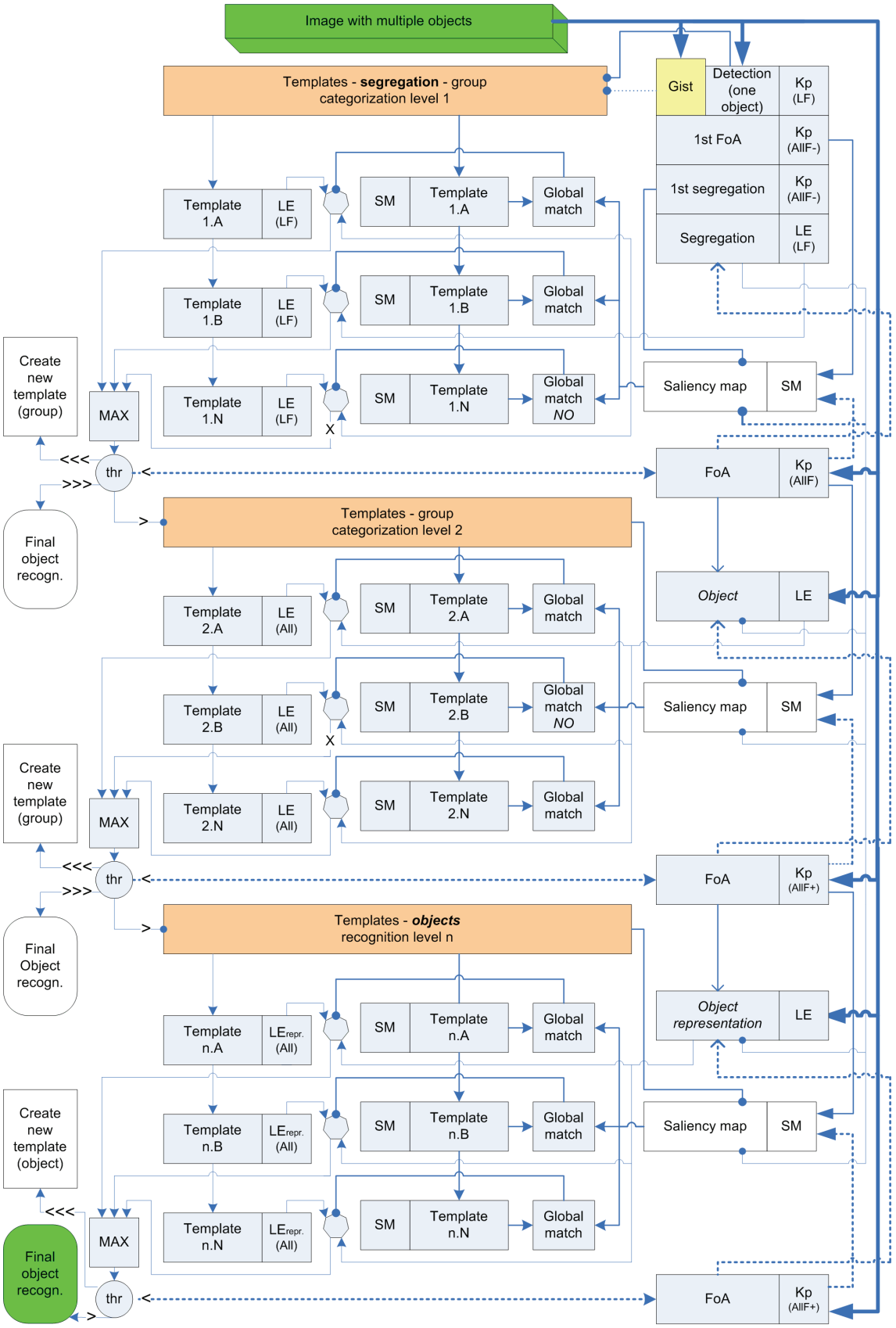

Figure 5.8: Generalized architecture: blocks, features and information flows.

LF meaning coarse scales only). The arrows show the information flow, the circles indicate activations, and dashed arrows represent feedback loops. If a template cannot reach a global match (NO), its output will be blocked (X) and cannot reach the MAX block; this is done to prevent the system from selecting some template when no template can match. Blocks marked "thr" perform thresholding, with four options: a very low value (<<<) implies the creation of a new template; a very high value (>>>) means final object recognition; if the value is not very much lower than the threshold (<), which means that more information is required to select the correct template, a feedback loop is activated (to the rightmost column of blocks, via FoA, in order to select more line/edge scales); if the value is not very much higher than the threshold (>), a specific template has been selected and this (group) template activates (selects) related (group) templates at the next level.

The heptagonal symbols between the LE and SM blocks of all templates represent comparisons (local and global correlations or matchings) between input and template features: line/edge events (LE) at categorization levels or their symbolic representations (LErepr.) at the final recognition level. A comparison is only activated when a global match occurs, and after the dynamic routing of events as explained before. In the rightmost column of blocks, the following abbreviations are used: LF refers to the coarsest scales, AllF– to many scales (coarse, medium and fine) but in octave intervals, AllF to more scales with sub-octave intervals, and AllF+ to the maximum number of scales with the smallest intervals. Instead of using only four selections, the number of scales is dynamic, i.e., more scales will be selected and used until the information provided by new scales becomes redundant.

When the objective is not to analyze a scene as a whole (as presented above), another usual task is to look for one specific object, like a person or a coffee cup. This task is much faster and easier, because the "what question" has already been solved and categorization levels might be skipped. The system could start by activating the recognition templates, i.e., only one object but represented by several views, and coarse scales. Finer scales can be added until detection occurs. The process also stops when no more information can be added and detection has not occurred, i.e., when the object is not present. This process is also controlled by higher cortical areas and by FoA, mainly in the where pathway. In the scheme shown in Fig. 5.8 information from the gist/segregation block at top-right passes directly to recognition level n at the bottom.

With respect to visual pathways, the where path is more related to the detection, segregation, FoA and object-representation blocks in the rightmost column in Fig. 5.8, whereas the what path consists of the other blocks, but it also includes the object-representation block. With respect to cortical areas involved, a strict attribution of the functional blocks to areas is still speculative, but a likely attribution is the following: simple, complex and end-stopped cells are located in area V1 [Olshausen and Field, 2005]. Line, edge and keypoint extractions also occur in V1, possibly also in V2. More complex object representations, at least of important objects like faces, are established in PP [Deco and Rolls, 2005] and IT [Zoccolan et al., 2005]. FoA processing may start at the LGN level (before the cortex!) but is most pronounced in V4 and beyond [Chelazzi et al., 2001], and figure-ground segregation may be achieved in V2, at least at the level of local occlusions [Qiu and von der Heydt, 2005]. Saliency maps may be present in MT [Born and Bradley, 2005] and PP [Deco and Rolls, 2004], and global matching using templates in IT. Templates of groups and objects are stored—or at least available—at PF46 [Miller, 2000].

5.6 Discussion

There are many properties of a real-world scene that can be defined independently of the objects. For instance, a forest scene with trees can be described in terms of the degree of roughness and homogeneity of its textural components. Oliva and Torralba [2006] conclude that there is converging evidence that natural scene recognition may not depend on recognizing objects, and that the gist does not need to be built on top of the processing of individual objects. Nevertheless, these processes are complementary. Initial gist can be the key for selecting the first group templates to start object recognition, but at some stage the objects should corroborate for the interpretation of the scene, and those objects must somehow be segregated. Any computational model of the cortical architecture should start with a model for getting the gist (forest scene), after which object recognition follows using segregated items, from generic information (trees) to more detailed information (tree type, leaf type). Only at the end of the entire process it may be possible to specify the gist, for example a Mediterranean forest with tall pine trees.

Gist has not yet been implemented in our architecture, because we think that segregation of complex environments like natural scenes and gist are very interconnected processes. These processes may be based on complementary information channels which address motion and disparity, but also surface properties instead of structural object shape: (a) color processing in the cytochrome oxidase blobs, which are embedded in the cortical hypercolumns with simple, complex and end-stopped cells for line, edge and keypoint coding, must attribute colors to homogeneous (line/edge-free but also textured) object surfaces, and (b) texture coding based on specific groupings of outputs of grating cells in the case of rather periodic patterns, or other but similar processes in the case of more stochastic patterns. As shown by du Buf [2007], groupings of outputs of grating cells is a straightforward, data-driven and therefore fast bottom-up process which provides a segmentation (segregation) of linear, rectangular and hexagonal textures. Therefore, a gist model, when seeing an image with blue and some white above green with a rather irregular pattern, may classify the scene, after sufficient training of course, as Mediterranean outdoor, thereby pre-selecting tree templates with a bias toward different pine trees (tall and more round etc.).

Not yet having a gist model, we simply assumed in our experiments that all group templates are available at the first categorization level, and that input objects are always seen against a homogeneous background (i.e., already segregated). At an early stage, only very coarse scales with big intervals are available, then medium scales with smaller intervals appear and finally the fine scales. The appearance and therefore the use of scales is directly related to all steps of the recognition process. The initial segregation starts with coarse scales, which provide a very diffuse object representation. This first segregation triggers a first categorization. When medium scales appear, and then fine scales, the segregation is improved and so is the categorization. The same occurs with the construction of the saliency map, first using keypoints detected at coarse scales and improving the map by adding keypoints detected at increasingly finer scales.

Invariance by neural routing from V1 via V2 to V4 etc. is based on the recurrent network layers used in the Deco and Rolls [2004] model, however with one big difference: instead of only using simple cells (Gabor model) we apply feature extractions and can use specific features to guide the routing. As a matter of fact, the routing can be seen as two vessels (input object and template) throwing anchors toward each other: the first, big anchor is the central object keypoint at very coarse scales and this is used to "position" the normalized template above the (shifted) input object. The second anchor is the most significant peak

of the saliency map, obtained by summing keypoints over many scales, and this is used to match rotation and size. Once "anchored together," the "ropes" are used to steer many more ropes that connect specific structures of the vessels, like bow, rail and stern, in order to check whether the structures are similar and the vessels are of the same type.

Our "anchoring" method is similar to the theory developed by Olshausen et al. [1993], suggesting that the position and size of the reference frame can be set by the position and size of the object in the scene, assuming that the scene is at least roughly segmented, and that the orientation of the reference frame can be estimated from relatively low-level cues. The computational advantage of such a system is obvious: only a few views of an object need to be stored for recognition under different viewing conditions. The disadvantage, of course, is that a scene containing multiple objects requires serial processing, the system only being able to attend one object at a time. The same happens in our model and that of Deco and Rolls: dynamic routing steers the information flow by adapting neural interconnections in V2 etc. for some time, until recognition has been achieved, after which the adapted steering can be released for the inspection of another object (or region around a fixation point). Psychophysical evidence suggests that the brain indeed employs such a sequential strategy [Rensink, 2000].

An interesting aspect of models is which features—and therefore which image representations—are being used. In our own model, explicit features are used: lines, edges and keypoints are detected on the basis of responses of simple, complex and end-stopped cells. The existence of other cells with very specific functions, like bar and grating cells, points at explicit feature extractions with increasing complexity at higher cortical areas [Rodrigues and du Buf, 2006d; du Buf, 2007]. The same idea, extended with increasing receptive field sizes, is supported by Deco and Rolls [2004], however without explicit feature extractions. By only using simple cells (Gabor model), higher features are represented implicitly: complex cells group outputs of simple cells; end-stopped cells group outputs of complex cells. Nevertheless, in principle—they did not test this—their model should also be able to achieve invariant object recognition by combining feedback effects of top-down attentional mechanisms in a hierarchically organized set of cortical areas with convergent forward connectivity, reciprocal feedback connections, and local intra-area competition. As a consequence, we may say that these two models are converging, but eventually the same will happen with other computational models [Olshausen et al., 1993; Hamker, 2005]).

Summarizing, the presented and tested architecture is a biologically plausible one. It is based on realistic multi-scale features which are extracted in the primary visual cortex. By employing feedback loops which are known to exist in abundance in the visual cortex, attention information based on keypoints and saliency maps is used to control the process. The entire process is composed of different categorization levels, recognition being the last level, with sequentially (but overlapping) coarse-to-fine-scale processing. Although not yet yielding perfect results, the architecture can deal with reasonable translations, rotations and scalings. In a next step, the maximally allowable transformations must be determined, which depend on the number of neural layers used in the routing, and this will provide information on how many views of objects must be stored in memory.

Chapter 6

Modeling brightness perception using line and edge representations

Abstract: A two-dimensional brightness model is presented. This model is a quantitative extension of the visual (re)construction principle, which is based on the multi-scale symbolic line/edge representation with an additional lowpass channel and nonlinear amplitude transfer functions. The brightness model is calibrated using psychophysical data, and it can predict most brightness effects and illusions, both the standard ones and variations: Mach bands, White's effect, Howe's and Anderson's patterns, Logvinenko's versions of Adelson's tile patterns, assimilation, simultaneous brightness contrast, grating induction and the Craik-O'Brien-Cornsweet illusion. Where possible, predictions are evaluated and discussed against real psychophysical data. The fact that the same line/edge representation can be used in the modeling of brightness perception and in object categorization and recognition suggests that both processes are correlated.

6.1 Introduction

The goal of visual psychophysics is to obtain a better insight into the process of visual perception by means of experiments and the modeling of measured data. The latter implies working in a quantitative way, but there is also the possibility of exploring models in a qualitative way, if the only aim is to construct a model that can account for a particular effect or a small group of related effects, for example some well-known illusions. In this chapter, we will focus on one aspect of visual psychophysics, namely the relation between the (physical) luminance and the (subjective) brightness of various spatial patterns.

The construction of a generally applicable brightness model, which can predict most if not all known brightness effects, is one of the most difficult aspects of visual psychophysics. This difficulty is caused by the fact that our visual system consists of many nonlinear subsystems called channels in parallel, and that our insight into this system is still far from complete. For example, we simply do not know at which level in the visual system the detection of spatial patterns at very low contrast takes place. Does it occur in the retina, at early cortical

layers, or even at a much higher level? In the case of detection, the contrast is so low that the pattern itself is not seen, only perhaps a small part or only "something" on the background. In the case of brightness perception, the image that we perceive, where is this image formed? Is it at one precise neural layer or is a big part of our brain involved? Perception is related to consciousness, and consciousness may mean our entire brain, which is a holistic view. In addition, brightness perception must somehow be linked to object recognition which involves syntax (lines, edges, textures) and semantics (meaning, like the functionality of a coffee mug).

The construction of brightness models is of paramount importance due to two main reasons: (a) models based on psychophysical data can be tested against these and model predictions, in particular inaccurate ones, may lead to a better insight into the processes involved (feedback leading to additional psychophysical experiments concerning unclear aspects of spatial interactions); and (b) practical applications such as image coding. A good brightness model can serve as the basis for codecs with high compression rates, because these may cause image deformations that are more natural and therefore more difficult to perceive if compared to standard codecs based on straightforward subband decomposition and quantization schemes [Ye et al., 2004]. In addition, a good brightness model can be used as a standard observer for assessing image quality, comparing the *perceived* input image with the *perceived* coded-decoded image [du Buf, 2001].

Although various brightness models have been published, exploring different possibilities, most models are restricted and can cope with just one effect like e.g. Mach bands or a few variations of a certain stimulus type, like White's effect with simultaneous brightness contrast or assimilation in grating patterns. The objective of this chapter is not to present a complete survey of existing brightness models. We will very briefly describe models which are more related to our own approach, with special focus on the ones that model real psychophysical data.

du Buf [1993] studied the responses of simple and complex cells to lines and edges, and proposed a detection operator on the basis of an abstraction of two simple cells, both centered at the same location. In [du Buf, 1994] he applied this analysis to luminance ramps and proposed the "syntactical reconstruction principle," where responding (active) cells are interpreted as Gaussian lines and errorfunction-shaped edges. He also showed explanations of Mach-band effects, e.g. the attenuation of Mach bands in the case of a bar located in the middle of a ramp edge between two luminance plateaus.

Also concerning Mach bands, Pessoa [Pessoa, 1996b,a] presented the results of a set of psychophysical experiments with and without adjacent stimuli placed near ramp edges, and explored a model based on the filling-in principle. Later, Neumann et al. [2001] demonstrated that filling-in from contrast estimates leads to a regularized solution of the computational problem posed by generating brightness representations from sparse estimates. They also proposed a new, improved version, namely confidence-based filling-in which generates even more robust brightness representations. Although the filling-in theory works quite well in many applications, it should be mentioned that there is no direct biological evidence of it (at least functional magnetic resonance imaging or fMRI) in early human visual cortex [Cornelissen et al., 2006]. Keil et al. [2006] presented a neural architecture model for explaining Mach bands. This model was designed and optimized for representing luminance gradients in real-world images, and it provides a novel approach to Mach bands with consistent predictions of real psychophysical data (but only related to Mach bands).

Moulden and Kingdom [Moulden and Kingdom, 1989; Kingdom and Moulden, 1991] investigated the properties of White's effect. They reported several experiments to reveal the effects of the height and weights of both flanking and coaxial bars on the brightness of the

grey bars. du Buf and Fischer [1995] extended the "syntactical reconstruction principle." They presented a one-dimensional (1D) brightness model based on the symbolic line/edge representation with an additional lowpass filter and nonlinear amplitude-transfer functions. The 1D model was shown to predict various effects such as White's one, simultaneous brightness contrast, the Craik-O'Brien-Cornsweet illusion and Mach bands.

Blakeslee and McCourt [1999] introduced an oriented, two-dimensional (2D) difference-of-Gaussian (ODOG) brightness model, where the filters are anisotropic and their outputs are pooled nonlinearly. They compared model predictions against a series of psychophysical experiments concerning White's effect, simultaneous brightness contrast and grating inductions. Later, in [Blakeslee and McCourt, 2004], they extended the multi-scale model to incorporate orientation selectivity of the filters and contrast normalization across channels. They showed several psychophysical data sets and related model predictions concerning White's effect, its shifted version, and "chalkboard" stimuli under different conditions. In [Blakeslee et al., 2005], the authors extended the data sets and predictions to more variations of White's effect, Howe's patterns and simultaneous brightness contrast.

Moulden and Kingdom [1990] studied the Craik-O'Brien-Cornsweet illusion which, for historical reasons, they called the Craik-Cornsweet-O'Brien (CCOB) illusion. Psychophysical experiments revealed how the CCOB effect was influenced by the width and amplitude of the brightness-inducing edges. They collected data using bar-shaped stimuli with sloping edges (gradients) on both in- and outsides and with positive and negative polarities. They also presented a model that assumes that a symbolic brightness description is generated separately by a number of differently-sized "second difference of a Gaussian" (2DOG) filters, and that the resulting brightness profile obtained by averaging across the separate descriptions. Schouten [1992] showed a two-dimensional luminance-brightness algorithm based on the following steps: the derivation of a multi-scale representation of the luminance distribution, the assemblage of the multi-scale signal into an illumination-insensitive version of the luminance distribution, and a local adjustment by means of a compressive brightness scale. The model could account for Craik-O'Brien-Cornsweet and some brightness-induction effects, but it did not predict Mach bands nor White's effect.

In the following sections a two-dimensional brightness model that can predict most known brightness effects is presented. The model is based on and a refinement of the earlier 1D model, so it also is an extension to two-dimensions of the one by du Buf and Fischer [1995]. This basis model was used because it was, to the best of our knowledge, the only one that already could explain Mach bands, brightness induction (assimilation and simultaneous brightness contrast) and the Craik-O'Brien-Cornsweet illusion. We show that the same image representation that was used for object recognition (see Chapters 4 and 5) can also be used in the modeling of brightness perception, i.e., image (re)construction. Finally, it will be shown that the 2D model can predict many more effects and not only the ones predicted by the 1D model. Section 6.2 explains the brightness model and simulation results are presented in Section 6.3. Section 6.4 deals with final concluding remarks.

6.2 Brightness model

In order to explain the brightness model, it is necessary to illustrate how our visual system can (re)construct, more or less, the input image. As already mentioned in Section 4.3, image reconstruction can be based on one lowpass filter plus a complete set of bandpass wavelet filters, such that the frequency domain is evenly covered. This concept is the basis of many

image coding schemes. It could also be used in the visual cortex, because simple cells in V1 are often modeled by complex Gabor wavelets, which are bandpass filters [Heitger et al., 1992], and lowpass information can be available through special retinal ganglion cells with photoreceptive dendrites [Berson, 2003] or through another "channel." Activities of all cells could be combined by summing them in one cell layer that would provide a reconstruction or brightness map. But then a big problem has been created: it is necessary to create *yet another observer* of this map in our brain.

The proposed solution is simple: instead of summing activities of all cells, we can assume that the visual system extracts lines and edges from responses of simple and complex cells, which is necessary for object recognition (see Sections 4.6 and 4.7), and that responding "line cells" and "edge cells" are interpreted symbolically. Responding line cells along a bar signal that there is a line with a certain position, orientation, amplitude and scale, the latter being interpreted by a Gaussian cross-profile with a size which is coupled to the scale of the underlying simple and complex cells. The same way a responding edge cell is interpreted, but with a bipolar, Gaussian-truncated errorfunction profile [du Buf, 1994].

The line and edge extraction method was explained in detail in Chapter 4. The 2D line and edge representations can be implemented from the 1D cross-profiles. For each detected event, the dominant orientation is computed (orientation of the maximum $C_{s,i}(x,y)$ response), and the corresponding 1D profile is rotated to this orientation. For generating 2D images, for example for illustrating illusions, it is necessary to interpolate values between two consecutive profiles (of neighboring cells). Using the normal definition of a Gaussian

$$G(x,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right), \tag{6.1}$$

a generalized positive line (in 1D) is described by

$$\Delta(x) = G(x, s\sigma_l) \tag{6.2}$$

where σ_l defines the width of the profile and s the scale, and an ideal (sharp) line by the Dirac function

$$\delta(x) = \lim_{\sigma\to 0} \Delta(x) \tag{6.3}$$

such that

$$\int_{-\infty}^{+\infty} \delta(x)dx = 1. \tag{6.4}$$

Similarly, a generalized positive edge with width σ_e is defined by

$$\Lambda(x) = \Phi\left(x/s\sigma_e\sqrt{2}\right) \tag{6.5}$$

where $\Phi(z)$ is the (generally complex) error function

$$\Phi(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt, \tag{6.6}$$

and an ideal (step) edge by

$$\vartheta(x) = \lim_{\sigma\to 0} \Lambda(x). \tag{6.7}$$

Hence, $\vartheta(x) = \pm 1$ for $x \lesseqgtr 0$ and $\vartheta(0) = 0$. Negative lines and edges are obtained by multiplication by -1.

