CAR-TR-722 DACA76-92-C-0009
CS-TR-3305 IRI-90-57934
June 1994 N00014-93-1-0257

Vision and Action

Cornelia Fermüller
Yiannis Aloimonos

Computer Vision Laboratory
Center for Automation Research
University of Maryland
College Park, MD 20742-3275

Abstract

Our work on Active Vision has recently focused on the computational modelling of navigational tasks, where our investigations were guided by the idea of approaching vision for behavioral systems in form of modules that are directly related to perceptual tasks. These studies led us to branch in various directions and inquire into the problems that have to be addressed in order to obtain an overall understanding of perceptual systems. In this paper we present our views about the architecture of vision systems, about how to tackle the design and analysis of perceptual systems, and promising future research directions. Our suggested approach for understanding behavioral vision to realize the relationship of perception and action builds on two earlier approaches, the Medusa philosophy [3] and the Synthetic approach [15]. The resulting framework calls for synthesizing an artificial vision system by studying vision competences of increasing complexity and at the same time pursuing the integration of the perceptual components with action and learning modules. We expect that Computer Vision research in the future will progress in tight collaboration with many other disciplines that are concerned with empirical approaches to vision, i.e. the understanding of biological vision. Throughout the paper we describe biological findings that motivate computational arguments which we believe will influence studies of Computer Vision in the near future.

The support of the Advanced Research Projects Agency (ARPA Order No. 8459) and the U.S. Army Topographic Engineering Center under Contract DACA76-92-C-0009, the National Science Foundation under Grant IRI-90-57934, and the Office of Naval Research under Contract N00014-93-1-0257, is gratefully acknowledged, as is the help of Sandy German in preparing this paper.

Written on the occasion of the 30^{th} anniversary of the Computer Vision Laboratory at Maryland.

1 Introduction

"The past two decades ... have led to a powerful conceptual change in our view of what the brain does ... It is no longer possible to divide the process of seeing from that of understanding ...". [69]. These lines of Zeki's article express in a concise way what has been realized in different disciplines concerned with the understanding of perception. Vision (and perception in general) should not be studied in isolation but in conjunction with the physiology and the tasks that systems perform. In the discipline of Computer Vision such ideas caused researchers to extend the scope of their field. If initially Computer Vision was limited to the study of mappings of a given set of visual data into representations on a more abstract level, it now has become clear that Image Understanding should also include the process of selective acquisition of data in space and time. This has led to a series of studies published under the headings of Active, Animate, Purposive, or Behavioral Vision. A good theory of vision would be one that can create an interface between perception and other cognitive abilities. However, with a formal theory integrating perception and action still lacking, most studies have treated Active Vision [2, 4] as an extension of the classical reconstruction theory, employing activities only as a means to regularize the classical ill-posed inverse problems.

Let us summarize the key features of the classical theory of Vision in order to point out its drawbacks as an overall framework for studying and building perceptual systems: In the theory of Marr [37], the most influential in recent times, Vision is described as a reconstruction process, that is, a problem of creating representations at increasingly high levels of abstraction, leading from 2D images through the primal sketch and the $2\frac{1}{2}$D sketch to object-centered descriptions ("from pixels to predicates") [48]. Marr suggested that visual processes—or any perceptual/cognitive processes—are information processing tasks and thus should be analyzed at three levels: (a) at the computational theoretic level (definition of the problem and its boundary conditions; formulation of theoretical access to the problem), (b) at the level of selection of algorithms and representations (specification of formal procedures for obtaining the solution), and (c) at the implementational level (depending on the available hardware).

In the definition of cognitive processing in the classical theory, Vision is formalized as a pure information processing task. Such a formalization requires a well-defined closed system. Since part of this system is the environment, the system would be closed only if it were possible to model all aspects of objective reality. The consequence is well-known: Only toy problems (blocks worlds, Lambertian surfaces, smooth contours, controlled illumination, and the like) can be successfully solved.

The strict formalization of representations at different levels of abstraction gave rise to breaking the problems into autonomous subproblems and solving them independently. The conversion of external data (sensor data, actuator commands, decision making, etc.) into an internal representation was separated from the phase of algorithms to perform computations on internal data; signal processing was separated from symbolic processing and action. Processing of visual data was treated, for the most part, in a syntactic manner and semantics was treated in a purely symbolic way using the results of the syntactic analysis. This is not surprising, since Computer Vision was considered as a subfield of Artificial Intelligence and thus studied using the same methodology, influenced by the ideas and computational theories of the last decades [12, 21, 44].

The strict hierarchical organization of representational steps in the Marr paradigm makes the development of learning, adaptation and generalization processes practically impossible (so that there hasn't been much work on "vision and learning"). Furthermore, the conceptualization of a vision system as consisting of a set of modules recovering general scene descriptions in a hierarchical manner introduces computational difficulties with regard to issues of robustness, stability, and efficiency. These problems lead us to believe that general vision does not seem to be feasible. Any system has a specific relationship with the world in which it lives, and the system itself is nothing but an embodiment of this relationship. In the Marr approach the algorithmic level has been separated from the physiology of the system (the hardware) and thus vision was studied in a disembodied, transcendental manner.

Of course, many of the solutions developed for disembodied systems may also be of use for embodied ones. In general, however, this does not hold. Given infinite resources, every (decidable) problem can be solved in principle. Assuming that we live in a finite world and that we have a finite number of possibilities for performing computations, any vision problem might be formulated as a simple search problem in a very high dimensional space. From this point of view, the study of embodied systems is concerned with the study of techniques to make seemingly intractable problems tractable.

Not the isolated modelling of observer and world (as closed systems), but the modelling of observer and world in a synergistic manner, will contribute to the understanding of perceptual information processing systems [58]. The question, of course, still remains how such a synergistic modelling should be realized. Or: How can we relate perception and action? What are the building blocks of an intelligent perceptual system? What are the categories into which the system divides its perceptual world? What are the representations it employs? How is it possible to implement such systems in a flexible manner to allow them to learn from experience and extend themselves to better ones? In this paper we present a formal framework for addressing these questions. Our exposition describes both some recent technical results and some of our future research agenda.

2 Where are we heading to?

2.1 Interdisciplinary research

Computer Vision is not the only discipline concerned with the study of cognitive processes responsible for a system's interaction with its environment. The last decade of the 20th century has been declared the decade of the brain. A number of new fields that together have established themselves as Neurosciences are providing us with results about the components of actually existing brains. In areas such as Neurophysiology, Neurogenetics, and Molecular Biology new techniques have been developed that allow us to trace the processes at the molecular, neural, and cellular levels. By now we have gained some insight into the various functional components of the brain. We are, however, far from understanding the whole. There are many other different disciplines concerned with the problem of perception from the biological point of view: Psychology, Cognitive Neurophysiology, Ethology, and Biology, to name a few of them.

For most of its history, cognitive modelling has focused almost exclusively on human abilities and capacities. In the past, however, the studies were guided by other ideas and a large number of psychological and psychophysical studies concentrated on the understanding of singularities in human perception, or visual illusions, as they are commonly called. The assumption was that the brain is designed in a modular, principled fashion, and thus from the study of perceptual malfunctions (illusions [24]), information about its design can be deduced. Recent results from Cognitive Neurophysiology—the discipline which is concerned, among other topics, with the study of visual agnosia (a condition exhibited by patients with partially damaged brains) [13, 30]—indicate that the human brain is not designed in a clean, modular fashion, but consists of several processes working in a cooperative, distributed manner. The findings from studies of illusions actually support this point, since a multitude of computational theories of different natures have been proposed for explaining the multitude of human visual illusions.

When referring to the intelligence of biological systems, we refer to the degree of sophistication of their competences and to the complexity of the behaviors that they exhibit in order to achieve their goals. Various disciplines have been concerned with the study of competences in biological organisms. Genetics and Evolution theory study how different species acquire their species-specific competences. Competences are classified into two categories: those genetically inherited (through phylogenesis) and those acquired individually, responsible for the specific categories that an individual distinguishes (through ontogenesis). In Ethology the relationship between the acquisition of individual and species-specific competences and the behavior of biological organisms is investigated. Organisms at various levels of complexity have been researched. The discipline of Neuroethology is concerned with the physical implementation of behaviors. By now it has given rise to a great deal of insight in the understanding of perceptual systems of lower animals, such as medusae, worms, and insects. In Computational Neuroethology (Neuroinformatics) researchers are copying the neuronal control found in such simple organisms into artificial systems with the hope of learning to understand in this way the dynamics responsible for adaptive behavior.