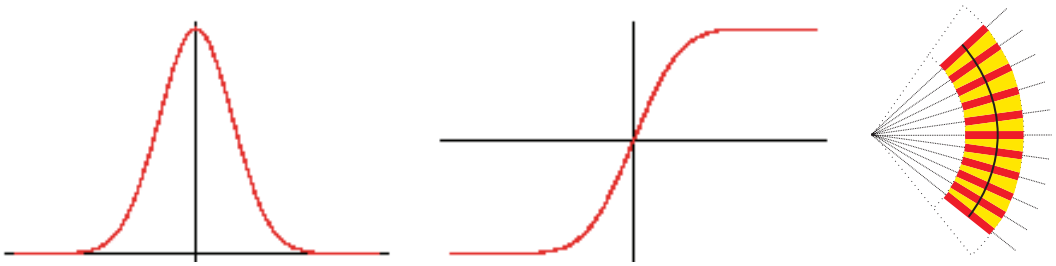

Figure 6.1: Left to right: 1D Gaussian line and errorfunction edge profiles, and the mapping and interpolation to obtain 2D profiles (see text).

Figure 6.1 illustrates the two generalized (scaled) functions, Gaussian profile (left) and the errorfunction profile (middle). Since the latter is not localized, in practice it is multiplied with a Gaussian window with a size which is also scaled. The right part of Fig. 6.1 illustrates the general idea of the 2D interpolation of the 1D profiles, in this case a group of connected events on an arc (thin back line). For each event, the 1D profile is placed perpendicular to the arc because of the dominant orientation (the profiles are represented by the red lines), and all the gaps are filled by local interpolation (represented in yellow). It should be emphasized that this solution is necessary for producing images; in reality such processes may not occur because of the learned symbolic interpretation of line and edge cells (syntax of a coffee mug and the semantics or purpose of the object).

Figure 6.2 illustrates the symbolic line/edge interpretation and the (re)construction process in 2D, in fact the basis for the brightness model. The top four rows show positive and negative edge and line representations, from fine (left) to coarse (right) scales. The 3rd row in Fig. 4.5 shows the line/edge coding. We used $\lambda = \{4, 8, 14, 20, 26, 32\}$, λ being the wavelength of the simple cells given in pixels, the same scales as used in the other chapters.

The bottom row illustrates visual reconstruction of the mug, from left to right: input image, lowpass-filtered image, the summation of the fine-scale symbolic line/edge representations (shown above at left), the same at coarse scale (shown above at right) and the reconstruction result. The number of scales used in this reconstruction result is the same as used in the brightness model (for more examples see also Section 4.3).

The model presented above provides a completely new way for image (re)construction, not like coding based on wavelets. An additional observation is that there is a lot of neural noise in the system and we do not know whether there exist simple and complex cells etc. at *all* retinotopic positions and tuned to *all* scales and *all* orientations (representation noise and completeness). Maps of stained hypercolumns and neural layers and pictures of dendritic/axonal fields of most if not all cells look quite random [Hubel, 1995]. Nevertheless, the image that we perceive looks rather stable and complete. It is very simple to simulate what happens when we suppress information, both in the brightness model as described here, and in wavelet coding as modeled by straightforward summation of responses of simple cells. For example, it is possible to suppress 50% of all information by a random selection. Figure 6.3 shows what happens: the result is a graceful degradation in the case of the brightness model (bottom-right), but a very disturbing rippling in the case of wavelet coding (bottom-left).

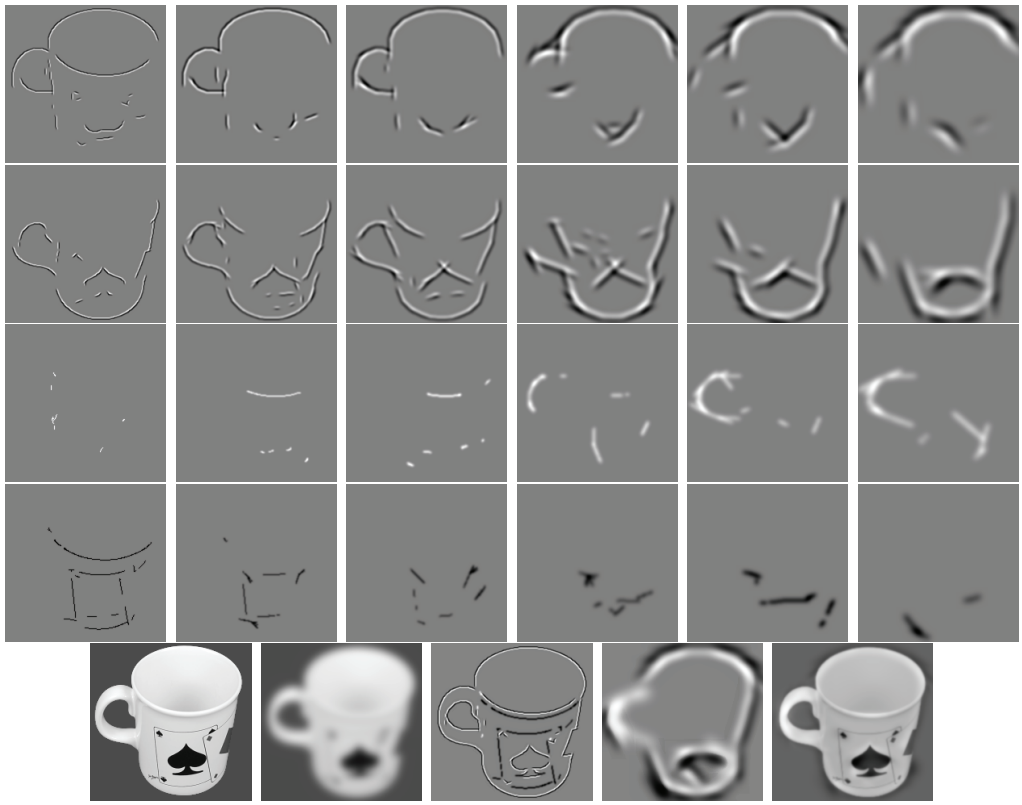

Figure 6.2: The top four rows show positive and negative edge and line representations, from fine (left) to coarse (right) scales ($\lambda = \{4, 8, 14, 20, 26, 32\}$). The bottom row illustrates visual (re)construction of the mug, from left to right: input image, lowpass-filtered image, the summations of symbolic line/edge representations at fine and coarse scales, and the (re)construction result.

6.2.1 The blocks of the model

As mentioned above, the brightness model is a 2D extension and refinement of the 1D brightness model from du Buf and Fischer [1995]. That model was already based on the multi-scale symbolic line and edge representation, with additional lowpass information, combined with nonlinear amplitude transfer functions of the various channels. The proposed model (Fig. 6.4) is based on the same principle, but it additionally includes calibration of the nonlinear amplitude transfer functions of the channels. The model is composed of five blocks, as follows:

Detection and representation: the first step consists of multi-scale line and edge detection and stabilization; see Chapter 4. We use six central scales, i.e., $\lambda = \{4, 8, 14, 20, 26, 32\}$, all linearly spaced except for the first one ($\lambda = 4$). Around each central scale, scale space was sub-divided into 9 scales (the central plus eight more) with $\Delta\lambda = 0.5$. Those additional scales were only used for stabilization of event detection. Only events which are stable over at least 5 of the 9 scales were preserved. Stabilization leads to the elimination of events which are not stable over enough neighboring scales, and therefore to fewer but more reliable events; see Chapter 4 or Rodrigues and du Buf [2006a,b] for stabilization examples with objects and faces. Similar image representations can be obtained by other multi-scale approaches [Lindeberg, 1994].

Figure 6.3: Top: wavelet coding (left) and visual (re)construction (right) using all information (six scales and eight orientations). Bottom: results after randomly suppressing 50% of all information.

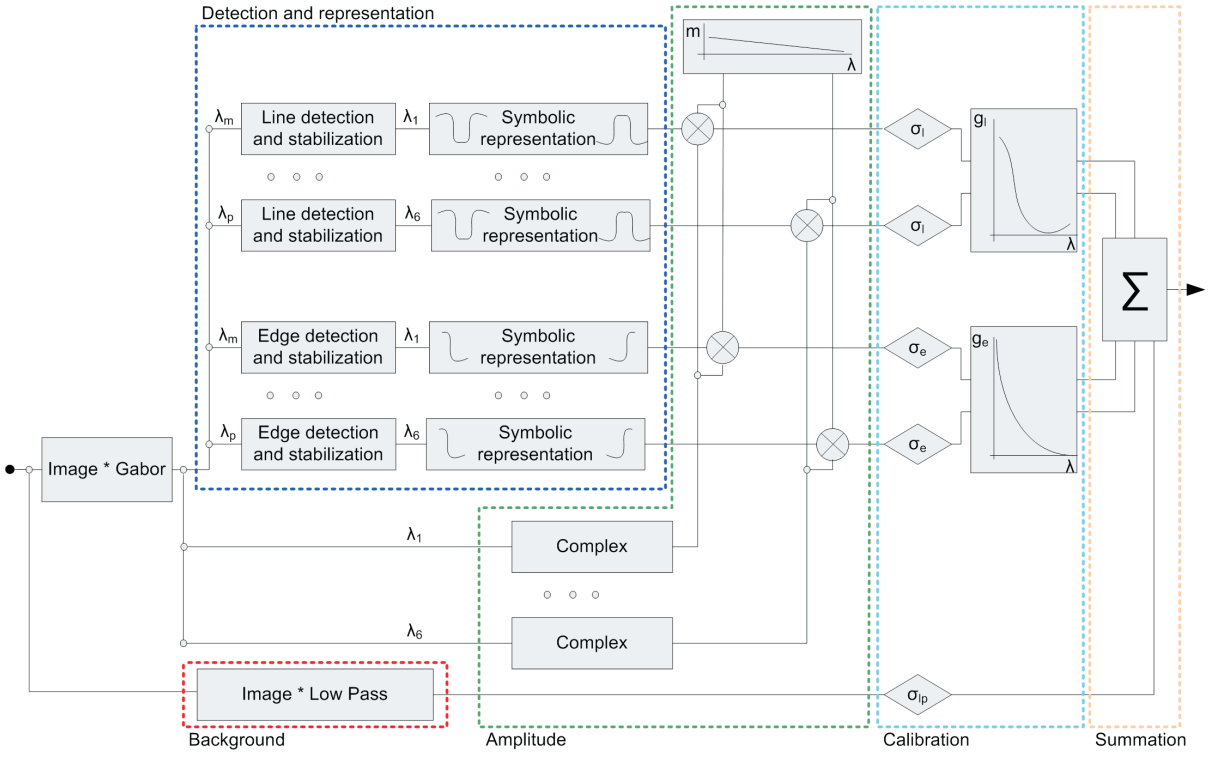

Figure 6.4: Block structure of the brightness model; "image" stands for a 2D stimulus pattern and $*$ denotes convolution (or correlation) with a filter kernel.

The use of more scales (central plus stabilization ones) will improve results, but using more scales implies more CPU time and more storage capacity. In this chapter we present results obtained with more than 130 different stimulus patterns, each sized 256×256 pixels,

and all processed with 54 scales (6 groups of 9 scales), with 8 orientations and 4 representation maps at each scale, plus the necessary storage for the "amplitudes" of simple and complex cells. The use of 6 central scales was a trade-off between, on the one hand, CPU time, storage capacity required and the number of different stimulus patterns, and, on the other hand, precision and quality of the results.

After line/edge detection and stabilization, the next step is the symbolic representation. As mentioned before, each line is represented by a Gaussian function with width σ_l, and each edge is represented by a bipolar, Gaussian-truncated errorfunction with width σ_e. All widths are coupled to the size of the receptive field of the complex cells at each scale. The representations were normalized: between $\{0, 1\}$ and $\{-1, 0\}$ for positive and negative lines, respectively, and between $\{-1, 1\}$ in the case of edges.

Background: lowpass information is obtained by the convolution of the stimuli with a 2D Gaussian filter kernel of size σ_{lp}. This information is used to create a diffuse background, on top of which the different event representations will be applied. This diffuse information could be used to initialize object categorization, perhaps by means of fast gist vision (see Chapter 5 and [Oliva and Torralba, 2006]). Later, all information composed of lines, edges and keypoints etc. at the different scales (from coarse to fine [Bar et al., 2006]) are complementing the lowpass information for visual (re)construction and object recognition.

Amplitude: real amplitudes must be applied to each of the symbolic representations after the normalization. There are two components: (i) actual amplitudes of the complex cells, and (ii) psychophysical data concerning the contrast of (co)sine gratings, i.e., the modulation depth m at different spatial frequencies required for an equal *subjective* contrast. The latter were from du Buf [1987] and du Buf [1992a]; see Fig. 6.6 (top-right, the blue line) and the explanation in the calibration section below.

Calibration: at this point several free parameters need to be determined and the entire model must be calibrated such that it can fit or reconstruct some basic stimulus patterns: (i) σ_{lp} in the lowpass channel; (ii) the width of event representations $\sigma_{\{l,e\}}$; plus (iii) the gain constants $g_{\{l,e\}}$ applied at each scale.

Summation: the final step consists of the linear summation of all the components, which results in a 2D prediction (image) for each tested stimulus.

6.2.2 Model calibration

An initial calibration was accomplished in two interactive steps: (i) all free parameters σ_l, σ_e, g_l and g_e (except for σ_{lp}) were adjusted such that the model can reconstruct basic patterns composed of line and edge structures (see Fig. 6.5 top-row), and at the same time the model must simulate psychophysical data concerning the contrast of cosine gratings (see Fig. 6.6 top-right) and the brightness of disks (at top-left). (ii) Then, only the size of the background kernel σ_{lp} was adjusted with the same objectives. These two interactive and iterative steps proved to be more efficient then calibrating all in a single step.

After the initial calibration a fine calibration was applied, but now only focusing on the background: (iii-1) adjusting parameters as best as possible for basic edge and line structures (Fig. 6.5 top-row). The best value obtained was $\sigma_{lp} = 5.0$, and this value will only be used in Mach-band simulations. (iii-2) For all other patterns, we focused on adjusting as best as possible the parameter such that the model fitted psychophysical data of cosine gratings and disks (Fig. 6.6 top row). The value obtained was $\sigma_{lp} = 3.8$. For a detailed explanation of Figs 6.5 and 6.6 see below.

The best values of the other free parameters were $\sigma_e = 2\lambda/5$ and $\sigma_l = \lambda/5$. Adjusting

the value of σ_l was quite subjective, because one aim was to produce a single sharp line and a pulse with sharp flanks. In this case the best compromise had to be found between a Gaussian bell curve (the contribution of the lowpass channel) and the sharp line/pulse; see Fig. 6.5 (top) for the simple input patterns and four cross-sections of output images (the blue curves). Small changes near the selected parameter values did not show any significant improvements of the output patterns.

The gain values applied at each scale (λ) to the line and edge representations are shown in Fig. 6.6 (bottom) by different symbols: $\square$ (g_l) and $\diamondsuit$ (g_e). All calibrated parameters are summarized in Table 6.1.

parameters	σ_{lp}	σ_l	σ_e	gains
Mach bands	5.0	$\lambda/5$	$2\lambda/5$	Fig. 6.6 bottom
others	3.8	$\lambda/5$	$2\lambda/5$	Fig. 6.6 bottom

Table 6.1: Model parameters after calibration.

In Fig. 6.5 and in many other figures—unless explicitly mentioned—the results are shown as a cross-section through the center line of a 2D image, i.e., the 128th line of 256 lines each with 256 pixels, with the following colors: the input signal is shown by a thin black line and the model prediction (output signal) in thick blue. The red curves correspond to the lowpass component and the green ones to the summation of all line and edge representations. In most cases the green curves have been shifted vertically to overlap the other curves (the line and edge representations are normally around zero because they do not include a DC component).

Figure 6.6 (top-right) shows psychophysical data concerning the contrast of concentric (radially-symmetric) cosine gratings. Data points were taken from du Buf [1987], Fig. 5, pp. 79. The different curves represent the modulation depth (m) for different spatial frequencies at five different contrast levels. Curve number 1 is the detection threshold at very low (visible) contrast, and the other curves are matching results obtained with a reference stimulus at four supra-threshold contrasts. For a detailed explanation of these data and the measurement errors see [du Buf, 1987]. The blue straight line corresponds to a model fit. In fact, this part of curve 5 was used to calibrate the model, i.e., adjust the channel gain factors g_l and g_e. The horizontal axis is given in cpd or cycles per degree of visual angle. Since half pixel corresponds to one minute of visual angle, 10 cpd (log equal to 1.0) corresponds to one cycle (period) in 3 pixels. This is the reason that no higher frequencies have been used: the input signal would be sampled too coarsely.

Figure 6.6 (top-left) shows psychophysical data concerning the brightness in the center of disks. Data points were taken from du Buf [1987] Fig. 15, pp. 69. The image shows decrement luminance (ΔL-) as a function of the area of the disks. Curves 1 to 5 show detection thresholds plus least-squares approximations by level-dependent point spread functions (PSFs) of center-brightness data. The darker blue curve shows a prediction. Basically, this curve should be about horizontal with an increase toward left at a disk size of about 30 square minutes (log equal 1.5), the edge of the yellow area to the left. Again, for smaller disks the approximation is too coarse and the result is not reliable. The darker blue curve has been shifted vertically such that it superimposes a data curve, since no real matching experiment with a reference stimulus was simulated, but in the future this must be done (it will require tremendous CPU times!).

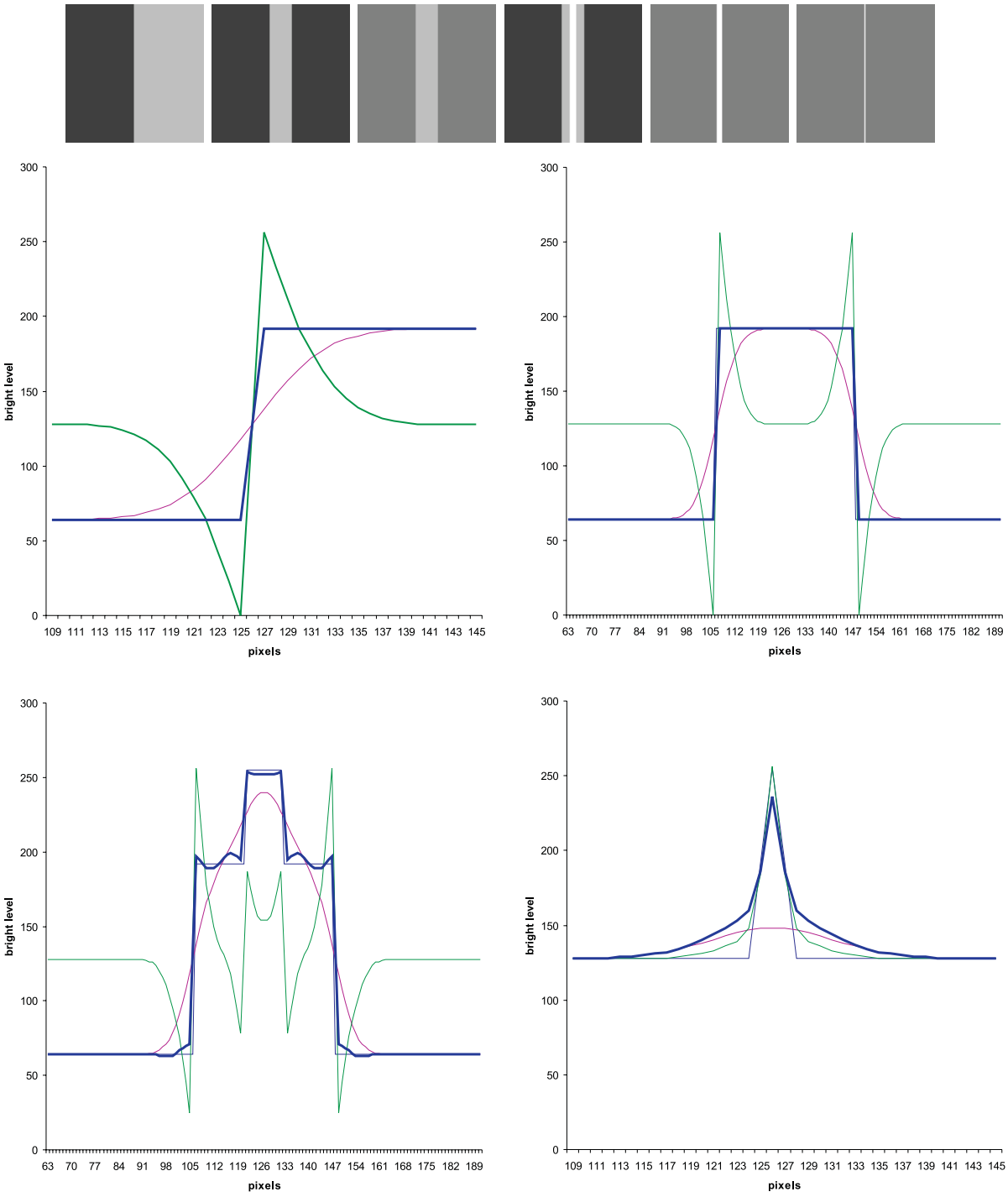

Figure 6.5: The top row shows examples of simple stimulus patterns used for model calibration. The bottom rows show cross-sections as signals: edge, bar, thin bar on wide bar and line, respectively. Color coding: model output in thick blue, lowpass component in red, summation of line/edge representations in green and input signal in thin black.

It should be mentioned that measurement errors are not shown in Fig. 6.6 (top-left and top-right). Such errors can be rather big and will clutter data points or graphs which present mean (averaged) values over many repeated experiments. Below, in many figures the errors

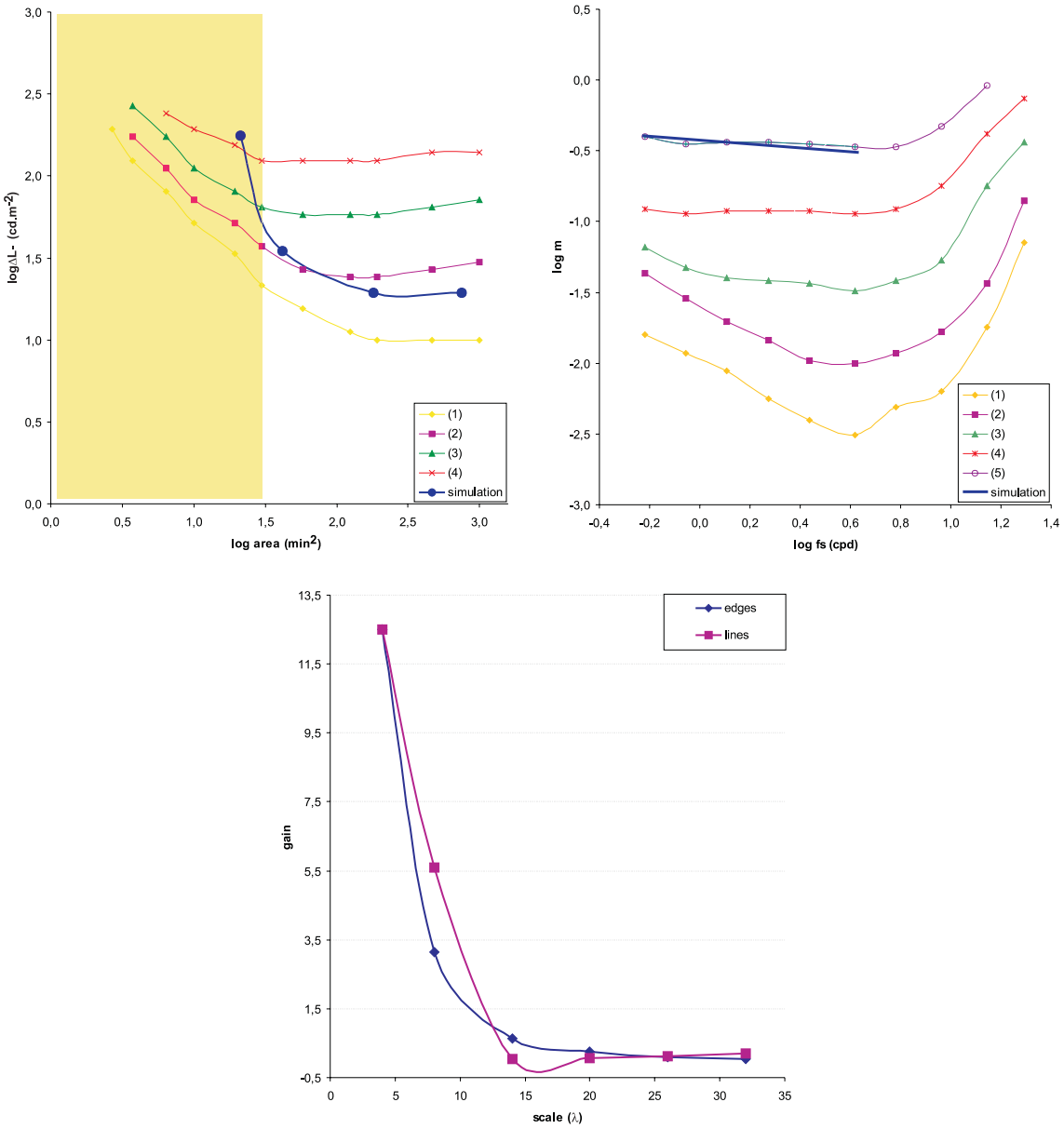

Figure 6.6: Top-left: psychophysical data of the brightness of decremental disks with a model prediction in dark blue superimposed on the data. Curves 1 to 5 show detection thresholds and least-squares approximations by level-dependent PSFs. Top-right: data concerning radially-symmetric cosine gratings, plus model calibration in dark blue. Curves 1 to 5 show detection thresholds and contrast-matching results. Bottom: the gain constants used in the line and edge representations as a function of scale.

will not be shown (unless they were very clearly specified in the original papers!), for three fundamental reasons: (i) As already mentioned, a cluttering of the data points or graphs. (ii) In almost all publications measurement errors are presented in a different way, and then either plotted in the graphs or mentioned in the text. In some publications it was simply impossible to retrieve the values of the errors with some precision. (iii) In order to present and explain some errors, especially big ones, it is necessary to explain at least part of the

experiment. Due to the amount of different data to be presented, detailed descriptions will distract the reader. In any case, the reader can have access to the original papers and we are only concerned with some general conclusions concerning the applicability of the model.

6.3 Experiments

In the previous section it was shown that the brightness model can predict some psychophysical curves concerning the contrast of concentric cosine gratings and the brightness of disks. As a matter of fact, the model has been linearized such that it may only be applied to patterns with high contrast; the reason is that the model was calibrated using the dark blue line in Fig. 6.6 (top-right). In this section predictions produced by the calibrated model will be compared with the results of psychophysical experiments on different brightness effects, such as Mach bands [Pessoa, 1996b,a; Keil et al., 2006], assimilation, simultaneous brightness contrast and several variations of White's effect [Blakeslee and McCourt, 1999, 2004; Blakeslee et al., 2005] and the Craik-O'Brien-Cornsweet illusion [Moulden and Kingdom, 1990]. It will be shown that the model can also predict other illusions, such as Chevreul steps, grating induction and variations of Adelson's tile patterns. Results will be discussed along the text.

6.3.1 Mach bands

The illusion referred to as Mach bands is named after the Austrian scientist Ernst Mach who studied it first. It refers to bands that appear adjacent to dark and light gradients, for example as bright and dark bands next to ramp edges (where a ramp meets a plateau, also called inflection point). Observed in astronomy and microscopy, it was long thought to be an optical effect, but the effect is created by our visual system. Figure 6.7 (top) shows four examples with linear ramps with increasing width. Many studies have been devoted to Mach bands, possibly because it may reflect processing at a very early stage, for example the bandpass retinal ganglion cells, although linear filtering is now excluded because Mach bands do not occur at ideal step edges but only at ramp edges. One of those studies was done by Ratliff and colleages, who investigated the appearance of Mach bands by varying spatial illumination patterns in adjacent positions.

Pessoa [1996b] summarized their results as follows: (a) a rectangular bar stimulus of sufficient contrast placed near inflection points attenuates the Mach band that normally is perceived at the inflection point; if the bar is positioned close enough, the Mach band may disappear; (b) a bar far away from the inflection point has no effect on Mach band appearance; (c) attenuation largely depends on the width of the adjacent stimulus; (d) attenuation is largely independent of the sign of contrast of the bar; (e) a bar with a triangular cross-section near the inflection point enhances the nearby band; when the bar is moved and its own associated Mach band approaches the stationary Mach due to the ramp edge, the two bands will fuse and produce an enlarged Mach band. This enhancement only occurs when both Mach bands are of the same polarity, light or dark. In the case when they have opposite polarities, light plus dark, they may attenuate or even cancel each other.

Summarizing, the three main features of interference with Mach bands are proximity, contrast and sharpness. These interferences are illustrated in Fig. 6.8). For a complete review of results, models and theories about Mach bands see [du Buf, 1994; Pessoa, 1996b,a; Keil et al., 2006].

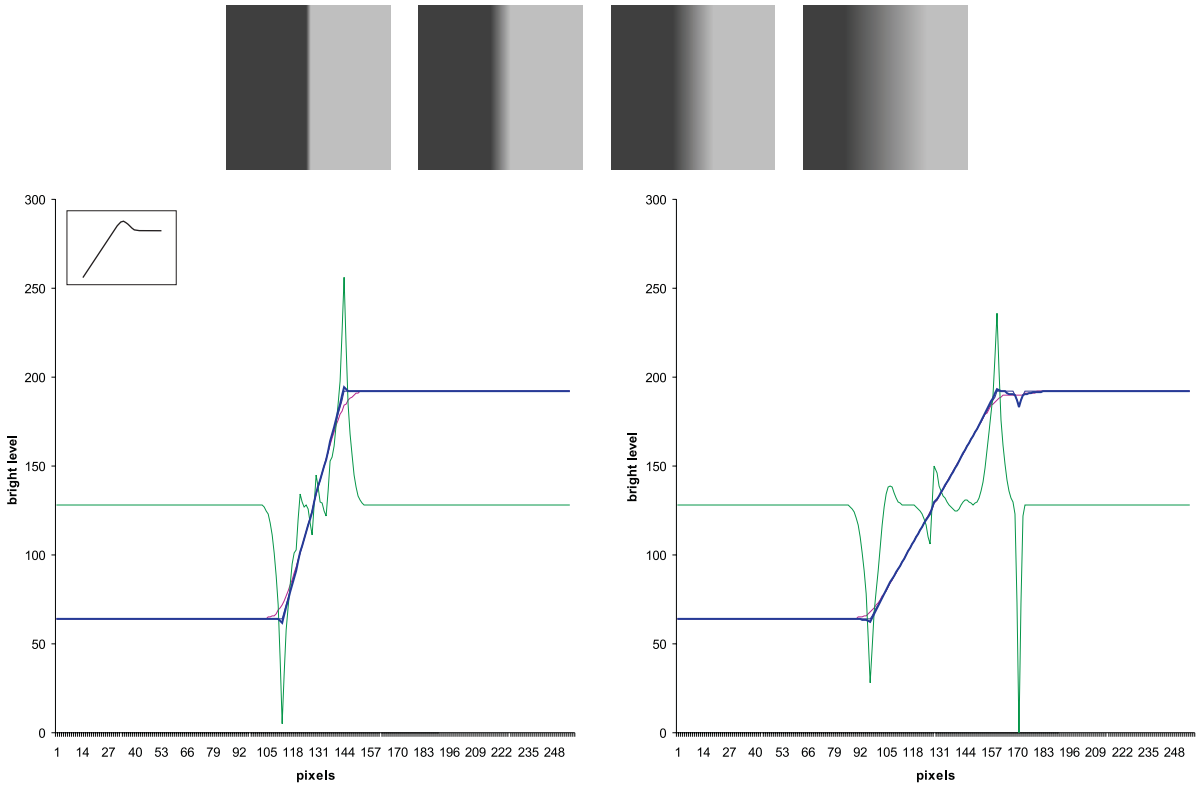

Figure 6.7: Top: four ramp widths used for predicting Mach-band effects. Bottom: simulated results in case of a ramp (left) and a ramp with a negative triangular bar as adjacent stimulus (right).

Figure 6.7 shows, apart from linear ramps with four widths at top, two model predictions at bottom: only a linear ramp at left (the input was the second image from left above), and a linear ramp plus an adjacent negative triangular bar at right (the ramp input without bar was the third image from left above). The small panel in the left graph shows the expected Mach-band effect, which is more difficult to show in detail in the main graph (blue line) due to resolution problems in the conversion of numerical results into graphics.

Figure 6.8 summarizes all model predictions (at right) and the original psychophysical data (at left). The first group of tests—shown at top—concerned the threshold contrast of trapezoidal gratings for seeing light and dark Mach bands. Data points were taken from Fig. 5 in Pessoa [1996b], but the experiment originates from the work of Ross and colleages in 1989. Such data are thought to reflect Mach-band strength or amplitude as a function of the spatial frequency of the trapezoidal grating, and a maximum perceived strength was found at some medium frequency. Both narrower and wider ramps lead to a decreased visibility of the Mach bands, and Mach bands were hardly seen or not visible at all with luminance steps (Keil and others refer to this as the "inverted-U" behavior because of the shape of the curves). The top-right graph shows that the model can predict this behavior. Since the model does not yet include asymmetrical processing of light and dark patterns on a same gray background, it cannot predict differences between light and dark Mach bands.

The bottom-left graph in Fig. 6.8 shows psychophysical data on the width of dark Mach bands with various adjacent stimuli as a function of the distance to the inflection point. The *width* is much easier to measure than the strength, but normally wider bands are also

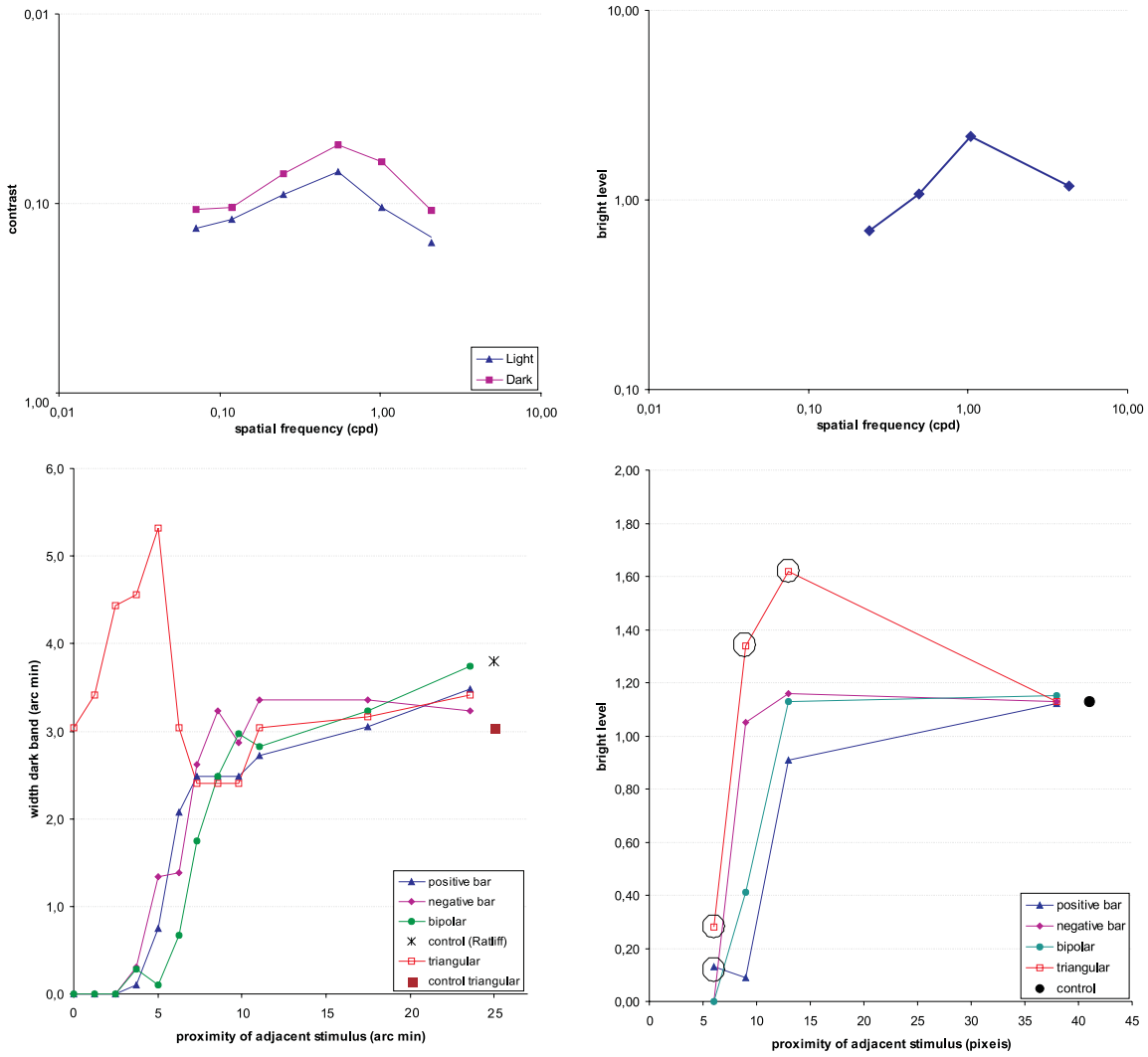

Figure 6.8: Mach bands. Psychophysical data (left) and model predictions (right). Top: Mach-band strength as a function of ramp slope. Bottom: influence of adjacent stimuli. See text.

stronger. Data points were taken from Figs 9 (left) and 10 (left) in Keil et al. [2006] (and these in turn originated from earlier publications by Ratliff and colleages in 1983). The data show the influence of a positive and negative bar (i.e., monophasic bars) and a bipolar (or biphasic) bar. The asterisk symbol (control) refers to the condition without adjacent stimulus. Also presented is the data in the case of a triangular bar and the respective control condition (open and solid square symbols). The latter differ from Ratliff's data because they were measured by another observer [Pessoa, 1996b].

The bottom-right graph in Fig. 6.8 shows model predictions. In the case of the positive, the negative and the bipolar bars the same trends are predicted if compared with the real data. The only exceptions are the encircled symbols. In case of the triangular bar (in red), the enhancement of the Mach band occurs at a (slightly) larger distance and no attenuation effect is predicted at larger distances (if the latter effect is significant (in the left graph) relative to measurement error). In case of the positive bar (in blue), the measured value at a distance of 6 pixels should be lower and/or the second point at 9 pixels should be higher.

Results could be improved by changing the balance between the Gaussian lines at the ramp edge and the bar on the one hand and the contribution of the lowpass channel, but this would worsen other predictions. Predictions of all individual effects can be improved by tweaking some model parameters, but our goal is to show predictions of all effects using some general compromise.

6.3.2 Brightness induction

The brightness of a region is not solely related to that region's luminance but it also depends on the luminances of adjacent regions. This phenomenon is known as brightness induction and it includes two opposite effects: simultaneous brightness contrast and assimilation [Blakeslee and McCourt, 2004]. Simultaneous brightness contrast (SBC) occurs when the brightness of a region shifts away from the brightness of the adjacent regions. A textbook example of SBC is that a gray test patch on a white background looks darker than an equiluminant gray test patch on a black background. SBC decreases with increasing size of the test patch [Blakeslee and McCourt, 1999]. Brightness assimilation refers to the opposite situation in which the brightness of a region shifts toward that of the surrounding regions. For a complete review of these effects and related theories see [Moulden and Kingdom, 1989; Blakeslee and McCourt, 1999; Blakeslee et al., 2005].

White's effect is a brightness illusion which illustrates the fact that the same target luminance can elicit different perceptions of brightness in different contexts. Usually it is illustrated by gray test patches or bars of identical luminance placed on the black and white bars of a squarewave grating. The effect still lacks a good explanation, because the induction effects of lateral and collinear bars are different and simple changes of the context can lead to completely different brightness levels. Blakeslee and McCourt [2004] reported that in White's effect the direction of the brightness change does not depend on the aspect ratio of the test patch, i.e., the direction of the brightness change does not correlate with the amount of black or white border in contact with the gray test patch, or in its direct vicinity. When the test patch is a vertically oriented rectangle sitting on the white stripe of a vertical grating, it has two short sides that are in contact with the collinear white bar on which it sits, and two long sides that are in contact with the flanking black bars. In this configuration the test patch has more extensive contact with the dark flanking bars, yet the gray patch appears darker than a similar gray patch situated on a dark bar and flanked by white bars. The effect cannot simply be attributed to assimilation, however, since the direction of the effect is unchanged even if the height of the test patch is reduced until it has more extensive border contact with the bar on which it is situated. In addition, although White's effect has been reported to increase with increasing spatial frequency of the squarewave grating, the effect does not disappear or reverse at low spatial frequencies.

Brightness induction was measured by Blakeslee et al. [2005], who used a set of stimuli to illustrate the relationship between the Howe stimulus [Howe, 2001], the White stimulus and the classical SBC stimulus. The White stimulus and the SBC stimulus occupy opposite ends of a continuum in which the Howe stimulus is the midpoint [Blakeslee et al., 2005]. Blakeslee and colleages replicated the Howe experiment and quantified brightness induction in the Howe stimulus relative to that in the SBC and White stimuli. Another variation of the White stimulus was introduced by Anderson [2001], who examined the influence of multiple test patches (bars). Blakeslee and McCourt [1999] determined the magnitude of SBC as a function of increasing patch size, and Blakeslee and McCourt [2004] examined the role of contrast and assimilation effects by measuring the magnitude of the White effect and

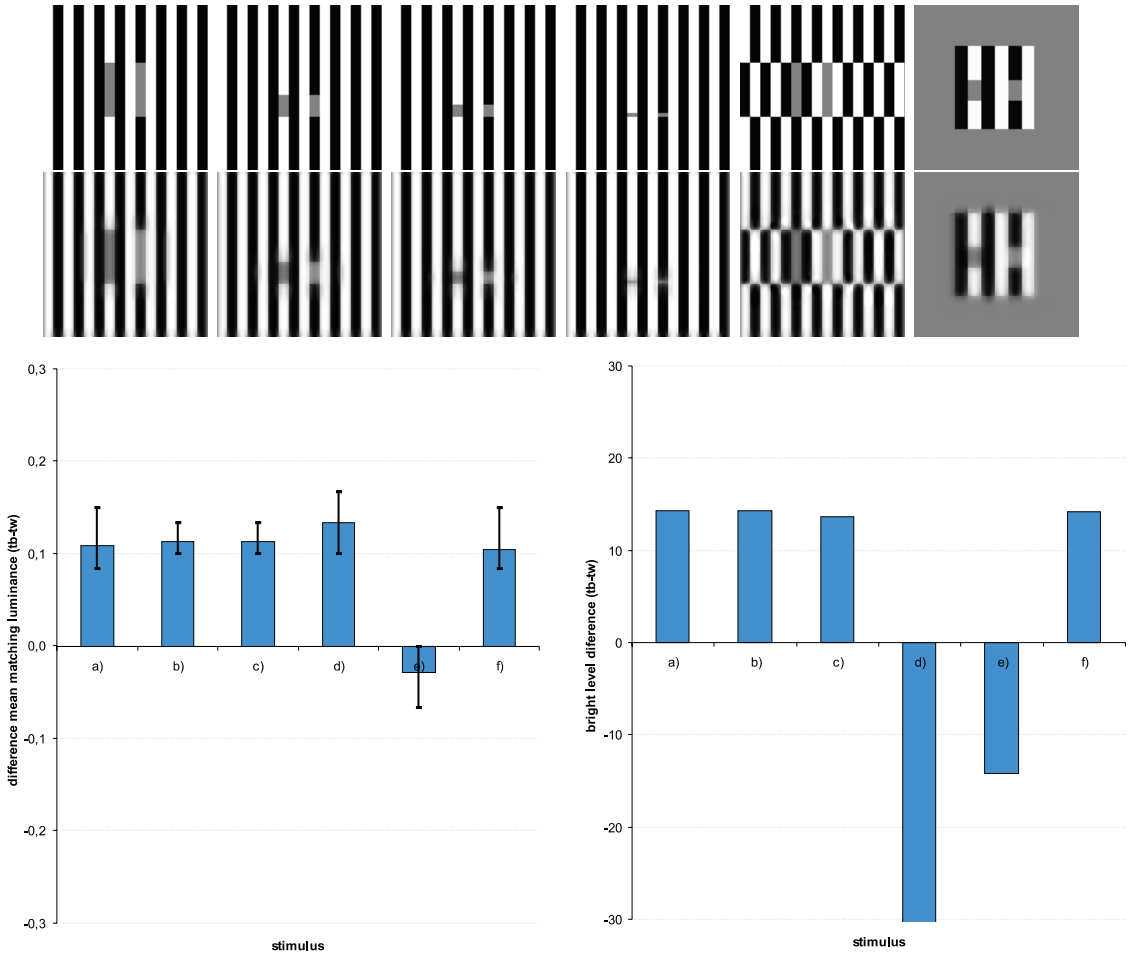

Figure 6.9: Top row, left to right, stimulus patterns (a) to (f): White's effect with different heights of test patches, a shifted White, and one with a different spatial frequency. Second row: corresponding model predictions. Bottom: psychophysical data (left) and model predictions.

the shifted White effect (also the checkerboard illusion, not discussed in this chapter) as a function of the inducing pattern's spatial frequency and the height of the test patch.