Two other fields concerned with the study of interactions of systems and their environments have also given rise to a number of new technical tools and mathematics. One of these is Cybernetics. Its goal is the study of relationships between behaviors of dynamical self-regulating systems (biological and artificial ones) and their structure. Cybernetics initiated many efforts in Control theory. The mathematics that has been employed involves integral and differential equations. The other discipline is Synergetics, which searches for universal principles in the interrelationship of the parts of a system that possesses macroscopic spatial, temporal, and functional structures.

2.2 The approach

After these discussions of biological sciences, one might assume that it is suggested here to define the scope of Computer Vision as copying biological vision in artificial systems. Not at all. Computer Vision is the discipline concerned with the study of the computational theories underlying vision. Its goal is to gain insight into perception from a computational point of view. The computations that could possibly exist have to be of a certain nature. Thus the problem is to understand the inherent properties of the computations that a framework which models the understanding of purposive, embodied systems will have.

To achieve this goal the study of perception has to be addressed at various levels of abstraction. Our approach here is two-fold: On the one hand we attempt to provide a global model—a working model—for explaining the abstract components of a vision system. On the other hand we propose an approach for achieving the study and building of actual vision systems. The interaction we expect with biological sciences will be of the following kind. Results from biological sciences should give us inspiration about the visual categories relevant for systems existing in environments like those of humans. The constraints imposed by the possible computations should tell the biological sciences what experiments to perform to find out how biological organisms can possibly function.

2.3 The modules of the system

Figure 1 gives a pictorial description of the basic components of a purposive vision system: The abstract procedures and representations of a vision system are: the procedures for performing visual perceptions, physical actions, learning, and information retrieval, and purposive representations of the perceptual information along with representations of information acquired over time and stored in memory.

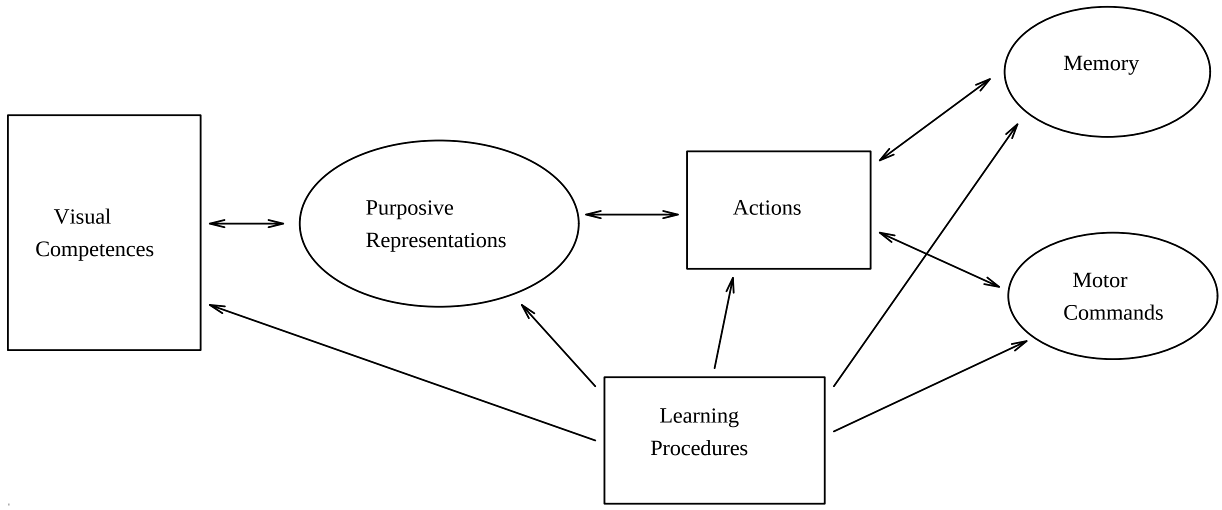


Figure 1: Working model: Basic components of a purposive vision system.

At any time a purposive vision system has a goal or a set of goals that it wishes to achieve as best as it can by means of its available resources. Thus at any time the system is engaged in executing a task. The visual system possesses a set of visual competences with which it processes the visual information. The competences compute purposive representations. Each of these representations captures some aspect of the total visual information. Thus compared with the representations of the old paradigm, they are partial. The representations are of different complexities with regard to the space they describe. The purposive representations themselves are purposive descriptions of the visual information organized in certain data structures. The purposive representations access programs which we call “action routines”. This collective name refers to two kinds of routines; the first kind are the programs that schedule the physical actions to be performed, i.e. they initialize motor commands and thus provide the interface to the body, and the second kind schedule the selection of information to be retrieved from the purposive representations and stored in long-term memory. An important aspect of the architecture is that the access of the visual processes to the actions is on the basis of the contents of the purposive representations; i.e., the contents of the purposive representations serve as addresses to the actions. Another class of programs is responsible for learning by providing the actions, the competences, and the representations with the means to change and adjust parameters.

As can be seen from the figure, learning takes place at various levels of, as well as in between, the modules of the system. For a flexible vision system, it should be possible to learn the parameters describing actions, to acquire new actions, to learn parameters describing visual competences, to acquire new visual competences that compute new purposive representations, and to learn the sequences of actions and perceptual competences to perform a task. In any case, learning is accomplished by means of programs—learning procedures—that allow the change and adaptation of parameters in order to learn competences, actions, and their interrelationships.

The purposive perceptual representations, as well as representations containing other kinds of information, are stored in memory. The storing must happen in an efficient way according to the available memory space. Different representations share common elements. Memory organization techniques have to be studied that allow information to be stored according to its content. Also, designing a memory for representations includes designing the procedures necessary for fast and reliable access.

The abstract components on which we focus our discussion are: (1) the visual competences, and (2) the organization of memory and the procedures for learning related to visual processing and the coupling of action and perception.

Let us summarize in which way the above model captures the study of perception and action in a synergistic way, and address some of the questions posed in Section 1: In this model the intelligence of a purposive system is embodied in its visual competences and its actions. Thus competences and actions are considered to be the building blocks of an intelligent system. In order to fulfill a purpose (a task which is stated in the form of events that can be perceived by means of the perceptual processes), a system executes behaviors. Thus, behaviors, which are an emergent attribute of the system, couple perception and action. They constitute some form of structure adaptation which might either be visible externally or take place only internally in the form of parameter adaptation.

2.4 Outline of the approach

If we aim to understand perception, we have to come up with some methodology to study it. The ideal thing would be to design a clearly defined model for the architecture of vision systems and start working on its components. However, we have few answers available when it comes down to actually talking about the visual categories that are relevant for visual systems. What kind of representations a system needs in order to perform a task depends on the embodiment of the system and the environment in which it lives. Answers to these questions cannot come as insights gained from the study of mathematical models. It must be empirical studies investigating systems (biological and artificial ones) that will tell us how to couple functionality, visual categories and visual processes. Up to now we haven't understood how we actually could develop visual competences for systems that work in environments as complex as our own, so we won't be able to obtain a global view of the overall architecture and functionality of vision systems. At this point in time it also wouldn't contribute much to the development of our understanding to just go ahead and develop particular systems that perform particular tasks—say, for example, to build a system that recognizes tables. Even if we were able to create such a system that has a success rate of 99%, this system would have the capacity of recognizing many things that are unknown to us, and not just tables. Thus by aiming to build systems that recognize certain categories that seem relevant to our symbolic language repertoire, we wouldn't gain much insight into perception.

It thus seems somehow natural that the only way out of this problem of where to start is to approach the study of vision systems in an "evolutionary" way. We call such an approach the synthetic (evolutionary) approach to Medusa (or Medusa synthesized). We give here a short outline of the ideas behind this approach, which we discuss in detail in the remainder of the paper. It means that we should start by developing individual primitive visual operations and provide the system in this way with visual capabilities (or competences). As we go on, the competences will become more and more complex. At the same time, as soon as we have developed a small number of competences, we should work on their integration. Such an endeavor throws us immediately into the study of two other major components of the system: How is visual information related to action and how is the information represented—how is it organized, how coordinated with the object recognition space. Thus we are confronted on the one hand with the study of activities and the integration of vision and action, and on the other hand with the study of the memory space with all its associated problems of memory organization, visual data representation, and indexing—the problem of associating data stored in the memory with new visual information. Furthermore we also have to consider the problem of learning from the very beginning.