Figures 6.9, 6.11 and 6.12 illustrate some of these situations. The top row in Fig. 6.9 shows, left to right, six stimuli (a) to (f). In the first four (a-d) the height of the test patch is varied for one spatial frequency of the standard White stimulus, the same height is used with two different spatial frequencies (b vs. f), and there is one example of a shifted White effect (e). Please notice that the test patches on white bars are always located to the left, while test patches on black bars are always on the right. The second row of images shows corresponding model predictions in 2D. The bottom row shows psychophysical data (left) and predicted results (right).

Data were taken from Fig. 3 in Blakeslee and McCourt [2004], and are plotted as the difference in mean matching luminance (over four observers) for test patches on a black (tb) and a white (tw) bar. The data bars represent the group means and the black vertical lines the intervals (maximum and minimum) of the individual observers.

A summary of stimulus parameters used in the psychophysical experiments and in our simulations is presented in Table 6.2. We note that where two values appear in the same

cell, the left one was used in the experiments and the right one in our simulations. In the case of stimulus (r), explained below, two psychophysical results were presented by Blakeslee and McCourt [1999], both quite different from the simulated configuration (most similar was the stimulus with a spatial frequency of 0.125 cpd and a patch height of 1^o). Therefore the experimental values are omitted.

stimuli	a	b	c	d	e	f	g-m	n	o	p	q	r
freq.	1.0	1.0	1.0	1.0	1.0	0.5/0.6	0.5/1.0					-/0.6
patch	3^o	1^o	*bar*	*bar*/2	3^o	1^o	3^o	3^o	-/1.5^o	-/1.2^o	1^o	-/0.7^o

Table 6.2: Summary of stimulus parameters in Figs 6.9, 6.10, 6.11 and 6.12. The spatial frequency is given in cpd (cycles per degree) and the height of the test patch in degrees. When two values are given, the left one was used in psychophysical experiments and the right one in model simulations.

The top row in Fig. 6.10 shows, left to right, seven stimuli (g) to (m): standard White with 1 (g) and 3 patches (m), also with changed sections of the inducing gratings of the standard White stimulus with homogeneous white and black bands with increasing height (h) to (k). Original stimuli: Howe (i), SBC (k) and Anderson (l). The second row shows the 2D model predictions. The bottom graphs are two cross-sections in the case of stimulus (m), through the upper (left graph) and lower (right graph) patches.

Figure 6.11 summarizes psychophysical data (left) and predicted results (right) in the case of stimuli (g) to (m) (Fig.6.11). Data were taken from Fig. 3 in [Blakeslee et al., 2005], and represent differences in mean matching luminance between the right and left patches over eight observers. The bars represent the group means and the black vertical lines the maximum and minimum of all observers. For a summary of stimulus parameters see Table 6.2.

The two bar graphs at center-right in Figure 6.12 show data (where available, at left) and model predictions (at right) in the case of the magnitude of SBC with decreasing patch size, stimuli (n) to (q), the (o) stimulus shown at top-left with the 2D model prediction. Data bars n1, n2, q1 and q2 refer to two observers. Data were extracted from Fig. 5 in Blakeslee and McCourt [1999]. The bars represent the mean deviation of matching luminance from the mean luminance, expressed as a proportion of mean luminance. The bars above the horizontal axis represent brightness matchings of test patches on the dark background, while the bars below represent the test patches on the bright background.

The images at top-right show an example of grating induction (GI), stimulus (r), i.e., input at left and model prediction at right. Grating induction, in contrast to SBC, is a brightness effect that produces a spatial brightness variation (a grating) in counter phase. Perceived contrast of the induced grating decreases with increasing inducing grating frequency and with increasing bar height [Blakeslee and McCourt, 1999]. Stimulus parameters are summarized in Table 6.2. The center-left graph in Fig. 6.12 illustrates the model prediction of SBC. The two graphs at the bottom illustrate grating induction, i.e., cross-sections through the center of an iducing grating (left) and through the center of the gray bar (right).

Model predictions can be analyzed focusing on two points, general trends and accuracy. Concerning general trends, almost all tested stimuli presented the same tendencies as the psychophysical data (exceptions are discussed below). In the case of SBC, stimuli (o) to (q), both data and model show increasing magnitude with decreasing patch size. Grating induction was correctly predicted, as were most variations of White's effect.

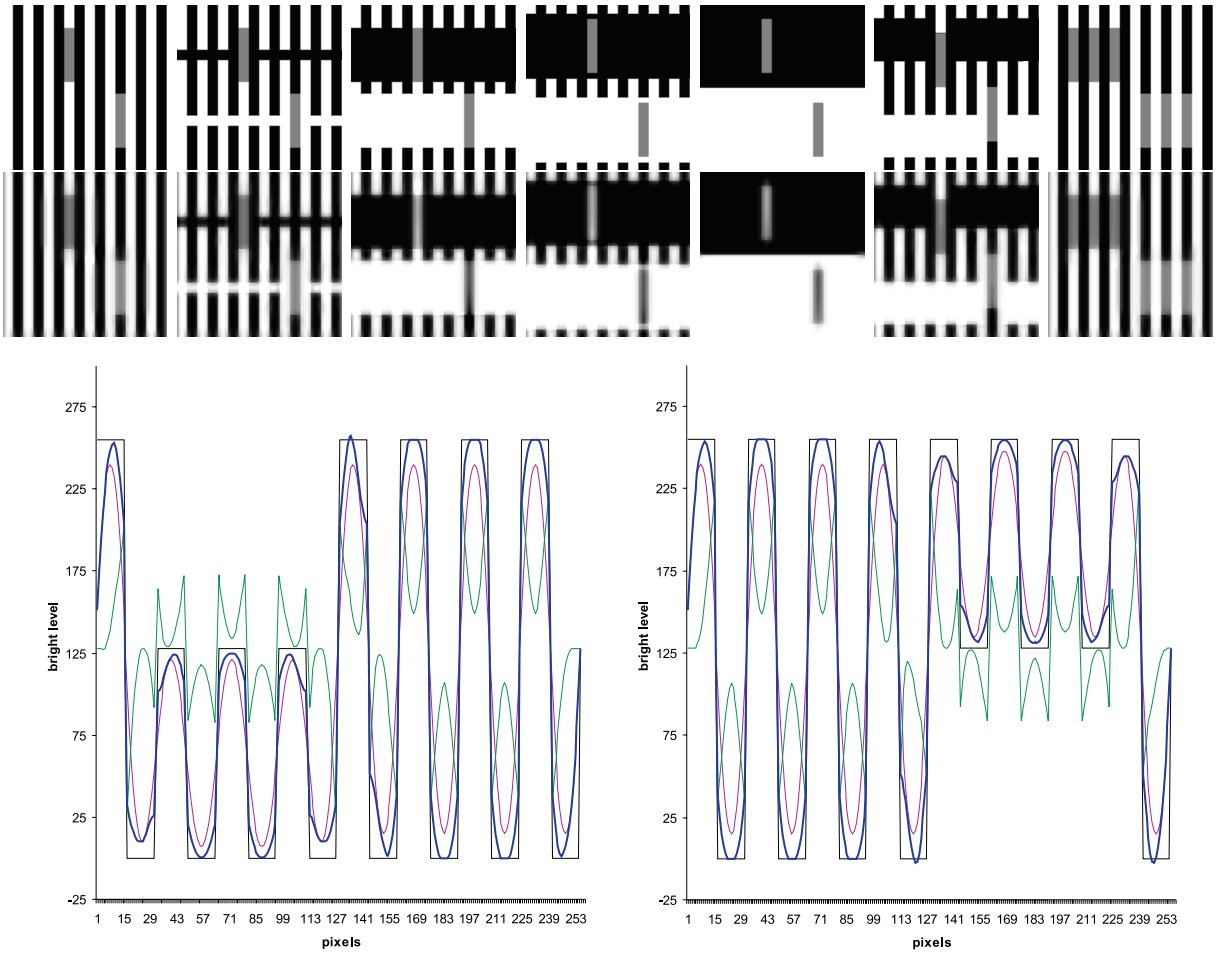

Figure 6.10: Top row, left to right, stimuli (g) to (m): standard White with 1 (g) and 3 (m) patches, and (h) to (k) involve replaced sections of the standard White inducing grating with homogeneous white and black bands with increasing height. Original stimuli: Howe (i), SBC (k) and Anderson (l). The two bottom graphs illustrate model prediction in case of stimulus (m), i.e. two cross-sections through the gray bars.

Exceptions were the following: deviation in the case of stimulus (d) is due to the size (height) of the patch. In this case the patch is very near to the absolute limit of small structures that the model can process (Gabor filters). In the case of the Howe stimulus (i) some doubt appears because psychophysical data present a mean of 0, with a positive maximum and a negative minimum. In this case the psychophysical data are far from conclusive. Analyzing the quantitative predictions by the computational model of Blakeslee et al. [2005], we verified that it produced the same tendency as our model.

The last case is the SBC stimulus (n) with a patch size of 3^o. This fails due to the very big size of the patch. For this stimulus no effect could be measured, because the scales used in the model are static, fixed and limited. As a consequence, there is a limitation to the overall size/area of the patch that can be tested. When using more and coarser scales, or a dynamic scale selection as a function of the area of the patch, the correct effect will be produced.

Analyzing results in terms of accuracy or quality, in the case of the stimuli (a) to (c), (f),

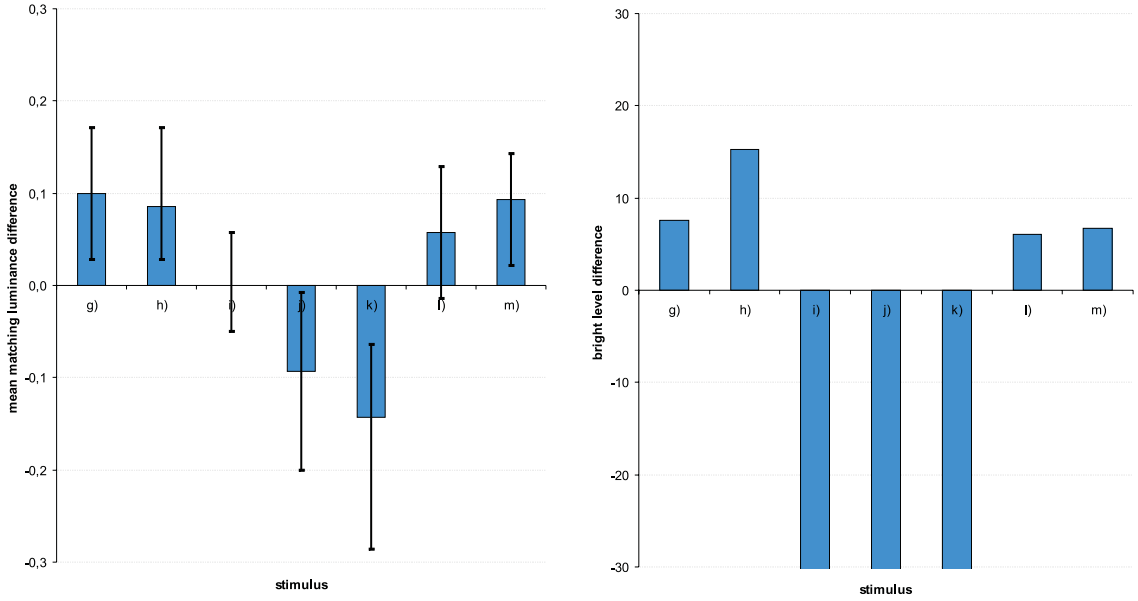

Figure 6.11: Psychophysical data (left) and model predictions (right) in case of stimuli (g) to (m) in Fig. 6.10.

(g) to (h) and (m) to (n) the model predications were inside the ranges of the psychophysical data. Stimulus (i) failed, as already discussed above. For stimulus (k) the prediction is a little bit lower than the lowest value of the psychophysical data. In case of stimulus (j) the result is much lower than the lowest psychophysical data. Concerning SBC, stimuli (o) to (q), the effect is difficult to analyze because the only corresponding stimulus was (q). In this case we may say that the bottom bar coincides with the data of subject MM, despite the fact that the upper bar is smaller than the one of the psychophysical data. Finally, for stimuli (o) to (p) and (r) no conclusions can be drawn in terms of quality because there are no psychophysical data available. Nevertheless, in the case of grating induction, stimulus (r), by considering the data available in [Blakeslee and McCourt, 1999], where the authors compared GI with SBC, and by extrapolation of the results, we might say that predictions should be a bit lower than those obtained.

In the above discussion we have not emphasized two aspects which are extremely important in psychophysics: (a) the model has been calibrated using data of one, other observer, and (b) the calibration data were measured at one high-luminance background level and at high contrast of cosine gratings. The nonlinear model has been linearized for these conditions. These two points imply that a lot of care should be exercized when comparing data over subjects and conditions. This aspect is supported by the measurement errors indicated in the graphs presented here, were available.

Summarizing, of all the eighteen stimuli tested in this section against psychophysical data, sixteen presented correct tendencies and from those only four were clearly outside the ranges of measured psychophysical data.

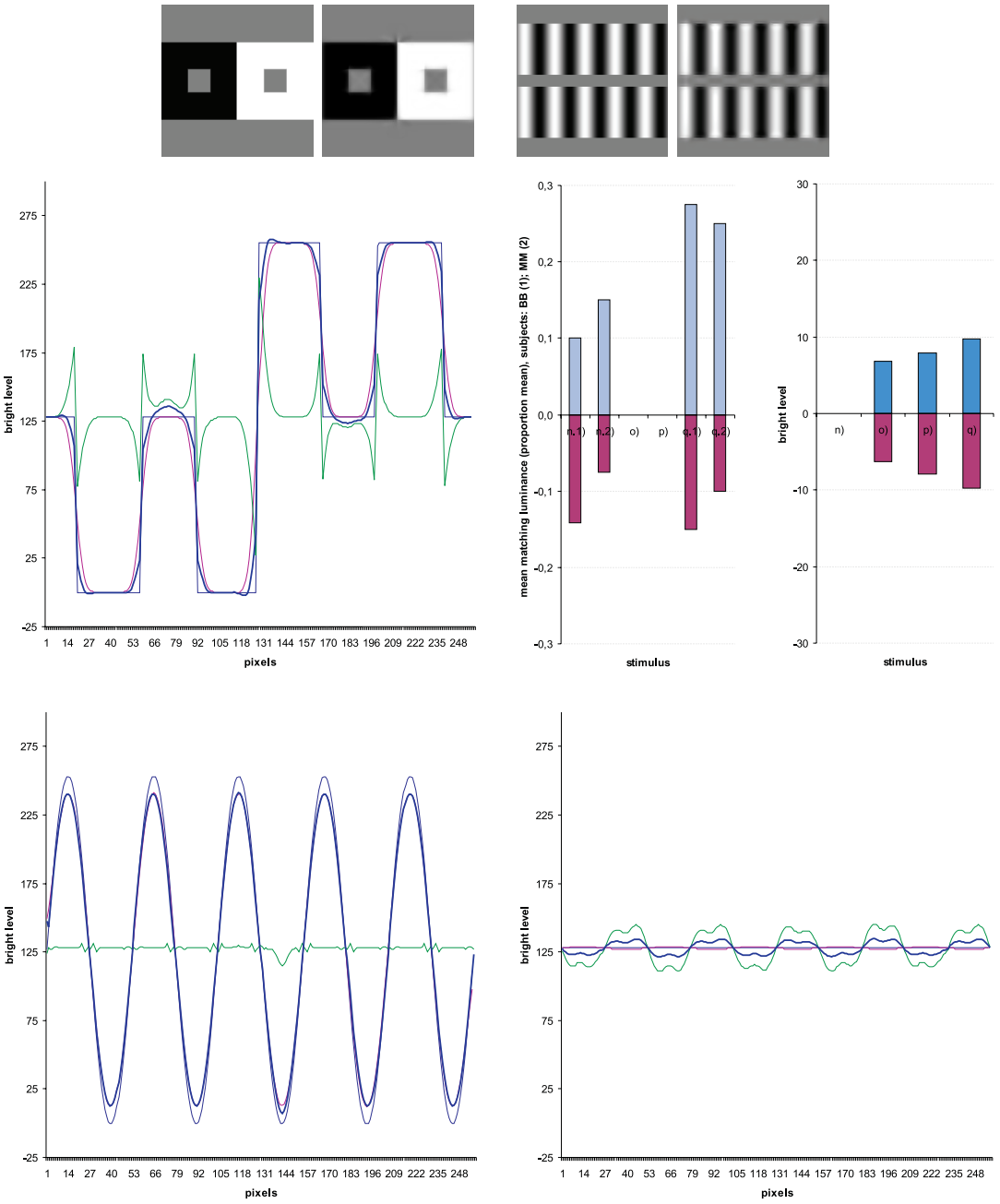

Figure 6.12: Top row: SBC (left) and GI (right) stimuli each with 2D model prediction to the right. Middle row: 1D cross-section through an SBC stimulus (left) and plots of psychophysical data (left panel, two observers) and model predictions (right panel) of SBC with decreasing patch size, stimuli (n) to (q). The bottom row illustrates model predictions in case of grating induction, stimulus (r).

6.3.3 Craik-O'Brien-Cornsweet

The best known illusion from the family of Craik-O'Brien-Cornsweet (COBC) illusions is the one demonstrated by Cornsweet in 1970. The figure looks like a bipartite field, with the left half darker than the right, even thought the two halves have identical reflectance except in a small region across the dividing contour, witch consists of a sharp and a gradual change

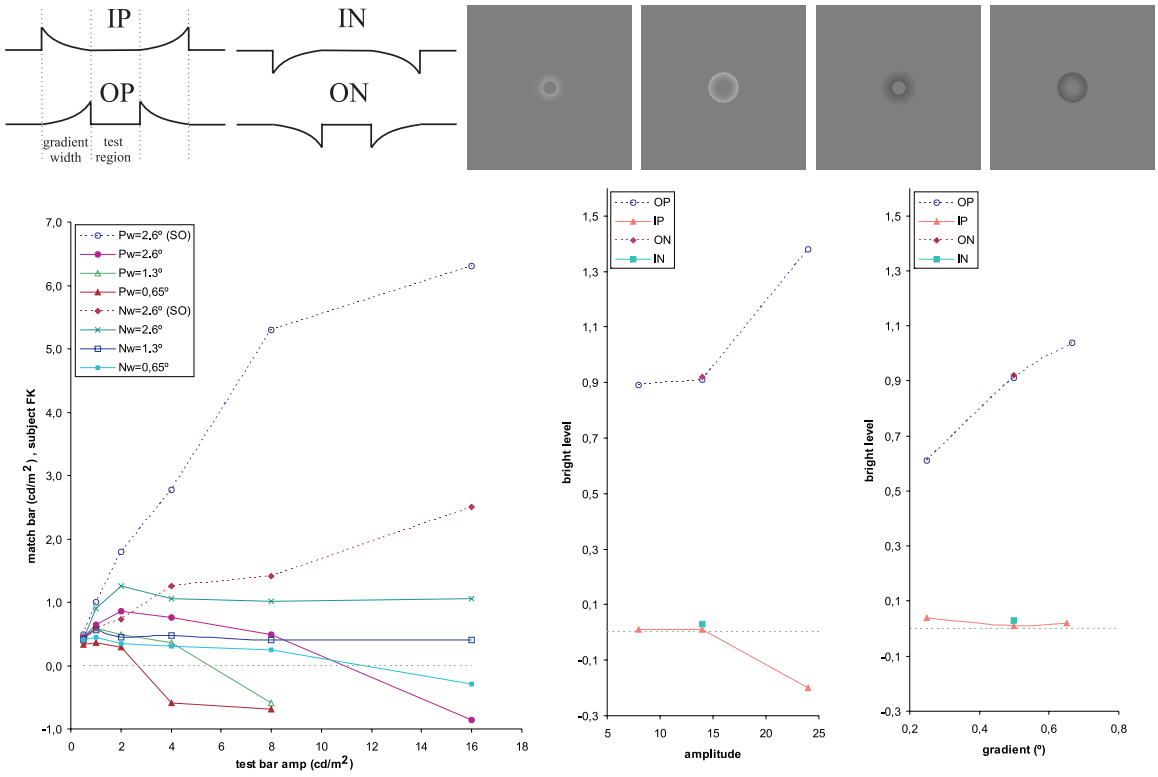

Figure 6.13: Top row, left to right: luminance profiles of Craik-O'Brien-Cornsweet disk stimuli and examples of 2D stimuli used in model predictions. Bottom row, left to right: psychophysical data and simulated results (see text).

in reflectance. This gradient induces a different brightness in the two fields.

Other versions of the illusion are shown at the top of Fig. 6.13. The edges in each disk consist of a sharp discontinuity bounded by a gradient on just one side. Each disk is characterized by a combination of two factors: the edge gradient is negative or positive and it can be inside or outside the sharp discontinuity. This gives four combinations: in-positive (IP), in-negative (IN), out-positive (OP) and out-negative (ON) [Moulden and Kingdom, 1990]. Figure 6.13 at top-left shows the reflectance or luminance (under homogeneous illumination) profiles of the four stimuli and at right examples of 2D images used in model predictions.

Figure 6.13 shows at bottom-left a set of psychophysical data. Data were extracted from Figs 3a and 4a in [Moulden and Kingdom, 1990]. The plots show the luminance of a matching bar which was adjusted to obtain a brightness match with the central area of the COBC stimulus, as a function of the amplitude of the edge gradient. Continuous lines represent stimuli with inner gradients; dashed lines stimuli with outer gradients (also labeled SO); stimuli with positive edges are denoted by P and those with negative edges by N. In the legend, w denotes the width of the edge. The (constant) test region was 1.6^o wide. For IP and ON stimuli, positive values of the matching bar imply that the central test region appeared brighter and negative values that the central test region appeared darker. For IN and OP stimuli the effects are reversed. For more details about the experiments see Moulden and Kingdom [1990]. This study reported data from two subjects (the authors), but in Fig. 6.13 we only show data from subject FK. Data from the other subject were about similar with some deviations.

Moulden and Kingdom [1990] summarized their findings as follows: (a) for both inner gradient (IP and IN) stimuli, an increase in gradient width results in a greater magnitude of induced brightness, whether the brightness induction was negative or positive; (b) in case of the IP stimuli, an increase in amplitude first results in an increase in induced brightness but then followed by a decrease. In many cases the downturn even crosses zero to negative values implying that the central test region appeared darkened; (c) in case of the IN stimuli there is no such downturn at high amplitudes; (d) for OP stimuli the magnitude of darkening is greater at all amplitudes if compared to the magnitude of lightening for IP stimuli; moreover, in the former case there is no downturn at high amplitudes; and (e) there is little difference between ON and IN conditions at the same gradient width.

The two graphs at bottom-right of Fig. 6.13 show model predictions of the brightness in the center of the disks, as a function of gradient amplitude (left) and width (right). The size of the (constant) test region was 0.6^o. In the left graph (function of gradient amplitude) the gradient width was kept at 0.5^o. In the right graph (function of gradient width) the amplitude was kept constant (value of 15), three gradient widths were applied: 0.3^o, 0.5^o and 0.7^o.

All model predictions showed correct tendencies, i.e., the central test area turned brighter in the case of IP and ON stimuli and darker in the case of OP and IN stimuli. All predictions were plotted positive to facilitate comparison with the psychophysical data. In the case of stimuli ON and IN, only two values were plotted as reference.

Comparing results with the psychophysical data (Fig. 6.13 bottom-left), the left graph with model predictions (function of gradient amplitude) shows for the OP stimulus an increasing brightness level in the test region as the amplitude of the stimulus increases, the same tendency as shown by the equivalent stimulus in the psychophysical data. The IP predictions, despite the faint response at low amplitudes, presents the downturn as shown in the psychophysical data; see e.g. Pw=0.65^o. Finally, responses of the IP stimuli are much lower than those of OP stimuli, which is also consistent with the psychophysical data.

The right graph with model predictions (function of gradient width) shows that the OP stimulus produces an increasing response in the test region for increasing gradient width. The IP stimulus shows a small fluctuation; all values remain positive but there is no significant amplitude increase. In contrast, the psychophysical data show an increasing amplitude in this case.

Summarizing, it was shown that the model can predict most of the effects in the case of positive gradients, both inner and outer. In the case of the negative gradients, despite the fact that the brightening and darkening tendencies were correct, resulting amplitudes were very similar to the positive case. It is not yet possible to differentiate (only in terms of amplitudes) between positive and negative gradients. As for other predictions presented in previous subsections, involving bright-dark effects, this is caused by the amplitude symmetry of the model.

6.3.4 Other patterns and effects

Figure 6.14 (top) presents a selection of other patterns for which no real psychophysical data are available. Stimuli (s) to (x), from left to right, concern: (s) the Chevreul step illusion, a rectangular luminance staircase where the visual system amplifies the steps such that over- and undershoots are created (these are *not* Mach bands!); (t) simultaneous brightness contrast in case of disks; (u) assimilation; (v) simultaneous brightness contrast in case of a long horizontal bar; (w) an Adelson tile pattern with luminace gradient and (x) a snake

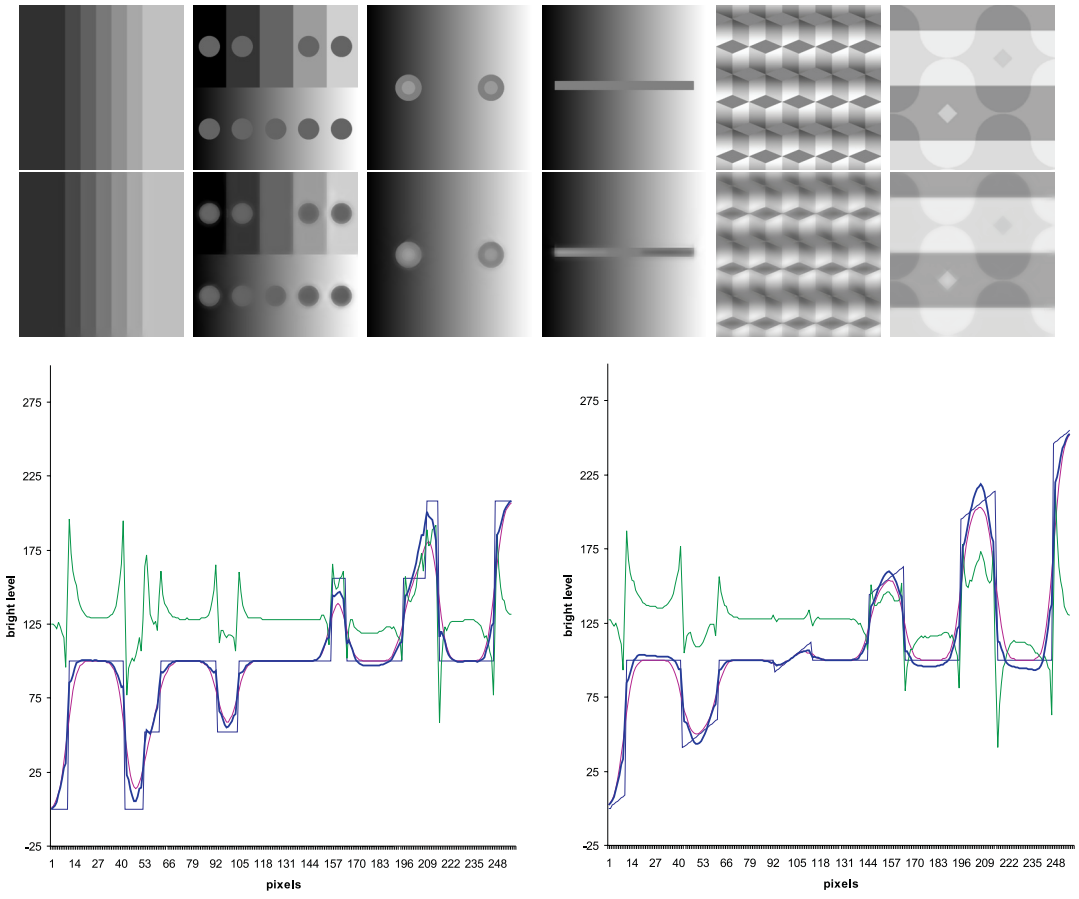

Figure 6.14: Top row, left to right: stimuli (s) to (x). The second row shows 2D model predictions, and the third row 1D cross-sections through the upper and lower disks of stimulus (t).

illusion also from Adelson. The first four patterns are classical textbook examples and the brightness effects are taken for granted. The last two stimuli are more recent, they have been created on the basis of brightness theories, and the two images were extracted and adapted from Logvinenko [2003] (Fig. 1) and Logvinenko and Ross [2005] (Fig. 2). We refer to the latter authors for a detailed explanation of the patterns, but in (w) the horizontal rows of lozenges (the "top faces of the cubes") are all physically equal but three rows appear much brighter than the other three rows, and in (x) the two rotated squares are also physically equal.

The second row in Fig. 6.14 shows the corresponding 2D model predictions. The bottom row shows the model predictions in the form of 1D cross-sections through the centers of the top and the bottom disks of stimulus (t). As can be seen, the latter model predictions are reasonable, and the ones on the linear ramp are better than those on the discrete staircase. In general, model predictions in the case of the other stimuli were better and conform our brightness impression.

The main problem was the detection and modeling of relatively small disks and annuli in the case of stimuli (t) and (u). Again, this problem is due to the "imposed" model limitations, i.e., using only eight orientations (Gabor filters) and only six central scales for event detection, with a few additional scales around the central ones to guarantee event

stability. Even after applying other stability criteria (see Section 6.2), a few incorrect events appeared around disks and a few correct ones disappeared. Only with many more scales (central, plus additional ones for stabilization) the scale space will be really covered, and a better image (re)construction could be achieved. We recall that the "imposed" model limitations were due to the trade-off between CPU time to process the data, the necessary storage capacity for the number of scales etc. and the total number of input images (stimuli).

Logvinenko and Ross [2005] discussed the possibility that in the case of stimuli (w) and (x) classical simultaneous brightness contrast is involved, like in the case of stimulus (t), or that perhaps also different visual phenomena may play a role. Now, after presenting all model predictions, we might say yes and no: both SBC and most if not all other effects are caused by balanced contributions of a lowpass background channel plus multi-scale line and edge representations that capture local detail but also more distant neighborhood influences.

6.4 General discussion

In this chapter we presented a 2D calibrated brightness model—a first 2D version—based on the summation of stable, symbolic line/edge representations with a lowpass channel. Although being a first version without any sophisticated processing, the model was able to produce correct predictions for a large number of brightness illusions, both standard and variations: Mach bands, White's effect, Howe's and Anderson's patterns, simultaneous brightness contrast, assimilation, the Craik-O'Brien-Cornsweet illusion and Logvinenko's versions of Adelson's tile patterns.

A very interesting feature of the brightness model is the principle on which it is based—the visual (re)construction process uses the same information that was used in previous chapters to solve other problems: object segregation (Chapter 4 or Rodrigues and du Buf [2006a]), object categorization (Chapter 5 or Rodrigues and du Buf [2006a]), object recognition (Chapter 5 or Rodrigues and du Buf [2006b]) and disparity estimation (Chapter 4 or Rodrigues and du Buf [2004b]). Indeed, the main conclusion may be that brightness perception and the other processes are in fact part of one integrated process, and that brightness is a relatively simple, fast and data-driven process, which may make many other theories obsolete.

Despite the large number of psychophysical data sets that the model can simulate, both quantitatively and qualitatively, there exist more data sets covering variations of those presented in this chapter that can be tested. This chapter not only deals with a first model version but also with first—and perhaps most important—data sets. Two major and pertinent problems were encountered when we attempted to test and integrate more data: (1) Many data sets are described using different parameters and presentations and in some cases parameters are missing and the experiment is difficult to replicate. (2) The excessive CPU time necessary to process the data and the storage (disk space) needed for each simulation. The last point mainly reflects the number of stimuli that can be used for model calibration, because these must be available fast to optimize model parameters, for example using a complete set of curves describing the contrast of cosine gratings and the brightness of disks. Once the model has been calibrated it can be applied sequentially to an unlimited number of other stimuli. Nevertheless, using an increased number of scales (a more detailed scale space) will improve model predictions in many of the cases that were tested, but at the expense of even longer CPU times.

Therefore, one of the first future steps will be the optimization and parallelization of

all the processing, from stimulus convolutions with Gabor filters to multi-scale line/edge detection to final brightness prediction. This will allow to use more scales, a better resolution than half a minute of visual angle per pixel, bigger images, but also conceptual ideas like the application of a dynamic scale selection, for instance as a function of the patch size/area. The last point will solve one of the big limitations of the model, namely the limited maximum patch size/area which varies from stimulus to stimulus.

In a later step additional syntactical information can be integrated, like transparency, texture and keypoints (very important for Focus-of-attention [Rodrigues and du Buf, 2006d]) enabling inclusion of 3D structural information to model 3D brightness perception. Also to be integrated in the model is the possibility to distinguish between positive and negative disks, and light and dark Mach bands. Very interesting will be the integration in a single model of supra-threshold brightness perception and threshold detection. Until now, the modeling of threshold detection is being done in parallel [du Buf, 1992b; Bobinger and du Buf, 2002].

The bottom line is that we are still looking for that one and very unique model which can explain *all* data sets. This becomes even more complicated if one of the goals is that that same model—or at least its basis—should at the same time be able to cope with other functions, such as invariant object recognition.

Chapter 7

Application: painterly rendering using human vision

Abstract: Painterly rendering has been linked to computer vision, but we propose to link it to human vision because perception and painting are two processes that are interwoven. Recent progress in developing computational models allows to establish this link. We show that completely automatic rendering can be obtained by applying four image representations in the visual system: (1) color constancy can be used to correct colors, (2) coarse background brightness in combination with color coding in cytochrome-oxidase blobs can be used to create a background with a big brush, (3) the multi-scale line and edge representation provides a very natural way to render finer brush strokes, and (4) the multi-scale keypoint representation serves to create saliency maps for Focus-of-Attention, and FoA can be used to render important structures. Basic processes are described, renderings are shown, and important ideas for future research are discussed.

7.1 Introduction

We combine two different research areas in this chapter: non-photorealistic rendering (NPR), in particular painterly rendering of images or photographs by means of discrete brush strokes; and visual perception, in particular computational models that have been—or are being—developed for color and brightness perception and Focus-of-Attention. Painterly rendering has been combined with computer vision [Gooch et al., 2002; Kovács and Szirányi, 2004; Shiraishi and Yamaguchi, 2000], but, in our view, it should be linked to human vision because painting is intrinsically related to the way we perceive the external world, i.e., real scenes or photographs.

NPR is a research area which aims at transforming input images (photographs, data) into output images which resemble the input, but an artistic impression can be created. For a recent NPR taxonomy and survey of stroke-based rendering techniques see [Hertzmann, 2003; Sousa, 2003]. In this chapter we address *painterly rendering* in the sense of Hertzmann [1998], i.e., completely automatic creation of paintings by using brush strokes of different

sizes. Since painters often select fine brushes in regions with important detail structures, we can combine painterly rendering with *stylization and abstraction* as introduced by DeCarlo and Santella [2002], but with automatic selection of the regions. In addition, since painters often exaggerate colors, we apply a method for transforming dull colors into more vivid ones.

The goal of this chapter is to show how we can apply models of visual perception. Painters have learned to observe and to select important structures to be painted, and do this quasi-automatically or intuitively. Rendering schemes must be developed that do the same, which requires good insight into our visual system and processes in the visual cortex. In this chapter we illustrate painterly rendering based on cortical image representations. In view of the many aspects involved, like disparity (stereo, depth), motion and texture perception, we will concentrate on only three aspects for which state-of-the-art models are available: color constancy, brightness perception on the basis of multi-scale line/edge representations, and saliency maps for Focus-of-Attention based on keypoint representations. Future extensions may involve texture perception for rendering realistic textures, simulating the hand of a painter while painting fine, repetitive textures, and motion perception for creating animations from video; see also Discussion.

This chapter is organized as follows: Section 7.2 introduces basic concepts, brightness perception and line/edge interpretation, Focus-of-Attention and keypoints. Section 7.3 deals with color constancy. Rendering procedures are explained and illustrated in Section 7.4, and we conclude with a discussion in Section 7.5.

7.2 From perception to rendering

In the present and following sections we will explain a few processes and show how perception models can be employed in NPR. These models cover four aspects:

1. Brightness perception based on the multi-scale line/edge representation provides a very natural way for painterly rendering [Hertzmann, 1998], with brush sizes that are coupled to the scale of image analysis, from coarse (global structures) to fine (local detail). Many painters start with big brushes in order to create a background with coarse regions, after which smaller brushes are used to paint small structures. Here we will not go into special techniques like *clair-obscur* and will only apply *wet-on-dry*, i.e., a new (wet) brush stroke will substitute previously applied (dry) strokes.

2. A global background level of brightness cannot be coded by cortical simple and complex cells if these are modeled by bandpass Gabor filters with a zero—or very small, residual—DC component. However, apart from the rods and cones, the common photoreceptors, retinal ganglion cells have been identified that have no (in)direct connection to rods and cones; their dendrites act as photoreceptors [Berson, 2003]. These ganglion cells transfer global luminance information to central brain regions, for controlling the circadian clock (solar day) and the eyes' iris muscles (pupil size). Because of their projections on ventral and dorsal areas of the LGN (lateral geniculate nucleus), it is assumed that they also project on the cytochrome-oxidase (CO) blobs in the cortical hypercolumns, where color information from the retinal cones is processed [Hubel, 1995]. This way, color coding in CO blobs is complemented with a global brightness level, and this representation will be used to paint a background with very large brush strokes. We will not devote a separate section to this aspect, because the solution is extremely simple; see Section 7.4.

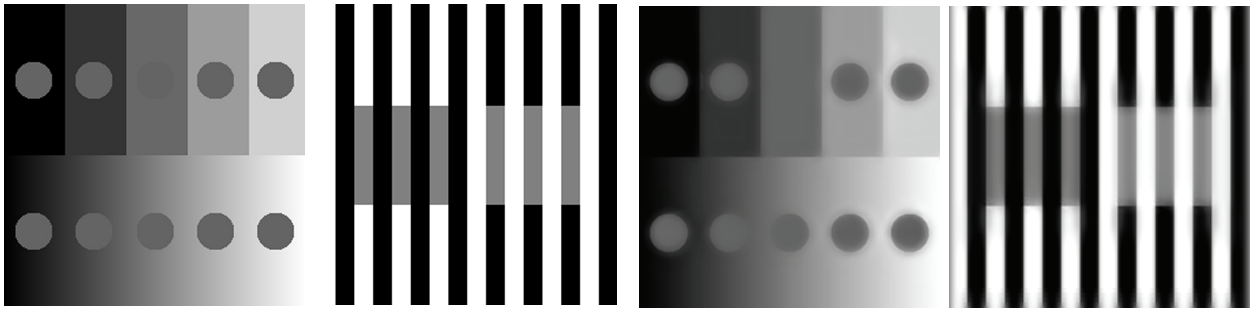

Figure 7.1: Brightness induction effects (left) and model predictions (right).

3. Color constancy, which means that the colors of a scene or image are perceived with little influence of the illumination spectrum (colors of light sources), can be obtained by applying the retinex theory [Land and McCann, 1971] or the more recent ACE (Automatic Color Equalization) model [Rizzi et al., 2003]. The ACE model can be applied to create more vivid colors, an effect that can be observed in many paintings.

4. Focus-of-Attention by means of the multi-scale keypoint representation and saliency maps can be used to guide the rendering such that line/edge-based brush strokes are only applied in image regions with a certain complexity. This aspect is related to image stylization and abstraction [DeCarlo and Santella, 2002], but automatic generation of saliency maps eliminates the need to record eye movements, i.e. fixation points, of persons who are actually looking at the image to be rendered.

There is a relation between color constancy and brightness perception. However, this relation is very complex and not sufficiently covered in the literature concerning computational models. Therefore we will not go into detail. Another aspect will be ignored too: DeCarlo and Santella [2002] employed the contrast sensitivity function (CSF) of sinewave gratings to suppress fixation points in regions with low contrast. We will not do this for two reasons: (1) A CSF depends on retinal eccentricity, the size of the gratings and background luminance, and can only be measured under very controlled experimental conditions and with very trained observers. (2) The sinewave CSF may *only* be used in the case of sinusoidal patterns. CSFs in the case of other gratings (squarewave, trapezoidal) are different, and all CSFs of periodic gratings differ completely from threshold curves of aperiodic patterns like blobs or discs [du Buf, 1992b]. Nevertheless, in NPR it may make sense to suppress very low-contrast patterns, but a unified detection model that can be applied to all patterns remains a *Holy Grail* of visual psychophysics [Bobinger and du Buf, 2002; du Buf, 2005].

Below, we illustrate the application of the perception models, without going into the mathematics of the models, but with first results created by painterly rendering. In the next two subsections we review briefly the brightness model, keypoints and FoA (for details see Chapter 3, 4 and 6), and in Section 7.3 the ACE model.

7.2.1 Lines, edges and brightness

Figure 7.1 shows two brightness induction effects, i.e., simultaneous brightness contrast (left) and assimilation (right); see Chapter 6 for more details. All nine circles have the same level of grey (pixel value), as have the six grey bars, which means that under homogeneous illumination they should appear equal in brightness. However, it can be seen that in simultaneous

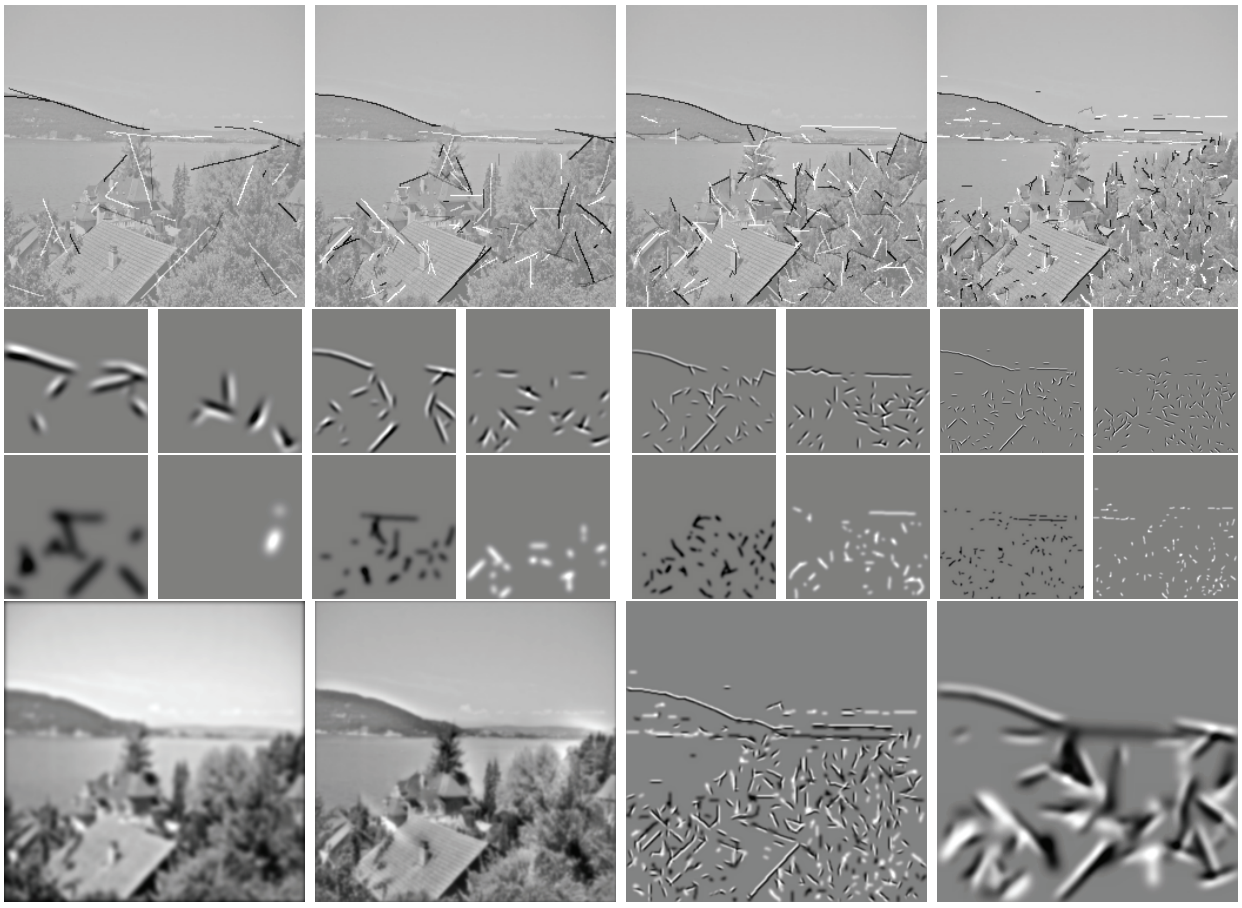

Figure 7.2: Top row: Event maps at four scales. Middle two rows: symbolic line/edge interpretations at the four scales. Bottom row: visual (re)construction (2nd) is based on lowpass filtering (1st at left) and summed line/edge interpretations (3rd at fine and 4th at coarse scale).

brightness contrast the background pushes the brightness of the circles in the opposite direction: grey circles against a dark background become brighter. On the other hand, in assimilation the black flanking bars pull the brightness of the grey bars in the same direction. Figure 7.1 shows correct model predictions.