3 The competences

3.1 Computational principles

Our goal is to study (or more precisely formulated: analyze in order to design) a system from a computational point of view. We argued earlier that the study of visual systems should be performed in a hierarchical manner according to the complexity of the visual processes. As a basis for its computations a system has to utilize mathematical models, which serve as abstractions of the representations employed. Thus, when referring to the complexity of visual processes, we mean the complexity of the mathematical models involved.

Naturally, the computations and models are related to the class of tasks the system is supposed to perform. A system possesses a set of capabilities which allow it to solve certain tasks. In order to perform a task the system has to extract and process certain informational entities from the imagery it acquires through its visual apparatus. What these entities are depends on the visual categories the system reacts to. The categories again are related to the task the system is engaged in. They are also related to the system's physiology, or amount of space (memory) and the time available to solve the task (the required reaction time).

The synthetic approach calls first for studying capabilities whose development relies on only simple models and then going on to study capabilities requiring more complex models. Simple models do not refer to environment- or situation-specific models which are of use in only limited numbers of situations. Each of the capabilities requiring a specified set of models can be used for solving a well-defined class of tasks in every environment and situation the system is exposed to. If our goal is to pursue the study of perception in a scientific way, as opposed to industrial development, we have to accept this requirement as one of the postulates, although it is hard to achieve. Whenever we perform computations, we design models on the basis of assumptions, which in the case of visual processing are constraints on the space-time in which the system is acting, on the system itself, and on their relationship. An assumption can be general with regard to the environment and situation, or very specific.

For example, the assumption about piecewise planarity of the world is general with regard to the environment (every continuous differentiable function can be approximated in an infinitesimal area by its derivatives). However, in order to use this assumption for visual recovery, additional assumptions regarding the number of planar patches have to made; and these are environment-specific assumptions. Similarly, we may assume that the world is smooth between discontinuities; this is general with regard to the environment. Again, for this assumption to be utilized we must make some assumptions specifying the discontinuities, and then we become specific. We may assume that an observer only translates. If indeed the physiology of the observer allows only translation, than we have made a general assumption with regard to the system. If we assume that the motion of an observer in a long sequence of frames is the same between any two consecutive frames, we have made a specific assumption with regard to the system. If we assume that the noise in our system is Gaussian or uniform, again we have made a system-specific assumption.

Our approach requires that the assumptions used have to be general with regard to the environment and the system. Scaled up to more complicated systems existing in various environments, this requirement translates to the capability of the system to decide whether a model is appropriate for the environment in which the system is acting. A system might possess a set of processes that together supply the system with one competence. Various of the processes are limited to specific environmental specifications. The system, thus, must be able to acquire knowledge about what processes to apply in a specific situation.

The motivation for studying competences in a hierarchical way is to increasingly gain insight into the process of vision, which is of high complexity. Capabilities which require complex models should be based on "simpler", already developed capabilities. The complexity of a capability is thus given by the complexity of the assumptions employed; what has been considered a "simple" capability might require complex models, and vice versa.

The basic principle concerning the implementation of processes subserving the capabilities, which is motivated by the need for robustness, is the quest for algorithms which are qualitative in nature. We argue that visual competences should not be formulated as processes that reconstruct the world but as recognition procedures. Visual competences are procedures that recognize aspects of objective reality which are necessary to perform a set of tasks. The function of every module in the system should constitute an act of recognizing specific situations by means of primitives which are applicable in general environments. Each such entity recognized constitutes a category relevant to the system. To give some examples from navigation:

The problem of independent motion detection by a moving observer usually has been addressed with techniques for segmenting optical flow fields. But it also may be tackled through the recognition of non-rigid flow fields for a moving observer partially knowing its motion [3, 41, 61]. The problem of obstacle detection could be solved by recognizing a set of locations on the retina that represent the image of a part of the 3D world being on a collision course with the observer. To perform this task it is not necessary to compute the exact motion between the observer and any object in the scene, but only to recognize that certain patterns of flow evolve in a way that signifies the collision of the corresponding scene points with the observer [42]. Pursuing a target amounts to recognizing the target's location on the image plane along with a set of labels representing aspects of its relative motion sufficient for the observer to plan its actions. Motion measurements of this kind could be relative changes in the motion such as a turn to the left, right, above, down, further away, or closer. In the same way, the problem of hand/eye coordination can be dealt with using stereo and other techniques to compute the depth map and then solve the inverse kinematics problem in order to move the arm. While the arm is moving the system is blind [6]. However the same problem can be solved by creating a mapping (the perceptual kinematic map) from image features to the robot's joints; the positioning of the arm is achieved by recognizing the image features [25].

Instead of reconstructing the world, the problems described above are solved through the recognition of entities that are directly relevant to the task at hand. These entities are represented by only those parameters sufficient to solve the specific task. In many cases, there exists an appropriate representation of the space-time information that allows us to directly derive the necessary parameters by recognizing a set of locations on this representation along with a a set of attributes. Since recognition amounts to comparing the information under consideration with prestored representations, the described approaches to solving these problems amount to matching patterns.

In addition, image information should be, whenever possible, utilized globally. Since the developed competences are meant to operate in real environments under actual existing conditions—just such as biological organisms do—the computations have to be insensitive to errors in the input measurements. This implies a requirement for redundancy in the input used. The partial information about the scene, which we want to recognize, will mostly be globally encoded in the image information. The computational models we are using should thus be such that they map global image information into partial scene information. Later in this section, we will demonstrate our point by means of the rigid motion model.

In order to speak of an algorithm as qualitative, the primitives to be computed do not have to rely on explicit unstable, quantitative models. Qualitativeness can be achieved in a number of ways:

The primitives might be expressible in qualitative terms, or their computation might be derived from inexact measurements and pattern recognition techniques, or the computational model itself might be proved stable and robust in all possible cases.

The synthetic approach to Medusa has some similarities at the philosophical level with Brooks' proposal about understanding intelligent behavior through the construction of working mechanisms [7]. In proposing the subsumption architecture, Brooks suggested a hierarchy of of competences such as avoiding contact with objects, exploring the world by seeing places, reasoning about the world in terms of identifiable objects, etc. This proposal, however, suffered from the same curse of generality that weakened Marr's approach. The subsumption architecture lacked a solid basis, since it did not provide a systematic way of creating a hierarchy of competences by taking into account the system's purpose and physiology.

3.2 Biological hierarchy

It remains to discuss what actually are the simple capabilities that we should concentrate our first efforts on. Other scientific disciplines give us some answer. Much simpler than the human visual system are the perceptual systems of lower animals, like medusae, worms, crustaceans, insects, spiders and molluscs. Researchers in neuroethology have been studying such systems and have by now gained a great deal of understanding. Horridge [28, 29], working on insect vision, studied the evolution of visual mechanisms and proposed hierarchical classifications of visual capabilities. He argued that the most basic capabilities found in animals are based on motion. Animals up to the complexity of insects perceive objects entirely by relative motion. His viewpoint concerning the evolution of vision is that objects are first separated by their motions, and with the evolution of a memory for shapes, form vision progressively evolves. The importance of these studies on lower animals becomes very clear when we take into account the commonly held view by leaders in this field, that the principles governing visual motor control are basically the same in lower animals and humans—whereas, of course, we humans and other primates can see without relative motion between ourselves and our surrounding.

In the last decades the part of the brain in primates responsible for visual processing—the visual cortex—has been studied from an anatomical, physiological, and also behavioral viewpoint. Different parts of the visual cortex have been identified and most of their connections established. Most scientists subscribe to the theory that the different parts perform functionally specialized operations. What exactly these functions are has not been clarified yet. In particular, opinions diverge about the specialization and the interconnections involved in later stages of processing of the visual data. Much more is known about the earlier processes. The visual signal reaches the cortex at the primary visual cortex—also called V1, or striate cortex, via the retina and the lateral geniculate body. From the primary visual cortex the visual signals are sent to about 30 extrastriate or higher-order visual cortical areas, among which about 300 connections have been reported. Figure 2, taken from [47], shows the major areas involved in visual processing. According to Orban the modules in the primate visual cortex can be divided into four hierarchical levels of processing. It seems to be pretty well accepted that there exist lower areas that are specialized for the processing of either static or dynamic imagery. MT (also called V5), MST, and FST seem to be involved in motion processing, and V4 in color processing. Form vision seems to be accomplished by different lower modules which use both static and dynamic information. Zeki [70], for example, suggests that V3 is responsible for the understanding of form from motion information, and V4 derives form and color information. At later stages the modules process both kinds of information in a combined way.