The multi-scale symbolic line and edge interpretation (see Section 4.3), in combination with lowpass information—the special retinal ganglion cells with photoreceptive dendrites—forms the basis for the brightness model predictions (we refer to Chapter 6 for details) shown in Fig. 7.1 and image (re)construction shown below.

Figure 7.2 illustrates visual (re)construction in case of the "lake" image, with event maps at four scales at the top, i.e., detected positions of positive and negative lines and edges, coded by grey level and superimposed on a low-contrast version of the input image (the "lake" input image is shown in B/W in Fig. 7.3 and in color in Figs 7.4 and 7.9 (top)). The actual detection process, with solutions to solve problems related to stability, completeness and curvature, has been described in Chapter 4 and in Rodrigues and du Buf [2005a]. The middle two rows in Fig. 7.2 show symbolic line/edge interpretations at the same four scales. The bottom row illustrates the (re)construction process by summing lowpass information (1st image) and combined interpretations with real event amplitudes at the four scales (only

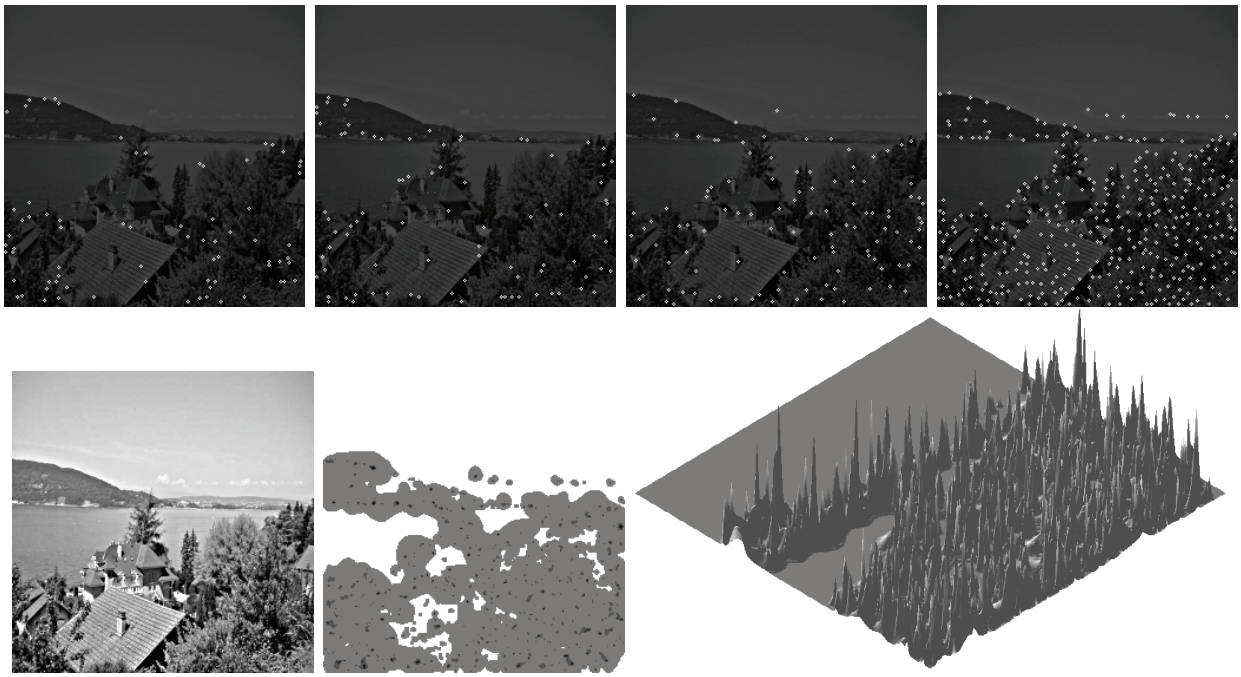

Figure 7.3: Top: keypoint maps at the four scales. Bottom: input image (B/W) and a saliency map for FoA.

two scales are shown, 3rd and 4th) and the result (2nd). Using more filter scales leads to better (re)constructions and the relative weighting of all information with a dynamic selection of scales is still under investigation. The basic idea for NPR follows from Fig. 7.2: event positions form strokes and the size of the profiles determines brush size, in the case of edges two parallel brush strokes.

7.2.2 Keypoints, saliency and FoA

Responses of end-stopped cells in V1 are very fuzzy and require optimized inhibition processes in order to detect, with high precision, keypoints at singularities like edge crossings; see Chapter 3. Figure 7.3 (top) shows detected keypoints in the case of the same image (Fig. 7.9 (top)) and at the same scales as used in Fig. 7.2. In general, keypoints are stable over certain scale intervals: over fine scales at fine image detail, over medium scales at coarser structures, etc. For a detailed analysis of keypoint behavior in scale space we refer to Chapter 3. Figure 7.3 (bottom) shows, apart from the input image (B/W), a saliency map as a normal 2D image and in projected 3D view.

Because saliency maps code complexity in terms of position and RoI (Region-of-Interest), such maps provide ideal information for Focus-of-Attention, a process used to steer our eyes and mental attention [Parkhurst et al., 2002]: to plan saccadic eye movements between fixation points (the peaks), and during fixation the stable information in the RoI (lines/edges, disparity) has time to be processed in V1 and, also during the next saccade, to propagate to area V2 and higher cortical areas. Here our point is that saliency maps can be used in NPR in order to control the rendering, using brushes of a certain size only inside the RoIs around the peaks. This eliminates the need to record eye movements [DeCarlo and Santella, 2002].

Figure 7.4: Colour correction by the ACE model (at right).

7.3 Color constancy

Color constancy is the effect that the colors of a scene or image are perceived with little influence of the color of the light source. The retinex theory [Land and McCann, 1971] explains this by assuming that local color depends on the surrounding regions or even on the entire image. The recent ACE (Automatic Color Equalization) model [Rizzi et al., 2003] achieves this in two processing steps in the RGB channels: first, chromatic and spatial adjustment of each pixel is applied by subtracting R, G and B values of all pixels, thereby employing a nonlinear saturation function to the differences and weighting the contributions of all pixels by a Euclidean distance function. Second, dynamic tone reproduction scaling serves to rescale, linearly but with clamping, the R, G and B values of all pixels to use the available range, normally from 0 to 255 (8 bits) in 24-bit images. Figure 7.4 shows the "lake" and "seals" images before (left) and after (right) applying the ACE model. As can be seen, contrast and color ranges have been stretched, and the colors are more balanced. This effect is ideal for automatic, unsupervised NPR on the basis of unprocessed photographs (we note that contrast stretching, although not on the basis of a perception model, can be easily achieved by packages like Adobe's Photoshop, Corel's Paint Shop Pro and GNU's GIMP).

7.4 Painterly rendering

The models described above are employed in a sequence of processing steps. Two steps can be seen as options: (1) if the colors of an image are very pleasing because they convey a certain mood, for example a sunset with reddish colors against a dark landscape and sky, the ACE model may change the mood drastically and it may be better to skip color equalization. (2) Stylization by means of saliency maps leads to less brush strokes in regions where there are no or few keypoints, for example along long edges between homogeneous regions. Application of stylization may therefore suppress important structures and the user can decide to experiment with and without stylization. Below we will illustrate a few effects and option selections, but the normal procedure is the following:

First, the ACE model is applied to the input image. Second, a background image is created. In order to obtain the effect of a painted background, a large brush size of 16 by 32 pixels, for example, is applied: (A) a position and an orientation are selected randomly. (B) Three colors are picked in the ACE image, at the actual position (the center of the stroke) and at the two end points. (C) If the color at one of the end points deviates significantly from the average of the other two colors, the stroke will not be applied. The reason is simple:

Figure 7.5: Created background images: lake (left) and seals (right).

the orientation of this stroke is such that it covers two distinct regions, and a wrong color would be introduced in one region in which no line/edge information may be available. This color may not be corrected by line/edge-related brush strokes, for example in homogeneous regions like the sky or in water. (D) If the three colors do not deviate significantly, they are averaged, the stroke will be painted, and we continue with step A. (E) This procedure is repeated until the entire image (all pixels) has been covered, counting the number of painted strokes, and then the same number of strokes will be applied again using the same procedure. This repetition with random positions and orientations will cover many previously painted strokes such that only parts remain visible, which yields more realistic results. Figure 7.5 shows created backgrounds in the case of the color-corrected "lake" and "seals" images shown in Fig. 7.4. The strokes which were painted last are clearly visible because they are complete.

Finally, detected lines and edges (event type, position, scale) are rendered, scale by scale, by continuous brush strokes with colors that are picked in the color-corrected image. This is accomplished in seven steps:

1. Detected events are checked for continuity, separated, and continuous sequences of (x, y) positions are stored in lists. Lists that contain more than 16 positions are divided into separate lists. If there are a few more positions, like 18, we keep the entire list, but a list of 21 will be separated into two lists of 10 and 11. This is done to prevent rendering very long strokes with one and the same color, which are often detected at coarse scales. In the future, the number of positions (16, 18) will be varied such that lists (strokes) will be shorter at fine scales.

2. Each list is filtered in order to transfer discrete lattice positions into smooth strokes with coordinates in floating point. The filtering applied consists of iteratively moving each point in the direction of the line defined by the point's four neighboring points, two on each side, until movements become very small [Nunes et al., 2005]. As a result, strokes are slightly "randomized," as if they were painted by hand.

3. The center point of each list is determined by counting. If this point is within a RoI of the saliency map at the list's scale, the entire list will be rendered as one stroke. If

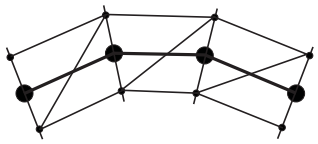

Figure 7.6: Polygon (triangle) construction using lines perpendicular to average left-right slopes.

not, a new list will be processed. As mentioned above, the selection can be skipped if stylization leads to incomplete structures.

4. Through each center point, a perpendicular line is computed on the basis of the average of the slopes of the two lines that connect the point with its two neighbors, see Fig. 7.6.

5. Point pairs are used to create polygons by (a) using the perpendicular lines and (b) a distance to the points that depends on brush size. In the case of rendering an edge, two but connected polygons are created, one on each side.

6. Colors are picked in the color-corrected image: a line stroke is rendered by averaging the colors at all the list's positions, and an edge stroke by two colors that are the averages of the colors at symmetric positions off the center list on the perpendicular lines.

7. All polygons (or triangles) are rendered using OpenGL, with texture (opacity) mapping in the alpha channel. Below we will illustrate two mappings: gradual opacity maps for creating clean, ellipsoidal-like strokes (as were used in Fig. 7.5), and real, digitized strokes composed of randomly-selected heads, bodies and tails.

Figure 7.7 shows discretized and painted strokes at four scales and Fig. 7.8 the combination with the background, using coarse-to-fine-scale color replacement which simulates painting wet-on-dry (at the moment, all individual brush strokes replace previously rendered positions). It can be seen that adding finer scales leads to many local corrections, i.e., to more realistic detail. Final results are shown in Fig. 7.9 in case of the lake (top) and the seals (bottom) images. These results were obtained by using only five scales, from medium to fine, and clean, artificial strokes. The middle images were created without saliency maps, the right ones with saliency maps. The difference is clearly visible in the case of the seals image, but not in the case of the lake image (only at two positions at the edge between the sky and the landscape).

Figure 7.11 shows the background (top-left) and final result (bottom-right) in the case of the lake image when rendered with real, digitized strokes (compare with Figs 7.5 and 7.9 (top)). These strokes are composed of randomly-combined heads, bodies and tails from digitized oil-painted strokes, see Fig. 7.10, which are used to control color opacity. As can be seen, the use of real brush strokes changes our impression of the "paintings." Figure 7.12 shows more examples, i.e., with real (top) and "clean" (middle and bottom) brush strokes. Finally, Fig. 7.13 shows results without (top) and with the application of color equalization. Two different parameter selections (minimum contrast, middle; maximum contrast, bottom) of the ACE model yield more vivid colors, for example the blue in the sky, but the mood of the reddish sunset has been lost. This is an example of a case where it may be better to *not* apply color equalization.

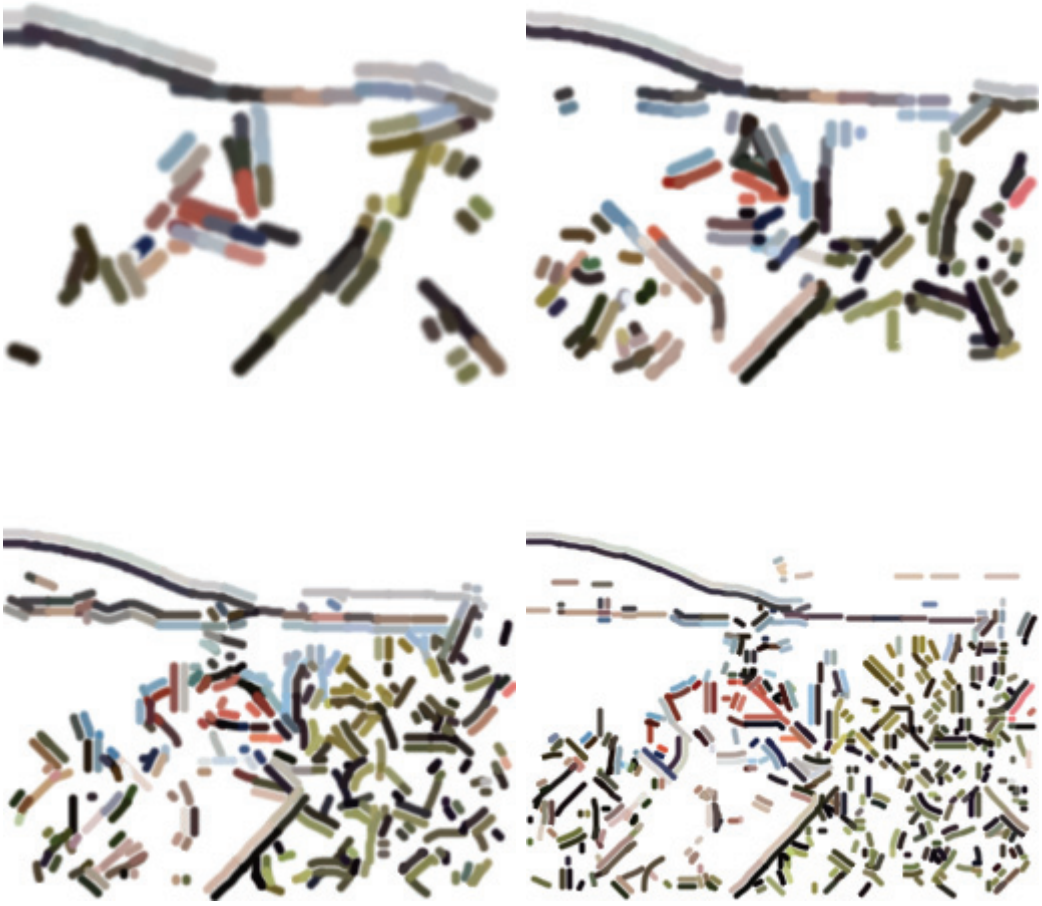

Figure 7.7: Discrete brush strokes at the four scales.

7.5 Discussion

We have seen that perception models can provide a solid basis for NPR, specifically for painterly rendering. Colors can be corrected to cover the available dynamic range, such that unprocessed photographs are optimized and colors become more vivid. Coarse brightness and color coding can be used to create a painted background with very large brush strokes, thereby avoiding painting across edges with different colors. The symbolic line/edge interpretation translates directly into brush strokes, and adding more scales (finer brushes) leads to more realistic renderings. Saliency maps can be used to steer the rendering by painting lines and edges only in regions with sufficient complexity. As mentioned in the Introduction, our goal is *completely automatic* painterly rendering by using human vision. Although all processing steps can be selected, there are a few options that the user can (de)select. The most important options, with possible solutions or alternatives, are:

1. Color correction by the ACE model often leads to brighter, more vivid colors, especially after experimenting with ACE's own parameters (different saturation and distance functions), but sometimes this effect is not desired. If the colors of the input image convey a desired mood or effect, for example, the ACE model can be skipped. Auto-

Figure 7.8: Rendering coarse to fine scale.

matically applying ACE or not, or selecting its parameters, requires *a priori* knowledge or years of research into local/global color histograms and their relation to subjective interpretations of many observers. The user is always free to edit the color gamut with standard techniques, like decreasing saturation in the case of preparing a watercolor rendering (transfer RGB space to HSV, reduce saturation and then transfer back to RGB).

2. Stylization by selecting brush strokes on the basis of keypoints and their ROIs may interrupt long lines and edges if no vertices or other structures like textures are present, and the painted background may be too random to convey continuity. The approach presented here is only one of many possibilities, because saliency maps at different scales can be used or combined in other ways. In addition, postprocessing of detected lines, edges and keypoints could be used to detect long and important structures. This solution might replace the use of image segmentation in conjunction with edge detection [DeCarlo and Santella, 2002].

3. Many line/edge scales can be selected: the more scales are used, the more realistic the rendering will be (from NPR to nPR to PR), which may not be desirable. Our

Figure 7.9: Input images (left) and rendering without (middle) and with (right) saliency. Lake (top) and seals (bottom).

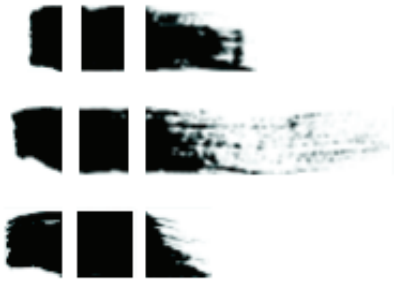

Figure 7.10: Examples of heads, bodies and tails.

impression is that, apart from the big brush strokes to render the background, only very few scales may lead to best results, but this requires many experiments with many images and renderings and subjective ratings of many observers. In addition, in order to improve the painterly effect, the few selected scales could be applied two or more times by overpainting already painted strokes using interpolations of stroke lists. Although such a process better resembles the technique applied by real artists, finer brushes may be necessary in some cases, for example when painting (small) faces. The multi-scale keypoint representation has already been used to detect faces, see Chapter 3 or Rodrigues and du Buf [2005c], and detected faces and their positions plus sizes could be used to control scale selection, but many other objects may require

Figure 7.11: Rendering with real brush strokes.

the same processing. This aspect is perhaps the most difficult one, because it relates to cognitive image interpretation and distinction between important and unimportant objects to be painted or to be left out.

Renderings shown here are first results produced to test the methods in an integrated framework. We intentionally did not present results obtained with different styles, in order to illustrate possibilities but also problems of the basic approach. Experiments with different styles and comparisons with renderings based on computer vision are beyond the scope of this chapter. Apart from improvements and future research as mentioned above, there are many other possibilities that can be explored to enhance artistic effects: (1) to test different ways of picking colors, for example only at one (two) central position(s) of a stroke, instead of by averaging colors under an entire stroke as done here, (2) to discretize and/or randomize positions, orientations, lengths and colors, for simulating styles like impressionism, expressionism or cubism, and (3) to replace the brush strokes by a library of strokes created by using different brush types, e.g. [Wang et al., 2004]. The latter aspect includes mixing painting wet-on-dry and wet-in-wet for simulating different types of oil paintings and watercolors, in addition to using opacity maps—possibly also bump or normal mapping—for simulating the texture of a canvas or paper in combination with different media like oil, crayon and charcoal.

A bigger challenge is to find new applications that may foster new styles in contemporary visual arts. An example is *symbolic pointillism* that exploits Gestalt laws of cognitive psychology [Krüger and Wörgötter, 2005]. Instead of combining segmentation and single-scale edge-detection methods in stylization and abstraction [DeCarlo and Santella, 2002], our line

Figure 7.12: Top row: using real brush strokes. Middle and bottom rows: using clean brush strokes.

and edge representation could be used for the same purpose, i.e., the creation of cartoon-like effects. In addition to rendering important edges on top of large, homogeneous regions, regions could be hatched (textured) such that an impression of 3D shape is obtained. This requires a cortical model for solving the shape-from-shading problem, which could be combined with a disparity model in the case of using a stereo camera, and could lead to automatic production of etchings. Such an approach can complement *suggestive contours* based on radial curvature, in both still images and animations [DeCarlo et al., 2004]. Texture segmentation is often based on complex Gabor filters (simple cells). For example, analysis of the symmetry order for separating linear, rectangular and hexagonal patterns can be achieved by complex moments but also by a simple cortical model [Bigun and du Buf, 1994; Santos and du Buf, 2002]. Within segmented regions, the symmetry order along with detected orientations at

Figure 7.13: Without (top) and with ACE.

different scales can be used to render realistic textures [Wang et al., 2004]. A good model of motion perception could be used to solve problems in creating animations from video. Instead of rendering frame by frame, effects of moving objects and a panning and tilting camera can be combined to create a stable background, adding moving objects and new information at the image borders in a consistent way [Collomosse et al., 2005]. Additional perception models and rendering techniques may be employed in the future, because *we just started looking through the eyes of the painter!*

Chapter 8

Concluding remarks

Abstract: This chapter summarizes the previous chapters and presents directions for further research.

8.1 Summary

After a short introduction (Chapter 1), a brief overview of the cortical aspects of vision with a special focus on the architecture and functionality was given in Chapter 2. How objects can be stored in memory and how invariance can be achieved were particularly important aspects.