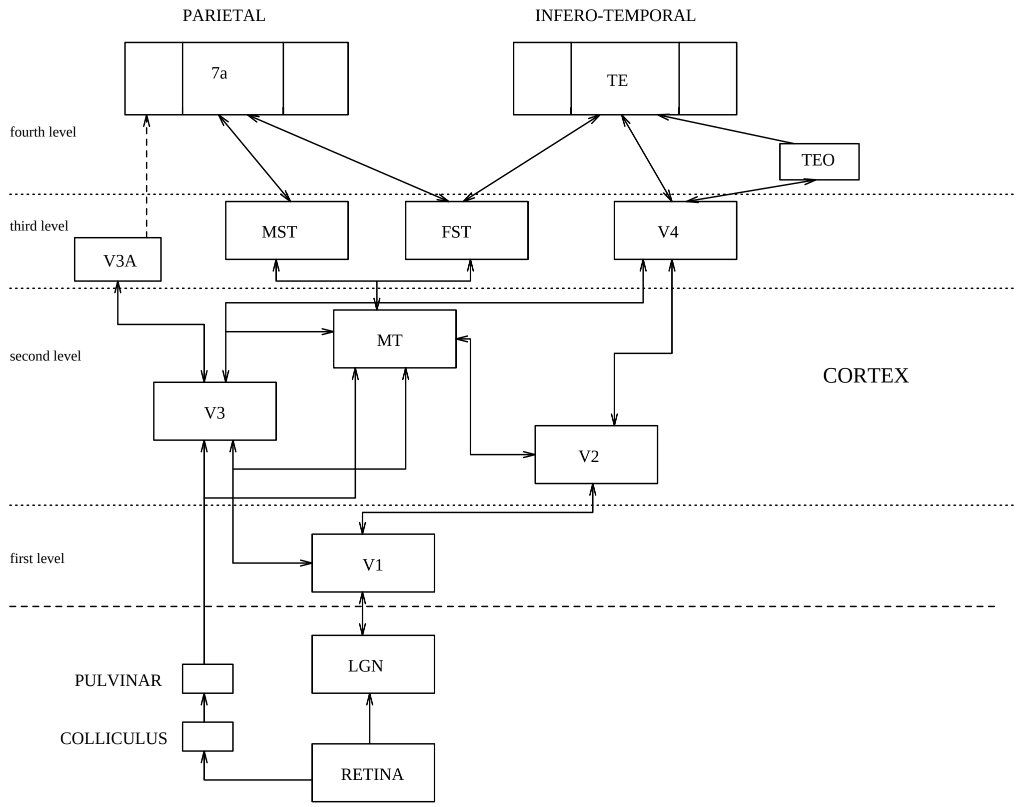


Figure 2: Diagram of the primate visual system indicating the subcortical structure as well as the four tentative levels of cortical visual processing (from [47]).

On the basis of anatomical evidence and behavioral studies (studies on patients with lesions of specific cortical areas) the hypothesis has been advanced [66] that there exist two visual pathways originating from V1: a dorsal one leading to the parietal cortex and a ventral one leading to the infero-temporal cortex. The dorsal path is concerned with either the computations concerned with "where" (object localization) or "how" (the visual guidance of movements [23]), and the ventral path with the computations concerned with "what" (object identification). It would be an oversimplification to conceive of these two pathways as being mutually exclusive and hierarchically organized [70]; one of the reasons is that this theory fails to provide an answer to where and how the knowledge of "what" an object is might be integrated with the knowledge of "where" it is. Also, recently the existence of a third pathway leading to the identification of actions has been suggested [5].

Results from the brain sciences show us that there doesn't exist just one hierarchy of visual processes, but various different computations are performed in parallel. Also, it isn't our intention to propose one strict hierarchy for developing visual competences. We merely suggest studying competences by investigating more and more complex models, and basing more complicated competences on simpler ones. Naturally, it follows that computations concerned with different cues and representations can and should be studied in parallel.

Inspired by the results from the natural sciences, we chose to study first the competences that only involve information resulting from motion. This led us to the problems of navigation. The competences we encounter in visual navigation encompass representations of different forms. To elucidate the synthetic approach, in the next section we will discuss a series of competences of increasing complexity employing representations of motion, shape, and space. In the following section we will then outline our realizations of the most basic competences in visual navigation, which only require motion information.

Next in the hierarchy follow capabilities related to the understanding of form and shape and the learning of space. Concerning form and shape, our viewpoint is that we should not try to adopt the classical idea of computing representations that capture the 3D world metrically. Psychological studies on the role of the eye movements suggest that fixations play an important role in our understanding of space. It seems to be that the level on which information from successive fixations is integrated is relatively abstract and that the representations from which organisms operate on the world are 3D only locally. Therefore, it will be necessary to study new forms of shape representations. In nature too there doesn't exist just one method of shape representation. As results from Neurobiology show, form perception in human brains takes place in more than just one part of the cortex and is realized with different kinds of hardware.

Space is also understood from the processing of various cues in a variety of ways. Furthermore, different tasks will require representations of space with regard to different reference systems—not just one, as often has been debated in the past. Representations might be object-centered, ego-centered, or action-driven.

Actions can be very typical for objects. Early perceptual studies have shown that humans are able to interpret moving scenes correctly, even when the static view does not contain information about the structure at all. In the experiments of Johansson [32] subjects were able to recognize animals, as well as specific human beings, given only the motions of light bulbs mounted on the object's joints. Since our viewpoint is that we should formulate competences as recognition procedures, the study of navigation also leads us to the study of action-driven visual processing. We propose to start modelling such competences by means of more complicated motion models (non-rigid motion models).

3.3 A hierarchy of models for navigational competences

Navigation, in general, refers to the performance of sensory mediated movement, and visual navigation is defined as the process of motion control based on an analysis of images. A system with navigational capabilities interacts adaptively with its environment. The movement of the system is governed by sensory feedback which allows it to adapt to variations in the environment. By this definition visual navigation comprises the problem of navigation where a system controls its single components relative to the environment and relative to each other.

Visual navigation encompasses a wide range of perceptual competences, including tasks that every biological species possesses, such as motion segmentation or kinetic stabilization (the ability of a single compact sensor to understand and control its own motion), as well as advanced specific hand-eye coordination and servoing tasks.

To explain the principles of the synthetic approach to Medusa, we describe six such competences, all of which are concerned only with the movement of a single compact sensor. These are: ego-motion estimation, partial object-motion estimation, independent motion detection, obstacle avoidance, target pursuit, and homing. These particular competences allow us to demonstrate a hierarchy of models concerned with the representation of motion, form and shape.

In the past, navigational tasks, since they inherently involve metric relationships between the observer and the environment, have been considered as subproblems of the general "structure-from-motion" problem [63]. The idea was to recover the relative 3D-motion and the structure of the scene in view from a given sequence of images taken by an observer in motion relative to its environment. Indeed, if structure and motion can be computed, then various subsets of the computed parameters provide sufficient information to solve many practical navigational tasks. However, although a great deal of effort has been spent on the subject, the problem of structure from motion still remains unsolved for all practical purposes. The main reason for this is that the problem is ill-posed, in the sense that its solution does not continuously depend on the input.

The most simple navigational competence, according to our definition, is the estimation of ego-motion. The observer's sensory apparatus (eye/camera), independent of the observer's body motion, is compact and rigid and thus moves rigidly with respect to a static environment. As we will demonstrate, the estimation of an observer's motion can indeed be based on only the rigid motion model. A geometric analysis of motion fields reveals that the rigid motion parameters manifest themselves in the form of patterns defined on partial components of the motion fields [16]. Algorithmically speaking, the estimation of motion thus can be performed through pattern recognition techniques.

Another competence, the estimation of partial information about an object's motion (its direction of translation), can be based on the same model. But, whereas for the estimation of egomotion the rigid motion model could be employed globally, for this competence only local measurements can legitimately be employed. Following our philosophy about the study of perception, it makes perfect sense to define such a competence, which seemingly is very restricted. Since our goal is to study visual problems in the form of modules that are directly related to the visual task the observer is engaged in, we argue that in many cases when an object is moving in an unrestricted manner (translation and rotation) in the 3D world, we are only interested in the object's translational component, which can be extracted using dynamic fixation [17].

Next in the hierarchy follow the capabilities of independent motion detection and obstacle avoidance. Although the detection of independent motion seems to be a very primitive task, it can easily be shown by a counterexample that in the general case it cannot be solved without any knowledge of the system's own motion. Imagine a moving system that takes an image showing two areas of different rigid motion. From this image alone, it is not decidable which area corresponds to the static environment and which to an independently moving object.

However, such an example shouldn't discourage us and drive us to the conclusion that egomotion estimation and independent-motion detection are chicken-and-egg problems: unless one of them has been solved, the other can't be addressed either. Have you ever experienced the illusion that you are sitting in front of a wall which covers most of your visual field, and suddenly this wall (which actually isn't one) starts to move? You seem to experience you yourself moving. It seems that vision alone does not provide us (humans) with an infallible capability of estimating motion. In nature the capability of independent motion detection appears at various levels of complexity. We argue that in order to achieve a very sophisticated mechanism for independent motion detection, various different processes have to be employed. Another glance at nature should give us some inspiration: We humans do not perceive everything moving independently in our visual field. We usually concentrate our attention on the moving objects in the center of the visual field (where the image is sensed with high resolution) and pay attention only if something is moving fast in the periphery. It thus seems to make sense to develop processes that detect anything moving very fast [41]. If some upper bound on the observer's motion is known (maximal speed), it is possible to detect even for small areas where motions above the speed threshold appear. Similarly, for specific systems, processes that recognize specific types of motion may be devised by employing filters that respond to these motions (of use, for example, when the enemy moves in a particular way). To cope with the "chicken-and-egg" problem in the detection of larger independently moving objects, we develop a process, based on the same principle as the estimation of egomotion, which for an image patch recognizes whether the motion field within the patch originates from only rigid motion, or whether the constraint of rigidity does not hold. Having some idea about the egomotion or the scene (for example, in the form of bounds on the motion, or knowing that the larger part of the scene is static) we can also decide where the independently moving objects are.

In order to perform obstacle avoidance it is necessary to have some representation of space. This representation must capture in some form the change of distance between the observer and the scene points which have the potential of lying in the observer's path. An observer that wants to avoid obstacles must be able to change its motion in a controlled way and must therefore be able to determine its own motion and set it to known values. As can be seen, the capability of egomotion estimation is a prerequisite for obstacle avoidance mechanisms, and general independent motion detection will require a model which is as complex as that used in egomotion estimation in addition to other simple motion models.

Even higher in the hierarchy are the capabilities of target pursuit and homing (the ability of a system to find a particular location in its environment). Obviously, a system that possesses these capabilities must be able to compute its egomotion and must be able to avoid obstacles and detect independent motion. Furthermore, homing requires knowledge of the space and models of the environment (for example, shape models), whereas target pursuit relies on models for representing the operational space and the motion of the target. These examples should demonstrate the principles of the synthetic approach, which argues for studying increasingly complex visual capabilities and developing robust (qualitative) modules in such a way that more complex capabilities require the existence of simpler ones.

3.4 Motion-based competences

In this section we describe the ideas behind some of the modules we have developed to realize the most basic competences for visual navigation: the competence of ego-motion estimation, a process for partial object motion estimation and a process for independent motion detection. This description should merely serve to demonstrate our viewpoint concerning the implementation of qualitative algorithms; more detailed outlines and analyses are found elsewhere.

First, let us state some of the features that characterize our approach to solving the above mentioned competences, and differentiates it from most existing work.

In the past, the problems of ego-motion recovery for an observer moving in a static scene and the recovery of an object's 3D motion relative to the observer, since they both were considered as reconstruction problems, have been treated in the same way. The rigid motion model is appropriate if only the observer is moving, but it holds only for a restricted subset of moving objects—mainly man-made ones. Indeed, all objects in the natural world move non-rigidly. However, considering only a small patch in the image of a moving object, a rigid motion approximation is legitimate. For the case of egomotion, data from all parts of the image plane can be used, whereas for object motion only local information can be employed.

Most current motion understanding techniques require the computation of exact image motion (optical flow in the differential case or correspondence of features in the discrete case). This, however, amounts to an ill-posed problem, additional assumptions about the scene have to be employed and as a result, in the general case, the computed image displacements are imperfect. In turn, the recovery of 3D motion from noisy flow fields has turned out to be a problem of extreme sensitivity with small perturbations in the input causing large amounts of error in the motion parameter estimates. To overcome this problem, in our approach to the development of motion related competences, we skip the first computational step. All the techniques developed are based on the use of only the spatio-temporal derivatives of the image intensity function—the so-called normal flow. As a matter of fact, in part only the sign of the normal flow is employed. It should be mentioned that a few techniques using normal flow have appeared in the literature; however, they deal only with restricted cases (only translation or only rotation [1, 27]).

Another characteristic is that the constraints developed for the motion modules, for which the rigid motion module is the correct one globally, are such that the input also is utilized globally. The basis of these computations form global constraints which relate the spatio-temporal derivatives of the image intensity function globally to the 3D motion parameters.

The global constraints are defined on classes of normal flow vectors. Given a normal flow field, the vectors are classified according to their directions. The vectors of each class have a certain structure that takes the form of patterns in the image (on the sphere or in the plane). For example, one can select in the plane normal flow vectors whose direction is defined with regard to a point with coordinates (r, s). These so-called copoint vectors (r, s) are vectors which are perpendicular to straight lines passing through the point (r, s). In addition, the normal flow vectors of a class are distinguished as to whether their direction is counter-clockwise or clockwise with respect to (r, s), in which case they are called positive or negative (see Figure 3). Since any point (r, s) in the image can be chosen as a reference point, there exists an infinite number of such classifications.

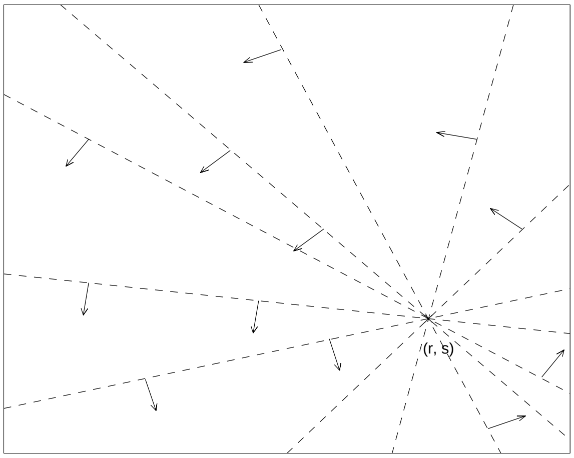
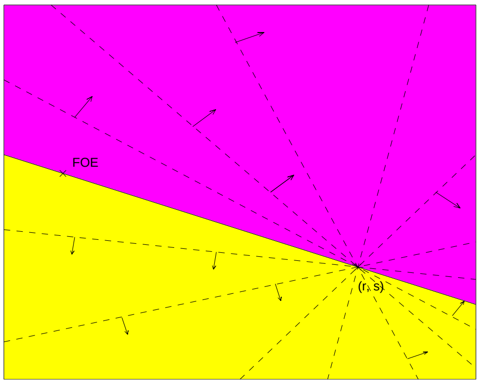
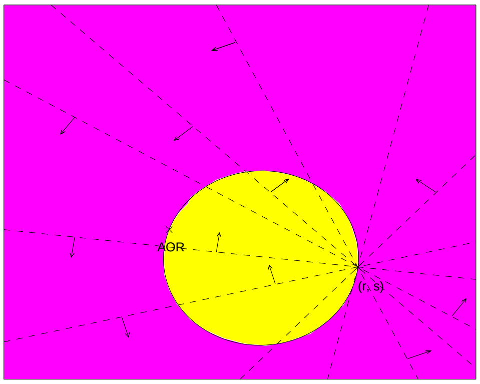


Figure 3: Positive (r, s) copoint vectors.