In Chapter 3 the multi-scale keypoint representation was studied, and it was shown that this representation provides very important information for object (and face) detection. End-stopped cells, which combine outputs of complex cells tuned to different orientations, were used to detect line and edge crossings, singularities and points with large curvature. These cells can also be used to construct retinotopic keypoint maps at different spatial scales. Different grouping operators can be used for object segregation and automatic scale selection. Saliency maps for Focus-of-Attention can be constructed, and such maps have been employed for detection of facial landmarks (eyes, nose and mouth) and thus faces.

In Chapter 4 an improved scheme for line and edge detection on the basis of responses of simple and complex cells was presented. This scheme is truly multi-scale with no free parameters. Also illustrated were automatic scale selection, visual (re)construction, and object segregation by coarse-to-fine-scale groupings. Two-level object categorization scenarios were tested in which pre-categorization was based on coarse scales only, and final categorization on coarse plus fine scales. A multi-scale face (object) recognition model based on the line and edge representations was tested. In addition, a new disparity model was proposed. This model allows to directly attribute depth information to detected lines and edges.

Chapter 5 introduced a novel method for obtaining 2D translation, rotation and size invariance, and partial and global saliency maps for face recognition were tested. It was shown that dynamic routing allows feature maps of input objects and of stored templates to converge. All the feature extractions and processing schemes from Chapters 2 to 5 were combined into an integrated cortical architecture.

Chapters 2 to 5 mainly covering object recognition, in Chapter 6 the scope was shifted to the modeling of brightness perception. An existing 1D brightness model was extended to 2D. The model is also based on the multi-scale symbolic representation of lines and edges, but with an additional low-pass channel and nonlinear amplitude transfer functions. The scheme was calibrated with psychophysical data, and it was shown that it can predict effects such as Mach bands, the White effect, simultaneous brightness contrast, grating induction and the Craik-O'Brien-Cornsweet illusion.

In Chapter 7 it was shown that one application can combine most processing described in the previous chapters. Painterly rendering has been linked to computer vision, but we proposed to link it to human vision, because perception and painting are two processes which are interwoven. It was shown that completely automatic rendering can be obtained: (1) by applying a model of color constancy, (2) coarse background brightness in combination with color coding in cytochrome-oxidase blobs can be used to create a background, (3) the multi-scale line and edge representation provides a natural way to render brush strokes in the foreground, and (4) saliency maps can be used to render important structures. In the meantime, new developments and results have been achieved (see Nunes et al. [2006a]), but these are beyond the focus of this thesis.

8.2 Achievements

There are three major achievements. These are summarized in this section and they are the basis for future research that will be presented in the next section.

(i) New insight into object recognition embedded into a cortical architecture. The invariant object recognition architecture can be characterized by two major aspects: (a) it is biologically plausible and based on multi-scale features extracted in V1 or beyond, and (b) it is based on explicit feature extractions in which cells have clearly defined functions, i.e., we know precisely how the output is obtained from the input, instead of having some "neural network" between input and output neurons with a few hidden layers.

Related to this achievement (but also to the other two) are a number of smaller contributions: (1) single (fine) scale keypoint detection combined with NCRF inhibition, extended with a vertex structure detection scheme; (2) multi-scale keypoint representation based on the responses of end-stopped cells; (3) object segregation based on the keypoint representation; (4) automatic scale selection; (5) saliency maps for Focus-of-Attention; (6) facial landmark detection using partial saliency maps and keypoints, plus a first application of partial and global saliency maps in face recognition; (7) a new multi-scale line and edge detection method, with no free parameters, based on the responses of simple and complex cells; (8) a visual (re)construction method, although still under investigation; (9) object segregation and (10) automatic scale selection using the line/edge representation; (11) an invariant categorization and recognition model based on the peaks in the saliency maps and on the multi-scale line and edge information; and (12) a proposal for a new disparity model.

Some aspects like the disparity model need much more research, but most can be seen as building blocks that can be integrated into future architectures. New and improved architectures must be developed in order to obtain better results, although the results already obtained with object and face detection, categorization and recognition are quite good.

(ii) A brightness model extended to 2D. It was the first time that a brightness model (1D or 2D) was calibrated using psychophysical data and that the 2D model could account for most known brightness illusions and real psychophysical data. The original 1D

model [du Buf and Fischer, 1995] could already account for a number of different effects but was not tested against real data. Perhaps more important is the fact that the image (re)construction model is based on the same cortical representation, i.e. object recognition and brightness perception have been shown to be related processes.

(iii) Painterly rendering using human vision. A new line of research was proposed and tested. Several authors already sporadically applied aspects of human vision in the artificial rendering of pictures, such as FoA [DeCarlo et al., 2004] or Gestalt laws in *symbolic pointillism* [Krüger and Wörgötter, 2005], but it was the first time that NPR was applied exclusively on the basis of models of human vision. This opens perspectives in empirical aesthetics, i.e. the perception of art, at the moment at low level (manipulations of brush strokes) but later undoubtedly also at medium and high levels (abstraction and composition, respectively).

8.3 Directions for further research

After all the research described in this thesis and the 3+ years it took, and despite the many achievements, toward the end it became clear that the entire processing as described only represents one data stream of two (or more) streams. Object segregation for categorization and recognition requires prior knowledge: the chicken-or-egg problem. This problem can only be solved by modeling a very fast gist system, not only a global scene gist system but also a local one together with spatial layout. Such a system must be based on very simple features which provide crucial information for entire objects: texture, color, motion and disparity. This information can be fed into a relatively simple neural network with some learning rule. For example, normally tree trunks are elongated, vertical and brown, whereas tree crowns are round, irregularly textured and green. Texture and color may be sufficient to detect trunks and crowns, but motion and disparity may indicate that one trunk belongs to one crown and that the whole thing is a tree, for example in the case when viewing a tree from a moving car such that the tree moves relative to the background. Such a local/global gist system can be really fast because of a few feature extractions (performed in parallel) and a feed-forward trained network, and it can therefore "bootstrap" the data stream described in this thesis, for example by biasing likely object templates in (associative) memory and preparing rough position estimates for segregation.

Instead of making a long list of topics which may or must be considered in future research, we will finish by mentioning a few research projects which involve collaborations at national and international levels. Developing a fast gist system for robotics, as decribed above, is the core of a European FP7 project "RoboGist" involving five partners. Two national projects have been submitted for support by the Portuguese Foundation for Science and Technology (FCT):

(1) "SmartVision: active vision for the blind." Continuing the development of the disparity model, which is an almost completely unexplored research topic in human vision, it will be possible to integrate depth information and the multi-scale line/edge and keypoint representations. The same applies to models of "texture cells" for extracting 3D gradients from curved object surfaces, which completes disparity information. The integration of all information sources is very important for developing a stable and reliable aid for the blind, i.e. for in- and outdoor navigation with obstacle avoidance and recognition of landmarks (zebra crossing, bus stop) and objects on a bathroom shelf or in a kitchen cupboard. A secondary problem which can be explored is the plasticity of the neural architecture, i.e.

the change of dendritic connections (weight factors) between different layers for dealing with invariance, with special focus on complex backgrounds or natural scenes, in relation to the minimum number of required views of objects in visual memory.

The modeling aspects described above can be complemented with psychophysical (eye tracking) and fMRI/EEG experiments to address two problems: (a) the optimization of the FoA model, and (b) further refining the cortical architecture. The latter is composed of many "components:" the what and where subsystems, information propagation from V1 to higher areas, coarse and fine scales, and last but not least the information flows in time. Further fMRI experiments may shed light on what actually happens in our brain.

(2) "Abstraction and color emotions in the perception of paintings." This is a continuation of the development of the painting software that semi-automatically converts an input photograph into a painting with discrete brush strokes. The software consists of two parts, the first one for image analysis using models of cortical processing, and the second for the graphics in which detected lines and edges at a certain scale are broken up into small coordinate lists and brush strokes are rendered using triangulation and texture mapping. Many more options will be implemented, such that the program is able to simulate most techniques as used by real painters. Direct extensions concern brush types (wax crayon, pastel, etc.), stroke types (pointillism, linear, curved, sigmoidal, etc.) and general styles (hatching of regions using pencils, realistic watercolor effects). Specific results concern optimized color shifts related to emotions (cold/warm, hard/soft, etc.). The goal is that the painting software will be of professional quality and can be made available to collaborating researchers specialized in empirical aesthetics.

Again, the modeling aspects mentioned above can be complemented with various psychophysical experiments concerning e.g. color (hue) shifts related to emotions cold and warm, and possible differences between men and women, between art lovers and ignorants. Eye-tracking experiments involve input images, but also paintings based on these with different levels of abstraction, even reproductions of real paintings. General trends of scan paths and fixation points are likely a function of the level of abstraction, and fMRI experiments may reveal the influence of abstraction on the activities of certain brain regions which are involved in abstract reasoning.

The projects and aspects mentioned above are selections of many possibilities or doors which have been opened by the research reported in this thesis. The study and modeling of a cortical architecture remains a continuing goal, even after the state-of-the-art has advanced: more cells with additional functionalities can be integrated, feature extractions can be further refined, small and big building blocks can be moved to other positions with other timings. This thesis is not a final period, not in basic research concerning the architecture nor in all applications that can be envisaged.

"We are continually faced with a series of great opportunities brilliantly disguised as insoluble problems," John W. Gardner, President, Carnegie Foundation.

Appendix A

List of publications

Papers in refereed journals

[1] Rodrigues, J. and du Buf, J.M.H. (2006) Multi-scale keypoints in V1 and beyond: object segregation, scale selection, saliency maps and face detection, BioSystems 86, pp. 75-90. DOI: 10.1016/j.biosystems.2006.02.019.

[2] du Buf, J.M.H., Rodrigues, J., Nunes, S., Almeida, D., Brito, V. and Carvalho, J. (2006) Painterly Rendering Using Human Vision, Virtual - Portuguese Journal of Computer Graphics, Special Edition: Advances of Computer Graphics in Portugal 2005, July 2006, Vol. AICG 2005 , pp. 12.

[3] du Buf, J.M.H. and Rodrigues, J. (2007) Image Morphology: from perception to rendering, Image - Journal of Interdisciplinary Image Science, Special Edition: Computational Visualistics and Picture Morphology, February 2007, Vol. 5, pp. 19.

[4] Rodrigues, J. and du Buf, J.M.H. (2007) A cortical framework for invariant object categorization and recognition, submitted to Cognitive Processing.

Refereed conference proceedings

[1] du Buf, J.M.H. and Rodrigues, J. (2000) Simple brightness models with lowpass and Gabor filters, Proc. 23rd European Conference on Visual Perception (ECVP2000), Groningen, The Netherlands, August 28-30, Perception Vol. 29 Suppl 52c.

[2] Rodrigues, J. and du Buf, J.M.H. (2004) A Vision Frontend With a New Disparity Model, Web Proc. Early Cognitive Vision Workshop (ECOVISION), Isle of Skye, Scotland, 28 May - 1 June (http://www.cn.stir.ac.uk/ecovision-ws/schedule.php).

[3] Rodrigues, J. and du Buf, J.M.H. (2004) Visual cortex frontend: integrating lines, edges, keypoints and disparity, Proc. International Conference on Image Analysis and Recognition (ICIAR2004), Porto, Portugal, 29 Sept. - 1 Oct., Springer LNCS 3211, pp. 664-671.

[4] Rodrigues, J. and du Buf, J.M.H. (2005) Multi-scale cortical keypoint representation for attention and object detection, Proc. 2nd Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA2005), Estoril, Portugal, June 7-9, Springer LNCS 3523, pp. 255-262.

[5] Rodrigues, J. and du Buf, J.M.H. (2005) Multi-scale keypoint hierarchy for Focus-of-Attention and object detection, Proc. 28º European Conference on Visual Perception (ECVP2005), Coruña, Spain, August 22-26, Perception Vol. 34 Suppl, pp. 240.

[6] Lam, R., Rodrigues, J. and du Buf, J.M.H. (2005) Artistic rendering of the visual cortex, Proc. 2º Workshop Luso-Galaico de Artes Digitais (ARTECH2005), Vila-Nova-Cerveira, Portugal, August 27, pp. 65-78.

[7] Rodrigues, J., Almeida, D., Nunes, S., Lam R. and du Buf, J.M.H. (2005) Building the

what and where systems: multi-scale lines, edges and keypoints, Proc. Workshop on Active Vision VII (WAVVII2005), University of Reading, Reading, UK, September 12, pp. 8.
[8] Rodrigues, J. and du Buf, J.M.H. (2005) Improved line/edge detection and visual reconstruction, Proc. 13º Encontro Português de Computação Gráfica (13EPCG), Vila Real, Portugal, October 12-14, pp. 179-184.
[9] Rodrigues, J. and du Buf, J.M.H. (2005) Multi-scale keypoints in V1 and face detection, Proc. 1st International Symposium Brain, Vision and Artificial Intelligence. (BVAI2005), Naples, Italy, October 19-21, Springer LNCS 3704, pp. 205-214.
[10] Lam, R., Rodrigues, J. and du Buf, J.M.H. (2006) Looking through the eyes of the painter: from visual perception to non-photorealistic rendering, Proc. 14th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG2006), Plzen, Czech Republic, January 30 - February 3, pp. 147-154.
[11] Rodrigues, J. and du Buf, J.M.H. (2006) Cortical object segregation and categorization by multi-scale line and edge coding, Proc. International Conference on Computer Vision Theory and Applications (VISAPP2006), Setúbal, Portugal, February 25-28, Vol. 2, pp. 5-12.
[12] Rodrigues, J. and du Buf, J.M.H. (2006) Face segregation and recognition by cortical multi-scale line and edge coding, Proc. 6th International Workshop on Pattern Recognition in Information Systems (PRIS2006), Paphos, Cyprus, May 23-24, pp. 5-14.
[13] Nunes, S., Almeida, D., Rodrigues, J., du Buf, J.M.H. (2006) Object categorisations using templates constructed from multi-scale line and edge representations, Proc. Workshop Visual Categorisation and Image Management Systems, University of Sunderland, UK, June 28, pp. 14-15.
[14] Nunes, S., Almeida, D., Brito, V., Carvalho, J., Rodrigues, J. and du Buf, J.M.H. (2006) Perception-based painterly rendering: functionality and interface design, Proc. Ibero-American Symposium in Computer Graphics (SIACG06), Santiago de Compostela, Spain, 5-7 July, pp. 14-15.
[15] Rodrigues, J. and du Buf, J.M.H. (2006) Face recognition by cortical multi-scale line and edge representations, Proc. International Conference on Image Analysis and Recognition (ICIAR2006), Póvoa do Varzim, Portugal, 18-20 September, Springer LNCS 4142, pp. 329-340.
[16] Rodrigues, J. and du Buf, J.M.H. (2007) Invariant multi-scale object categorisation and recognition, Proc. 3rd Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA2007), Girona, Spain, June 6-8, Springer LNCS 4477, pp. 459-466.

Journal papers in preparation

[1] Rodrigues, J. and du Buf, J.M.H. (2007) Multi-scale lines and edges in V1 and beyond: object reconstruction, segregation, categorization, recognition and disparity.
[2] Rodrigues, J. and du Buf, J.M.H. (2007) Modelling 2D brightness perception from multi-scale cortical line and edge representations.

Bibliography

Afraz, S., Kiani, R., Esteky, H., 2006. Microstimulation of inferotemporal cortex influences face categorization. Nature 442 (7103), 692–695.

Aggelopoulos, N., Rolls, E., 2005. Natural scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene. Europ. J. Neuroscience 22, 2903–2916.

Anderson, B., 2001. Contrasting theories of White's illusion. Perception 30 (12), 1499–1501.

Ban, S., Skin, J., Lee, M., 2003. Face detection using biologically motivated saliency map model. Proc. Int. Joint Conf. Neural Netw. 1, 119–124.

Bar, M., 2003. A cortical mechanism for triggering top-down facilitation in visual object recognition. J. Cogn. Neuroscience 15 (4), 600–609.

Bar, M., 2004. Visual objects in context. Nature Rev.: Neuroscience 5, 619–629.

Bar, M., Kassam, K., Ghuman, A., Boshyan, J., Schmid, A., Dale, A., Hämäläinen, M. S., Marinkovic, K., Schacter, D., Rosen, B., Halgren, E., 2006. Top-down facilitation of visual recognition. Proc. National Academy of Sciences 103 (2), 449–454.

Barth, E., Zetzsche, C., Krieger, G., 1998. Endstopped operators based on iterated nonlinear center-surround inhibition. Human Vision and Electronic Imaging SPIE Vol. 3299, 67–78.

Berson, D., 2003. Strange vision: ganglion cells as circadian photoreceptors. Trends in Neurosciences 26 (6), 314–320.

Bierderman, I., 1987. Recognition-by-components: a theory of human image understanding. Psychological Rev. 94 (2), 115–147.

Bigun, J., du Buf, J., 1994. N-folded symmetries by complex moments in Gabor space and their application to unsupervised texture segmentation. IEEE Tr. PAMI 16, 80–87.

Blakeslee, B., McCourt, M., 1999. A multiscale spatial filtering account of the White effect, simultaneous brightness contrast and grating induction. Vision Res. 39 (26), 4361–4377.

Blakeslee, B., McCourt, M., 2004. A unified theory of brightness contrast and assimilation incorporating oriented multiscale spatial filtering and contrast normalization. Vision Res. 44 (21), 2483–2503.

Blakeslee, B., Pasieka, W., McCourt, M., 2005. Oriented multiscale spatial filtering and contrast normalization: a parsimonious model of brightness induction in a continuum of stimuli including White, Howe and simultaneous brightness contrast. Vision Res. 45 (5), 607–615.

Bobinger, U., du Buf, J., 2002. In search of the Holy Grail: a unified spatial detection model. Proc. 25th ECVP, Glasgow (UK), Perception Vol. 31 (Suppl.), 137.

Born, R., Bradley, D., 2005. Structure and function of visual area MT. Annual Rev. of Neuroscience 28, 157–189, doi:10.1146/annurev.neuro.26.041002.131052.

Bruce, V., Green, P., Georgeson, M., 2000. Visual Perception. Physiology, Psychology and Ecology. Psychology Press Ltd.

Carmi, R., Itti, L., 2006. The role of memory in guiding attention during natural vision. NeuroImage 6 (9), 898–914.

Chelazzi, L., Miller, E., Duncan, J., Desimone, R., 2001. Responses of neurons in macaque area V4 during memory-guided visual search. Cerebral Cortex 11 (8), 761–772.

Churchland, P., Sejnowski, T., 1992. The computational brain. MIT Press, Cambridge, MA, USA.

Collomosse, J., Rowntree, D., Hall, P., 2005. Stroke surfaces: temporally coherent non-photorealistic animations from video. IEEE Tr. Vis. Comp. Graphics 11 (5), 540–549.

Corchs, S., Deco, G., 2005. Feature based attention in human visual cortex: simulation of fMRI data. NeuroImage 21, 36–45.

Cornelissen, F., Wade, A., Vladusich, T., Dougherty, B., Wandell, B., 2006. No fMRI evidence for brightness and colour filling-in in early visual cortex. J. Neuroscience 26, 3634–3641.

Cox, D., Meier, P., Oertelt, N., DiCarlo, J., 2005. 'Breaking' position-invariant object recognition. Nature Neuroscience 8 (9), 1145–1147.

Csurka, G., Bray, C., Dance, C., Fan, L., 2004. Visual categorization with bags of keypoints. Proc. Int. Worksh. Statistical Learning in Comp. Vision, Prague (Czech Republic), 1–16.

Cumming, B., Parker, A., 2000. Local disparity not perceived depth is signaled by binocular neurons in cortical area V1 of the macaque. J. Neuroscience 20 (12), 4758–4767.

DeCarlo, D., Finkelstein, A., Rusinkiewicz, S., 2004. Interactive rendering of suggestive contours with temporal coherence. Proc. ACM/SIGGRAPH-Eurographics NPAR, Annecy (France), 15–24.

DeCarlo, D., Santella, A., 2002. Stylization and abstraction of photographs. Proc. ACM SIGGRAPH02, 769–776.

Deco, G., Rolls, E., 2004. A neurodynamical cortical model of visual attention and invariant object recognition. Vision Res. 44 (6), 621–642.

Deco, G., Rolls, E., 2005. Attention, short term memory, and action selection: a unifying theory. Prog. in Neurobiol. 76, 236–256.

Delorme, A., Thorpe, S., 2001. Face identification using one spike per neuron: resistance to image degradations. Neural Netw. 14 (6-7), 795–804.

du Buf, J., 1987. Spatial characteristics of brightness and apparent contrast perception. PhD Thesis, Technical University, Eindhoven, The Netherlands.

du Buf, J., 1992a. Brightness versus apparent contrast 3: Blurred disks and concentric cosine gratings. Spatial Vis. 6 (4), 265–284.

du Buf, J., 1992b. Modelling spatial vision at the threshold level. Spatial Vis. 6 (1), 25–60.

du Buf, J., 1993. Responses of simple cells: events, interferences, and ambiguities. Biol. Cybern. 68, 321–333.

du Buf, J., 1994. Ramp edges, Mach bands, and the functional significance of simple cell assembly. Biol. Cybern. 70, 449–461.

du Buf, J., 2001. Modeling brightness perception. Chapter in: Vision models and applications to image and video processing. C.J. van den Branden Lambrecht (ed.), Kluwer Academic, pp. 21-36.

du Buf, J., 2005. Modelfest and contrast-interrelation-function data predicted by a retinal model. Proc. 28th ECVP, A Coruña (Spain). Perception Vol. 34 (Suppl.), 240.

du Buf, J., 2007. Improved grating and bar cell models in cortical area V1 and texture coding. Image and Vision Comput. 25 (6), 873–882.

du Buf, J., Fischer, S., 1995. Modeling brightness perception and syntactical image coding. Optical Eng. 34 (7), 1900–1911.

Ekenel, H., Sankur, B., 2005. Multiresolution face recognition. Image and Vision Comput. 23 (5), 469–477.

Elder, J., Sachs, A., 2004. Psychophysical receptive fields of edge detection mechanisms. Vision Res. 44, 795–813.

Elder, J., Zucker, S., 1998. Local scale control for edge detection and blur estimation. IEEE Tr. PAMI 20, 699–716.

Fleet, D., Jepson, A., Jenkin, M., 1991. Phase-based disparity measurement. Image Understanding 53 (2), 198–210.

Freedman, D., Riesenhuber, M., Poggio, T., Miller, E., 2002. Visual categorization and the primate prefrontal cortex: Neurophysiology and behavior. J. Neurophysiol. 88 (2), 929–941.

Freedman, D., Riesenhuber, M., Poggio, T., Miller, E., 2003. Comparison of primate prefrontal and inferior temporal cortices during visual categorization. J. Neuroscience 23 (12), 5235–5246.