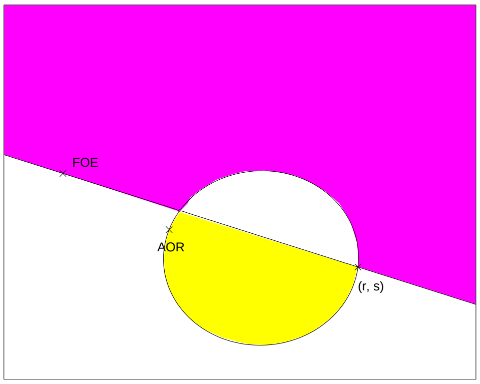
Every class of copoint vectors has the following property: Considering only translational vectors, we find that the positive and negative vectors are separated by a line. In one half-plane the vectors



(a)



(b)



(c)

Figure 4: (a) The translational (r, s) copoint vectors are separated by a line that passes through the FOE (the point which denotes the direction of translation); in one half-plane all vectors have positive values (light grey), in the other half-plane negative values (dark grey). (b) The rotational (r, s) copoint vectors are separated by a second order curve that passes through the AOR (the point where the rotation axis pierces the image plane). (c) A general rigid motion separates the (r, s) copoint vectors into an area of negative vectors, an area of positive vectors, and an area that may contain vectors of any value (white).

are positive, in the other the vectors are negative, and on the line they are zero (Figure 4a). Vectors due to rotation, on the other hand, are separated by a conic section into positive and negative ones (Figure 4b). Vectors of a general rigid motion (rotation and translation) thus obey the structure shown in Figure 4c. In one area the vectors are positive, in a second they are negative, and the vectors in the third area can take any value. This structure on the normal flow vectors is called the copoint pattern. Similar patterns exist for other classifications.

These findings allow us to formulate the problem of ego-motion estimation as a pattern recognition problem. By localizing for different classes of normal flow vectors the positive and negative areas in the image plane, the parameters for the axis of translation and direction of rotation can be derived [16].

Also, based on the same basic constraints, a process for the detection of independent motion has been designed. Since the observer is moving rigidly, an area with a motion field not possibly due to only one rigid motion must contain an independently moving object. The constraints are defined for the whole visual field, but also the motion vectors in every part of the image plane must obey a certain structure. Our approach consists of comparing the motion field within image patches with prestored patterns (which represent all possible rigid motions)

By considering patches of different sizes and using various resolutions, the patterns may also be of use in estimating the motion of objects. Differently sized filters can first be employed to localize the object and then an appropriately sized filter can be used to estimate the motion. Objects, however, do not always move rigidly. Furthermore, in many cases the area covered by the object will not be large enough to provide satisfying, accurate information. In the general case, when estimating an object's motion, only local information can be employed. In such a case, we utilize the observer's capability to move in a controlled way. We describe the object's motion with regard to an object centered coordinate system. From fixation on a small area on the object the observer can derive information about the direction of the object's translation parallel to its image plane. By tracking the object over a small amount of time, the observer derives additional information about the translation perpendicular to the image plane. Combining the computed values allows us to derive the direction of an object's translation [18].

3.5 A look at the motion pathway

There is a very large amount of literature [11, 38, 60, 65] on the properties of neurons involved in motion analysis. The modules which have been found to be involved in the early stages of motion analysis are the retinal parvocellular neurons, the magnocellular neurons in the LGN, layer 4Cβ of V1, layer 4B of V1, the thick bands of V2 and MT. These elements together are referred to as the early motion pathway. Among others they feed further motion processing modules, namely MST and FST, which in turn have connections to the parietal lobe. Here we concentrate on two striking features: the change of the spatial organization of the receptive fields and the selectivity of the receptive fields for motion over the early stages of the motion pathway. The computational modelling of the visual motion interpretation process that we described above appears consistent with our knowledge about the organization and functional properties of the neurons in the early stages motion pathway of the visual cortex. In addition our computational theory creates a hypothesis about the way motion is handled in the cortex and suggests a series of experiments for validating or rejecting it.

Figure 5 (from [40]) shows an outline of the process to be explained which involves four kinds of cells with different properties. In the early stages, from the retinal Pa ganglion cells through the magnocellular LGN cells to layer 4Ca of V1 the cells appear functionally homogeneous and respond almost equally well to the movement of a bar (moving perpendicularly to its direction) in any direction (Figure 5a). Within layer 4C of V1 we observe an onset of directional selectivity. The receptive fields of the neurons here are divided into separate excitatory and inhibitory regions. The regions are arranged in parallel stripes and this arrangement provides the neurons with a preference for a particular orientation of a bar target (which is displayed in the polar diagram) (Figure 5b). In layer 4B of V1 another major transformation takes place with the appearance of directional selectivity. The receptive fields here are relatively large and they seem to be excited everywhere by light or dark targets. In addition, these neurons respond better or solely to one direction of motion of an optimally oriented bar target, and less or not at all to the other (Figure 5c). Finally, in MT neurons have considerably large receptive fields and in general the precision of the selectivity for direction of motion that the neurons exhibit is typically less than in V1 (Figure 5d).

One can easily envision an architecture that, using neurons with the properties listed above, implements a global decomposition of the normal motion field. Neurons of the first kind could be involved in the estimation of the local retinal motion perpendicular to the local edge (normal

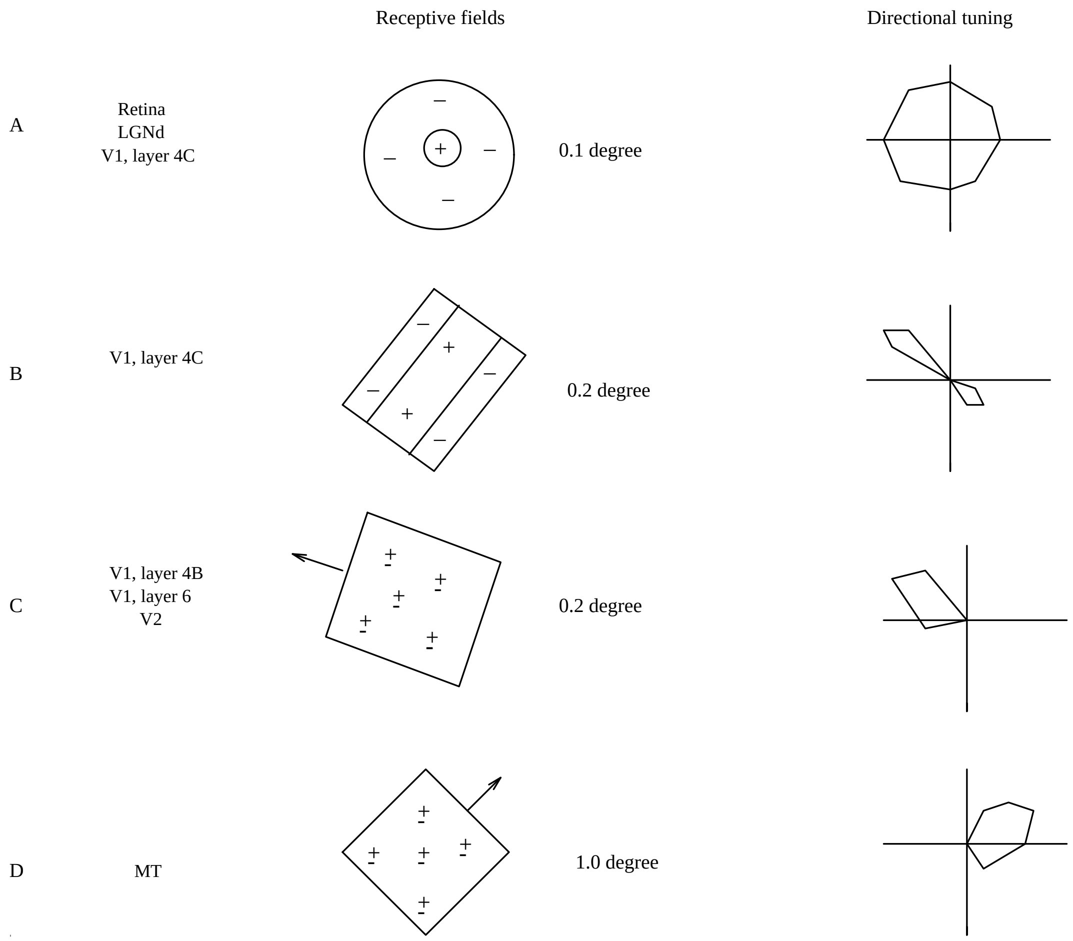


Figure 5: The spatial structure of visual receptive fields and their directional selectivity at different levels of the motion pathway (from [40]: The spatial scales of the receptive fields (0.1 degree, etc.) listed here are for neurons at the center of gaze; in the periphery these dimensions would be larger. The polar diagrams illustrate responses to variation in the direction of a bar target oriented at right angles to its direction of motion. The angular coordinate in the polar diagram indicates the direction of motion and the radial coordinate the magnitude of the response.

flow). Neurons at this stage could be thought of as computing whether the projection of retinal motion along some direction is positive or negative. Neurons of the second kind could be involved in the selection of local vectors in particular directions as parts of the various different patterns discussed in the previous section, while neurons of the third kind could be involved in computing the sign (positive or negative) of pattern vectors for areas in the image; i.e., they might compute for large patches of different sizes, whether the normal flow in certain directions is positive or negative. Finally, neurons of the last kind could be the ones that piece together the parts of the patterns developed already into global patterns that are matched with prestored global patterns. Matches provide information about egomotion and mismatches provide information about independent motion.