Gauthier, I., Hayward, W., Tarr, J., Anderson, A., Skudlarski, P., Gore, J., 2002. Bold activity during mental rotation and viewpoint-dependent object recognition? Neuron 34 (1), 161–171.

Gegenfurtner, K., Kiper, D., Levitt, J., 1997. Functional properties of neurons in macaque area V3. J. Neurophysiol. 77 (4), 1906–1923.

Geisler, W., Perry, J., Super, B., Gallogly, D., 2001. Edge co-occurrence in natural images predicts contour grouping performance. Vision Res. 41 (6), 711–724.

Ghosh, A., Petkov, N., 2005. A cognitive evaluation procedure for contour based shape descriptors. Int. J. Hybrid Intelligent Systems 2 (4), 237–252.

Gooch, B., Coombe, G., Shirley, P., 2002. Artistic vision: painterly rendering using computer vision techniques. Proc. ACM/SIGGRAPH-Eurographics NPAR, Annecy (France), 83–91.

Goodale, M., Milner, A., 1992. Separate visual pathways for perception and action. Trends in Neuroscience 15 (1), 20–25.

Grigorescu, C., Petkov, N., Westenberg, M., 2003. Contour detection based on nonclassical receptive field inhibition. IEEE Tr. IP 12 (7), 729–739.

Grill-Spector, K., Kanwisher, N., 2005. Visual recognition: as soon as you know it is there, you know what it is. Psychological Science 16 (2), 152–160, doi:10.1111/j.0956–7976.2005.00796.x.

Hamker, F., 2005. The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. Cereb. Cortex 15, 431–447.

Hansen, T., Neumann, H., 2002. A biologically motivated scheme for robust junction detection. Proc. Int. Worksh. Biol. Motivated Comp. Vision, Springer LNCS Vol. 2525, 16–26.

Hansen, T., Sepp, W., Neumann, H., 2001. Recurrent long-range interactions in early vision. Lecture Notes in Computer Science 2036, 127–138.

Heath, M., Sarkar, S., Sanocki, T., Bowyer, K., 2000. A robust visual method for assessing the relative performance of edge-detection algorithms. IEEE Tr. PAMI 19 (12), 1338–1359.

Hegarty, M., Waller, D., 2004. A dissociation between mental rotation and perspective-taking spatial abilities. Intelligence 32 (2), 175–191.

Heitger, F., Rosenthaler, L., von der Heydt, R., Peterhans, E., Kubler, O., 1992. Simulation of neural contour mechanisms: from simple to end-stopped cells. Vision Res. 32 (5), 963–981.

Heitger, F., von der Heydt, R., 1993. A computational model of neural contour processing: figure-ground segregation and illusory contours. Proc. Int. Conf. Comp. Vision, Berlin (Germany), 32–40.

Henricsson, O., Heitger, F., 1994. The role of key-points in finding contours. Proc. 3rd Europ. Conf. Comp. Vision 2, 371–382.

Hertzmann, A., 1998. Painterly rendering with curved brush strokes of multiple sizes. Proc. ACM SIGGRAPH98, 453–460.

Hertzmann, A., 2003. A survey of stroke-based rendering. IEEE Comp. Graphics Appl. 23 (4), 70–81.

Hotta, K., Mishima, T., Kurita, T., Umeyama, S., 2000. Face matching through information theoretical attention points and its applications to face detection and classification. IEEE Proc. 4th Int. Conf. Automatic Face and Gesture Recogn. 34–39.

Howe, P., 2001. A comment on the Anderson (1997), the Todorovic (1997), and the Ross and Pessoa (2000) explanations of White's effect. Perception 30 (8), 1023–1026.

Hubel, D., 1995. Eye, Brain and Vision. Scientific American Library.

Hummel, J., 2000. Where view-based theories break down: the role of structure in shape perception and object recognition. Cognitive Dynamics: Conceptual Change in Humans and Machines, 157–185.

Hung, C. P., Kreiman, G., Poggio, T., Dicarlo, J., 2005. Fast readout of object identity from macaque inferior temporal cortex. Science 310 (5749), 863–866.

Hupe, J., James, A., Girard, P., Lomber, S., Payne, B., Bullier, J., 2001. Feedback connections act on the early part of the responses in monkey visual cortex. J. Neurophysiol. 85 (1), 134–144.

Itti, L., Koch, C., 2001. Computational modeling of visual attention. Nature Rev. Neuroscience 2 (3), 194–203.

Kaas, J., Lyon, D., 2001. Visual cortex organization in primates: theories of V3 and adjoining visual areas. Prog. Brain Res. 134, 285–295.

Keil, M., Cristóbal, G., Neumann, H., 2006. Gradient representation and perception in the early visual system - A novel account of Mach band formation. Vision Res. 46, 2659–2674.

Kingdom, F., Moulden, B., 1991. White's effect and assimilation. Vision Res. 31 (1), 151–159.

Koenderink, J., 1984. The structure of images. Biol. Cybern. 50 (5), 363–370.

Kovács, L., Szirányi, T., 2004. Painterly rendering controlled by multiscale image features. Proc. 20th Spring Conf. Comp. Graphics, Budmerice (Slovakia), 177–184.

Kovesi., P., 1999. Image features from phase congruency. J. Comp. Vision Res. 1 (3), 2–27.

Kovesi, P., 2003. Phase congruency detects corners and edges. Proc. Australian Patt. Recogn. Society Conf. 309–318.

Krüger, N., Peters, G., 1997. Object recognition with banana wavelets. Proc. 5th Europ. Symp. Artif. Neural Netw., 61–66.

Krüger, N., Wörgötter, F., 2005. Symbolic pointillism: computer art motivated by human brain structures. Leonardo 38 (4), 337–341.

Kruizinga, P., 1999. Computational models of texture processing neurons. PhD Thesis, Rijksuniversiteit Groningen, The Netherlands.

Kruizinga, P., Petkov, N., 1995. Person identification based on multiscale matching of cortical images. Proc. Int. Conf. and Exhib. High-Perf. Comp. Netw. Springer LNCS 919, 420–427.

Kruizinga, P., Petkov, N., 1999. Nonlinear operator for oriented texture. IEEE Tr. IP 8 (10), 1395–1407.

Lampl, I., Ferster, D., Poggio, T., Riesenhuber, M., 2004. Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. J. Neurophysiol. 92, 2704–2713.

Land, E., McCann, J., 1971. Lightness and retinex theory. J. Opt. Soc. America 61 (1), 1–11.

Lee, T., 1996. Image representation using 2D Gabor wavelets. IEEE Tr. PAMI 18 (10), 959–971.

Leibe, B., Schiele, B., 2003. Analyzing appearance and contour based methods for object categorization. IEEE Proc. Int. Conf. Comp. Vis. Patt. Recogn. 2, 409–415.

Lindeberg, T., 1994. Scale-Space Theory in Computer Vision. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Lindeberg, T., 1999. Automatic scale selection as a pre-processing stage for interpreting the visual world. Proc. Fundamental Structural Properties in Image and Patt. Analysis. Schriftenreihen der Österreichischen Computer Gesellschaft, Vol. 130, 9–23.

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

Logvinenko, A., 2003. Does the bandpass linear filter response predict gradient lightness induction? A reply to Fred Kingdom. Perception 32, 621–626.

Logvinenko, A., Ross, D., 2005. Adelson's tile and snake illusions: a Helmholtzian type of simultaneous lightness contrast. Spatial Vis. 18 (1), 25–72.

Lourens, T., Nakadai, K., Okuno, H., Kitano, H., 2001. Automatic graph extraction from color images. Proc. Europ. Symp. Artif. Neural Netw., 329–334.

Lourens, T., Würtz, R., 1997. Object recognition by matching symbolic edge graphs. Proc. 3rd Asian Conf. Comp. Vision, Springer LNCS Vol. 1352, 193–200.

Lourens, T., Würtz, R., 2003. Extraction and matching of symbolic contour graphs. Int. J. Patt. Recogn. Artif. Intell. 17 (7), 1279–1302.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comp. Vision 2 (60), 91–110.

Miikkulainen, R., Bednar, J., Choe, Y., Sirosh, J., 2005. Computational maps in the visual cortex. Springer Science Business+Media, Inc.

Miller, E., 2000. The prefrontal cortex and cognitive control. Nature Rev. Neuroscience 1 (1), 59–65.

Miller, E., Cohen, J., 2001. An integrative theory of prefrontal cortex function. Annual Rev. Neuroscience 24, 167–202.

Moulden, B., Kingdom, F., 1989. White's effect: a dual mechanism. Vision Res. 29 (9), 1245–1259.

Moulden, B., Kingdom, F., 1990. Light-dark asymmetries in the Craik-Cornsweet-O'Brien illusion and a new model of brightness coding. Perception 5 (2), 101–128.

Neumann, H., Pessoa, L., Hansen, T., 2001. Visual filling-in for computing perceptual surface properties. Biol. Cybern. 85 (5), 355–369.

Nunes, S., Almeida, D., Brito, V., Carvalho, J., Rodrigues, J., du Buf, J., 2006a. Perception-based painterly rendering: functionality and interface design. Proc. Ibero-American Symp. in Comp. Graphics, Santiago de Compostela (Spain), 5-7 July, 53–60.

Nunes, S., Almeida, D., Loke, E., du Buf, J., 2005. Polygon optimization for the modelling of planar range data. Proc. 2nd Iberian Conf. on Patt. Recogn. and Image Anal., Estoril (Portugal), Springer LNCS Vol. 3522, 128–136.

Nunes, S., Almeida, D., Rodrigues, J., du Buf, J., 2006b. Object categorisations using templates constructed from multi-scale line and edge representations. Proc. Worksh. Visual Categorisation and Image Management Systems, University of Sunderland (UK), 28 June, 14–15.

Ohzawa, I., DeAngelis, G., Freeman, R., 1997. Encoding of binocular disparity by complex cells in the cat's visual cortex. J. Neurophysiol. 18 (77), 2879–2909.

Oliva, A., 2005. Gist of the scene. Chapter in: Neurobiology of Attention, L. Itti, G. Rees and J. Tsotsos (eds), Academic Press, Elsevier, 251–256.

Oliva, A., Schyns, P., 1997. Coarse blobs or fine edges evidence that information diagnosticity changes the perception of complex visual stimuli. Cognitive Psychology 34, 72–107.

Oliva, A., Torralba, A., 2006. Building the gist of a scene: the role of global image features in recognition. Progress in Brain Res.: Visual Perception 155, 23–26.

Oliva, A., Torralba, A., Casthelano, M., Henderson, J., 2003. Top-down control of visual attention in object detection. IEEE Proc. Int. Conf. Image Processing 1, 253–256.

Olshausen, B., Anderson, C., van Essen, D., 1993. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J. Neuroscience 13 (11), 4700–4719.

Olshausen, B., Field, D., 2005. How close are we to understanding V1? Neural Computation 17 (8), 1665–1699.

Parasuraman, R., 1998. The attentive brain. MIT Press, Cambridge, Massachusetts.

Parkhurst, D., Law, K., Niebur, E., 2002. Modelling the role of salience in the allocation of overt visual attention. Vision Res. 42 (1), 107–123.

Pasupathy, A., Connor, C., 2001. Shape representation in area V4: position-specific tuning for boundary conformation. J. Neurophysiol. 86 (5), 2505–2519.

Perrett, D., Oram, M., Ashbridge, E., 1998. Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. Cognition 67 (1,2), 111–145.

Pessoa, L., 1996a. Mach-band attenuation by adjacent stimuli: experiments and filling-in simulations. Perception 24 (4), 425–442.

Pessoa, L., 1996b. Mach bands: how many models are possible? Recent experimental findings and modeling attemps. Vision Res. 36, 3205–3227.

Peters, G., 2000. Theories of three-dimensional object perception - A survey. Recent Research Developments in Patt. Recogn., (Part-I), Transworld Research Netw. 1, 179–197.

Peters, G., 2004. Efficient pose estimation using view-based object representations. Machine Vision and Applications 16 (1), 59–63.

Petkov, N., Kruizinga, P., 1997. Computational models of visual neurons specialised in detection of periodic and aperiodic visual stimuli. Biol. Cybern. 76, 83–96.

Petkov, N., Kruizinga, P., Lourens, T., 1993a. Biologically motivated approach to face recognition. Proc. Int. Worksh. Artif. Neural Netw., 68–77.

Petkov, N., Lourens, T., Kruizinga, P., 1993b. Lateral inhibition in cortical filters. Proc. Int. Conf. Dig. Signal Proc. and Int. Conf. on Comp. Appl. to Eng. Sys., Nicosia (Cyprus), 122–129.

Pomplun, M., Rieser, H., Ritter, H., Velichkovsky, B., 1997. Augenbewegungen als kognitionswissenschaftlicher forschungsgegenstand. Kognitionswissenschaft: Strukturen und Prozesse intelligenter Systeme, Kluwe, R.H., Deutscher Universitätsverlag, 65–106.

Prime, D., Ward, L., 2006. Cortical expressions of inhibition of return. Brain Res. 1072 (1), 161–174.

Qian, N., 1997. Binocular disparity and the perception of depth. Neuron 18, 359–368.

Qiu, F. T., von der Heydt, R., 2005. Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. Neuron 47 (1), 155–166.

Rasche, C., 2005. The making of a neuromorphic visual system. Springer.

Rehn, M., Sommer, F., 2006. Storing and restoring visual input with collaborative rank coding and associative memory. Neurocomputing 69 (10-12), 1219–1223.

Rensink, R., 2000. The dynamic representation of scenes. Visual Cogn. 7 (1-3), 17–42.

Riesenhuber, M., 2005. Object recognition in cortex: neural mechanisms and possible roles for attention. Chapter in: Neurobiology of Attention, L. Itti, G. Rees, and J. Tsotsos (eds), Academic Press, Elsevier, 279–287.

Riesenhuber, M., Poggio, T., 1999. Models of object recognition. Nature Neuroscience 2 (11), 1019–1025.

Riesenhuber, M., Poggio, T., 2000a. CBF: A new framework for object categorization in cortex. IEEE Proc. Int. Worksh. Biol. Motivated Comp. Vision, Seoul (Korea), May 15-17, 1–9.

Riesenhuber, M., Poggio, T., 2000b. Computational models of object recognition in cortex: a review. CBCL Paper 190/AI Memo 1695, Massachusetts Inst. Technology, Cambridge, MA.

Riesenhuber, M., Poggio, T., 2000c. Models of object recognition. Nature Neuroscience 3, 1199–1204.

Rizzi, A., Gatta, C., Marini, D., 2003. A new algorithm for unsupervised global and local color correction. Pattern Recogn. Lett. 24 (11), 1663–1677.

Rodrigues, J., du Buf, J., 2004a. Vision frontend with a new disparity model. Early Cogn. Vision Worksh. Isle of Skye (Scotland). `http://www.cn.stir.ac.uk/ecovision-ws/`.

Rodrigues, J., du Buf, J., 2004b. Visual cortex frontend: integrating lines, edges, keypoints and disparity. Proc. Int. Conf. Image Anal. Recogn., Porto (Portugal), Springer LNCS Vol. 3211, 664–671.

Rodrigues, J., du Buf, J., 2005a. Improved line/edge detection and visual reconstruction. Proc. 13th Portuguese Comp. Graphics Meeting, Vila Real (Portugal), 179–184.

Rodrigues, J., du Buf, J., 2005b. Multi-scale cortical keypoint representation for attention and object detection. Proc. 2nd Iberian Conf. on Patt. Recogn. and Image Anal., Estoril (Portugal), Springer LNCS 3523, 255–262.

Rodrigues, J., du Buf, J., 2005c. Multi-scale keypoints in V1 and face detection. Proc. 1st Int. Symp. Brain, Vision and Artif. Intell., Naples (Italy), Springer LNCS Vol. 3704, 205–214.

Rodrigues, J., du Buf, J., 2006a. Cortical object segregation and categorization by multi-scale line and edge coding. Proc. Int. Conf. Comp. Vision Theory Applicat., Setúbal (Portugal), 2, 5–12.

Rodrigues, J., du Buf, J., 2006b. Face recognition by cortical multi-scale line and edge representations. Proc. Int. Conf. Image Anal. Recogn., Póvoa do Varzim (Portugal), Springer LNCS Vol. 3211, 329–340.

Rodrigues, J., du Buf, J., 2006c. Face segregation and recognition by cortical multi-scale line and edge coding. Proc. 6th Int. Worksh. Patt. Recogn. in Information Systems, Paphos (Cyprus), May 23-24, 5–14.

Rodrigues, J., du Buf, J., 2006d. Multi-scale keypoints in V1 and beyond: object segregation, scale selection, saliency maps and face detection. BioSystems, 75–90, doi:10.1016/j.biosystems.2006.02.019.

Rolls, E., Aggelopoulos, N., Zheng, F., 2003. The receptive fields of inferior temporal cortex neurons in natural scenes. J. Neuroscience 23 (1), 339–348.

Rosenthaler, L., Heitger, F., Kübler, O., von der Heydt, R., 1992. Detection of general edges and keypoints. Proc. 2nd Europ. Conf. Comp. Vision, Springer LNCS Vol. 588, 78–86.

Ruzon, M., Tomasi, C., 2001. Edge, junction, and corner detection using color distributions. IEEE Tr. PAMI 23 (11), 1281–1295.

Santos, L., du Buf, J., 2002. Computational cortical cell models for continuity and texture. Proc. Int. Worksh. Biol. Motivated Comp. Vision, Tuebingen (Germany) Springer LNCS Vol. 2525, 90–98.

Schouten, G., 1992. Luminance-brightness mapping: the missing decades. PhD Thesis, Eindhoven Technical University, The Netherlands.

Schwartz, S., 1994. Visual Perception: a clinical orientation. East Norwalk, Appleton and Lange, Connecticut.

Serre, T., Kreiman, G., Poggio, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., Poggio, T., 2005. A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. CBCL Paper 259/AI Memo 2005-036, Massachusetts Institute of Technology, Cambridge, USA.

Serre, T., Riesenhuber, M., 2004. Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex. CBCL Paper 239/AI Memo 2004-017, Massachusetts Institute of Technology, Cambridge, MA.

Shiraishi, M., Yamaguchi, Y., 2000. An algorithm for automatic painterly rendering based on local source image approximation. Proc. ACM/SIGGRAPH-Eurographics NPAR, Annecy (France), 53–58.

Smith, S., Brady, J., 1997. Susan - A new approach to low level image processing. Int. J. Comp. Vision 23 (1), 45–78.

Sousa, M., 2003. Theory and practice of Non–Photorealistic graphics: algorithms, methods, and production systems. Course Notes for SIGGRAPH03.
`http://pages.cpsc.ucalgary.ca/~mario/`.

Stringer, S., Perry, G., Rolls, E., Proske, H., 2006. Learning invariant object recognition in the visual system with continuous transformations. Biol. Cybern. 94, 128–142.

Tarr, J., 1995. Rotating objects to recognize them: a case study of the role of viewpoint dependency in the recognition of three-dimensional objects. Psychonomic Bulletin and Rev. 4 (1), 55–82.

Tarr, J., 2005. How experience shapes vision. Psychological Science Agenda 19 (7).

Tarr, J., Bülthoff, H., 1995. Is human object recognition better described by geon-structural-descriptions or by multiple-views? J. Experimental Psychology: Human Perception and Performance 21 (6), 1494–1505.

Tarr, J., Gauthier, I., 1998. Do viewpoint-dependent mechanisms generalize across members of a class? Cognition 67 (1,2), 71–108.

Tarr, J., Williams, P., Hayward, W., Gauthier, I., 1998. Three-dimensional object recognition is viewpoint dependent. Nature Neuroscience 1 (4), 275–277.

Thorpe, S., Guyonneau, R., Guilbaud, N., Allegraud, J., 2004. Real-time visual processing with one spike per neuron. Neurocomputing 58-60, 857–864.

Triggs, B., 2004. Detecting keypoints with stable position, orientation and scale under illumination changes. Proc. Europ. Conf. Comp. Vision 4, 100–113.

Valentin, D., Abdi, H., Edelman, B., 1997. What represents a face? A computational approach for the integration of physiological and psychological data. Perception 26 (10), 1271–1288.

van Deemter, J., du Buf, J., 2000. Simultaneous detection of lines and edges using compound Gabor filters. Int. J. Patt. Recogn. Artif. Intell. 14 (6), 757–777.

Verbeek, P., van Vliet, L., 1992. Line and edge detection by symmetry filters. Proc. 11th Int. Conf. Patt. Recogn. III, 749–753.

von der Heydt, R., Peterhans, E., Dursteler, M., 1992. Periodic-pattern-selective cells in monkey visual cortex. J. Neuroscience 12 (4), 1416–34.

Wallis, G., Bülthoff, H., 2001. Effects of temporal association on recognition memory. Proc. Natl. Acad. Sci. USA, 4800–4804.

Wallis, G., Rolls, E., 1997. Invariant face and object recognition in the visual system. Prog. in Neurobiol. 51 (2), 167–194.

Walther, D., Itti, L., Riesenhuber, M., Poggio, T., Koch, C., 2002. Attentional selection for object recognition - a gentle way. Proc. Int. Worksh. Biol. Motivated Comp. Vision, Springer LNCS Vol. 2525, 472–479.

Walther, D., Rutishauser, U., Koch, C., Perona, P., 2005. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. Comp. Vision and Image Understanding 100 (1-2), 41–63.

Wang, B., Wang, W., Yang, H., Sun, J., 2004. Efficient example-based painting and synthesis of 2D directional texture. IEEE Tr. Vis. Comp. Graphics 10 (3), 266–277.

Würtz, R., Lourens, T., 2000. Corner detection in color images by multiscale combination of end-stopped cortical cells. Image and Vis. Comp. 18 (6-7), 531–541.

Yang, M., Kriegman, D., Ahuja, N., 2002. Detecting faces in images: a survey. IEEE Tr. PAMI 24 (1), 34–58.

Ye, S., Sun, Q., Chang, E., 2004. Edge directed filter based error concealment for wavelet-based images. IEEE Proc. Int. Conf. on Image Processing 2, 809–812.

Zacks, J., Gilliam, F., Ojemann, J., 2003. Selective disturbance of mental rotation by cortical stimulation. Neuropsychologia 41 (12), 1659–1667.

Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P., 2003. Face recognition: a literature survey. ACM Computing Surveys 35 (4), 399–458.

Zhaoping, L., 2003. V1 mechanisms and some figure-ground and border effects. J. Physiology, 503–515.

Zoccolan, D., Cox, D., DiCarlo, J., 2005. Multiple object response normalization in monkey inferotemporal cortex. J. Neuroscience 25 (36), 8150–8164.