In this architecture we are not concerned with neurons that possibly estimate the motion field (optic flow). This is not to say that optic flow is not estimated in the cortex; several neurons could be involved in approximating the motion field. However, if the cortex is capable of solving some motion problems without the use of optic flow, whose estimation amounts to the solution of an optimization problem, it is quite plausible to expect that it would prefer such a solution. After all, it is important to realize that at the low levels of processing the system must utilize very reliable data, such as for example the sign of the motion field along some direction. It is worth noting that after deriving egomotion from normal flow, information about 3D motion is available, and the cortex could involve itself with approximating optic flow, because in this way the problem is not ill-posed any more (at least for background scene points).

3.6 Form-based competences

Since Computer Vision was considered to have as a goal the construction of 3D descriptions of the world, a lot of effort was spent on developing techniques for computing metric shape and depth descriptions from 2D imagery. Studies that are concerned with this kind of work are collectively referred to as "shape from X" computations, where by X is meant cues such as shading, texture, pattern, motion, or stereo. However, exact, quantitative 3D structure is hard to compute, and in the models employed, explicit assumptions about the scene (smoothness, planarity, etc.) usually have to be made.

Considering all the work that has been expended on the computation of metric shape and that has not yet given rise to any system working in a real environment, a glance at nature might give us some inspiration. Maybe it is a hopeless task to aim at deriving metric shape or depth information. Psychophysical experiments indicate that binocular stereopsis in the human visual system does not produce an explicit representation of the metric depth structure of the scene. Psychophysical evidence [10, 33] suggests that human performance in tasks involving metric structure from binocular disparities is very poor. Also, other cues don't seem to allow humans to extract the kind of depth information that has usually been considered. In their experiments, Todd and Reichel [62] had subjects estimate the depths of points on a drape-like surface shown on video images. Subjects could accurately report the relative depth of two points if they were on the same surface on the same side of the "fold", but were quite poor at determining the relative depth if the points were on different "folds". This experiment leads to the conclusion that humans possess a relative depth judgment for points within a local area lying on a surface; however, they cannot estimate even relative depth correctly for large distances in the visual field, when depth extrema are passed.

We also know that in humans the area of the eye in which detailed (high resolution) information can be extracted covers only a small region around the fovea (about five degrees of visual angle at normal viewing distance). The low resolution at the periphery does not allow us to derive accurate depth information. Human eyes, however, are seldom not in motion. The eyes are engaged in performing fixations, each lasting about 1/4 of a second. Between the fixations, saccadic movements are carried out, during which no useful information is extracted.

The biological evidence gives us good reason to argue for alternative shape models. The experiments mentioned above give rise to the following conclusions:

(a) Shape or depth should not be computed in metric form, but only relative depth measurements (ordered depth) should be computed.

(b) Shape/depth information should be computed only locally. Then the information derived for different patches has to be integrated. This integration, however, should not take place in the usual form, leading to complete, coherent spatial descriptions. The result should not be a complete reconstructed 3D shape model, obtained by exactly putting ("glueing") together the local shape representations into a global one. Instead, we have to look for alternative representations that suffice for accessing the shape information one needs to solve particular tasks.

These or similar arguments also find support from computational considerations. Concerning argument (b), one might ask why one should compute only local information, if from a technical standpoint there is no difference whether the sensors have different or the same resolution everywhere. If stereo systems are used—the most obvious for deriving shape information—and the two cameras fixate at a point, the disparity measurements are small only near the fixation point, and thus can also be computed exactly only there. In particular, if continuous techniques are employed to estimate the displacement (due to stereo or due to motion), the assumption of continuity of the spatio-temporal imagery does not have to be greatly violated. The measurements which are due to rotation increase with the distance from the image center and the translational measurements are proportional to the distance from the epipole or the point denoting the direction of translation. Another argument is that computing shape only locally gives legitimacy to the the orthographic projection model for approximating the image formation. The exact perspective projection model makes the computation of distance and shape very hard, since the depth component appears inversely in the image coordinates, which in turn leads to equations that are non-linear in the unknown parameters.

However, concerning argument (a), we don't want to prescribe the computation of ordered, as opposed to metric, shape information. Why should we limit ourselves to ordered depth and not be even less restrictive? Throughout this paper, we have argued for task-dependent descriptions. This also applies to shape descriptions; a variety of shape descriptions subserving different tasks can be accepted. To derive metric depth or shape means to compute exact values of the distance between the camera and the scene. In order to solve, for example, the general structure from motion problem, theoretically we require at least three views of the scene, or two views and some additional information, such as the length of the baseline for a stereo setting. From two perspective views, only scaled distance, or distance up to the so-called relief transformation, can be derived. To compute only ordered depth measurements would mean that in addition, scaled depth is derived only up to a positive term (i.e. it would result in deriving functions of the depth measurement Z of the form $f(Z) = \frac{1}{Z}a + b$, where a and b are constants). We argue that one could try to compute even less informative depth or shape information by aiming at deriving more complicated depth functions.

Under the influence of the reconstructionists' ideas, all effort in the past has been devoted to deriving metric measurements. A new look at the old research with a different goal in mind might give us new insights. From different cues, depth and shape information of different forms might be computed and then appropriately fused. A representation less than an ordered one by itself does not seem to be sufficient for 3D scene understanding. However, by combining two or more such representations, additional information can be obtained. It seems that the study of fusion of information for the purpose of deriving form and shape description will definitely be of importance.

It should be noted that whereas shape and depth measurements are equivalent for a metric 3D representation, they are not for ordered representations. Dealing with metric measurements, if absolute depth is given, shape (defined as the first order derivatives of depth) can be directly computed, and vice versa. The same, however, does not hold for ordered, or even less informative representations.

Our goal is to derive qualitative, as opposed to quantitative representations, because the computations to be performed should be robust. This requires that we don't make unreasonable assumptions and employ computations that are ill-posed. Qualitativeness, for example, does not mean performing the same computations that have been performed under the reconstruction philosophy, making the same assumptions about the 3D world, and at the end separating the computed values by a threshold in order to end up with "qualitative" information in the form of "greater or smaller than some value". Our effort should be devoted to deriving qualitative shape descriptions from well-defined input. For example, it wouldn't make sense to assume exact optical flow or stereo disparity measurements—which are impossible to obtain—in order to derive shape descriptions less powerful than the one of scaled depth. If we had exact 2D image measurements, we could compute scaled shape, and we would gain nothing computationally from computing less.

By concentrating on simpler shape descriptions, new mathematical models and new constraints might be found. Purely mathematical considerations can reveal what kind of information could possibly be computed from a certain input allowing a defined class of operations. The study of Koenderink and van Doorn [35] on affine structure from motion might serve as an inspiration; in it they investigated a hierarchy of shape descriptions based on a stratification of geometries.

3.7 Space understanding

Since in the past the actions of the observer were not considered as an integral part of perceptual investigations, computational modelling, and in particular AI research, has dealt with space only at a symbolic level. For example, some early systems [68] dealt with the spatial relationship of objects in a blocks world. Assuming that objects can be recognized and thus can be stored as symbols, the spatial configuration of these objects under changing conditions was studied. Also, in existing studies on spatial planning (e.g. path planning), solutions to the problems of recognizing the objects and the environment are assumed to be available for the phase of coordinating motions.

Within the framework of behavioral vision a new meaning is given to the study of space perception. The understanding of the space surrounding an observer results from the actions and perceptions the observer performs and their relationships. For a static observer that does not act in any way, space does not have much relevance. But in order to interact with its environment it has to have some knowledge about the space in which it lives, which it can acquire through actions and perceptions. Of course, the knowledge of space can be of different forms at various levels of complexity, depending on the sophistication of the observer/actor and the tasks it has to perform. At one end of the scale, we find a capability as simple as obstacle avoidance, which in the most parsimonious form has to capture only the distance between the observer and points in the 3D world; and at the other end of the scale, the competence of homing, which requires the actor to maintain some kind of map of its environment.

To obtain an understanding of space by visual means requires us to identify entities of the environment and also to localize their positions; thus both basic problems, the one of "where" and the one of "what", have to be addressed.

The problem of recognizing three-dimensional objects in space is by itself very difficult, since the object's appearance varies with the pose it has relative to the observer. In the Computer Vision literature two extreme views are taken about how to address the 3D recognition problem, which differ in the nature of the models to be selected for the descriptions of objects in the 3D environment. One view calls for object-centered models and the other for descriptions of the objects by means of viewer-centered views (3D vs. 2D models). In most of the work on object-centered descriptions the form of objects is described using simple geometric 3D models, such as polyhedra, quadrics, or superquadrics. Such models are suited to represent a small number of man-made (e.g. industrial) parts. However, to extend 3D modelling to a larger range of objects will require models of more complex structural description, characterizing objects as systems and parts of relations. Recently a number of studies have been performed on viewer-centered descriptions, approaching the problem from various directions. To name a few of them: Based on some results in the literature of structure from motion, that show that under parallel projection any view of an object can be constructed as a linear combination of a small number of views of the same object, a series of studies on recognition using orthographic and paraperspective projections have been conducted [31, 64]. The body of projective geometry has been investigated to prove results about the computation of structure and motion from a set of views under perspective projection [14]. The learning of object recognition capabilities has been studied for neuronal networks using nodes that store viewer-centered projections [52], and geometric studies on so-called aspect graphs have investigated how different kinds of geometric properties change with the views the observer has of the geometric model [34].

The problem of solving both localization and recognition is exactly the antagonistic conflict at the heart of pattern recognition. From the point of signal processing, it has been proved [20] that any single (linear) operator can answer only one of these questions with sufficient accuracy. In theory, thus, a number of processes are required to solve tasks related to space perception.

Results from the brain sciences reveal that the receptive field sizes of cells are much larger in the specialized visual areas involved in later processing than in those of the early stages. Many cells with large receptive field sizes respond equally well to stimuli at different positions. For example, in V5 cells with large receptive fields respond to spots of lights moved in certain directions, no matter where the stimulus in the receptive field occurs; nevertheless, the position of the light in the visual field can be localized accurately. Neurobiologists have suggested several solutions to this problem. Very interestingly, we find the following results: In the visual cortex cells have been found which are "gaze-locked", in the sense that they only respond to a certain stimulus if the subject is gazing in a particular direction. These cells probably respond to absolute positions in the ego-centric space [70].

It seems that nature has invented a number of ways for perceiving space through recognition and localization of objects in the 3D world. Also, neurophysiological studies have been conducted that give good reason to assume that the perception of space in primates is not only grounded on object-centered or ego-centered descriptions, but that some descriptions are with regard to some action. For example, in an area called TEA, cells have been reported which are involved in the coding of hand movements [50]. These cells respond when an action is directed towards a particular goal, but they do not respond to the component actions and motions when there is no causal connection between them. Monkeys were shown on video film arrangements of hand movements and object movements contiguous or separated in space or time—for example, a hand and a cup. The hand was retracted and after a short delay the cup moved (by itself) along the same trajectory as the hand. As the discrepancy between hand and object movement widened the impression of causality weakened. The cells tuned to hand actions were found to be less responsive when the movement of the hand and the object were spatially separated and appeared not to be causally related.

Humans possess a remarkable capability for recognizing situations, scenes, and objects in the space surrounding them from actions being performed. In the Computer Vision literature a number of experiments are often cited [32] in which it has been shown that humans can recognize specific animals and humans that move in the dark and are visible only from a set of light bulbs attached to their joints. These experiments demonstrate very well the power of motion cues. Since actions give rise to recognition, and actions are largely understood from motions, it seems to be worthwhile to investigate other motion models, more complicated than the rigid one, to describe actions. For examples, situations occurring in manipulation tasks might be modelled through non-rigid motion fields. The change of the motion field or parts of it may be expressed in form of space-time descriptions that can be related to the tasks to be performed. It should be mentioned that recently some effort along this line has started; a few studies have been conducted exploiting motion cues for recognition tasks. In particular, periodic movements, such as the motion of certain animal species, have been characterized in frequency space [43, 57]. Statistical pattern recognition techniques have been applied in the time domain to model highly structured motions occurring in nature, such as the motions of flowing water or fluttering leaves [54]. Attempts have been made to model walking or running humans by describing the motion of single limbs rigidly [55], and various deformable spatial models like superquadrics and snakes have been utilized to model non-rigid motions of rigid bodies [49], for example for the purpose of face recognition.

Representations used for understanding space should be allowed to be of any of three kinds: with regard to the viewer, with regard to an object, or action-driven. An appropriate representation might allow us to solve tasks straightforwardly that would require very elaborate computations and descriptions otherwise. Perrett et al. [51] give a good example supporting this point of view. A choreographer could, for example, use a set of instructions centered on the different dancers (such as to dancer M. who is currently lying prostrate and oriented toward the front of the stage "raise head slowly", and to dancer G., currently at the rear of the stage facing stage left, "turn head to look over left shoulder"). Alternatively the choreographer could give a single instruction to all members of the dance troupe ("Move the head slowly to face the audience"). To allow for the choice of different systems of representation will be a necessity when studying space descriptions. These descriptions, however, must be related in some form. After all, all measurements are taken in a frame fixed to the observer's eye. Thus a great deal of work in space understanding will amount to combining different representations into an ego-centered one.

The competence of homing is considered to be the apogee of spatial behavior. The amazing homing capabilities of some animals have attracted the attention of researchers for many years. In particular, effort has been spent on investigating the sensory basis of animals' perception; discoveries were made about the use of sensory guidance by sunlight, light patterns in the sky, and moonlight, such as the use of ultraviolet light by ants [36] and polarized light by bees [19]. Recently, research has also started on investigations of how particular species organize the spatial information acquired through their motor sequences and sensors [56, 59].

Zoologists differentiate two mechanisms of acquiring orientation: the use of ego-centered and geo-centered systems of reference. Simple animals, like most arthropods, represent spatial information in the form of positional information obtained by some kind of route integration relative to their homes. The route consists of path segments each of which takes the animal for a given distance in a given direction. This form of representation related to one point of reference is referred to as an ego-centered representation.[1] More complicated than relying on only information collected en route is the use of geo-centered reference systems where the animal in addition relies on information collected on site (recognition of landmarks) and where it organizes spatial information in a map-based form.

However, research from studies on arthropods [8, 9, 67] shows that already in these simple animals, the competence of homing is realized in many ways. A large variety of different ways employing combinations of information from action and perception have been discovered. In what way the path is stored, in what way landmarks are recognized, etc., is different for every species. Not many general concepts can be derived; it seems that the physical realizations are tightly linked to the animal's physiology and overall performance. This has to apply to artificial systems as well. Computations and implementations cannot be separated. Obviously, the more storage capability a system has, the more complex operations it can perform. The number of classes of landmarks that a system can differentiate and the number of actions it can perform will determine the homing capability of a system. Our suggested strategy is thus to address competences involving space representations (and in particular the homing competence) by synthesizing systems with increasing action and perception capabilities and study the performance of these systems, considering constraints on their memory.

4 Learning and memory

To reiterate, a flexible system that perceives and acts in its environment is considered here to consist of (a) competences, (b) representations, (c) action routines, and (d) learning programs. All these components can considered as maps. The visual competences are maps from retinotopic representations or space-time representations to other space-time representations. Actions are maps from space-time representations to motor commands or to other representations residing in memory. For any map $m : A \rightarrow B$ in the system, the learning programs are maps from $A \times B \rightarrow m$. Without loss of generality, and to be consistent with the literature, we will call A the stimulus set and B the response set, and we will consider the map m as a behavior. It might be the case, of course, that a behavior amounts to a composition of maps, but for simplicity we will use behavior and map interchangeably.

[1] In the Computer Vision literature the term "ego-centered" reference system is used with a different meaning than in Zoology.

The learning programs, like the competences, are not of a general nature and they help the various different maps of the system develop. Synthesizing mappings from examples is an instance of a supervised learning problem and it is basically a problem of interpolation. In this case the system is trained by being told the correct response to each stimulus. Since this cannot be done for all possible stimuli, the system will have to generalize from a set of sufficiently representative examples. When dealing, however, with the learning of difficult tasks, it is impossible to bound the stimulus space with a large enough set of examples, and the problem becomes one of extrapolation as opposed to interpolation. Standard neural networks (such as feed-forward networks) have been shown to yield unpredictable results when required to extrapolate [22].

Learning an input output behavior amounts to learning the probability distribution $p(a, b)$ of the input output pairs. Thus, given an input a_0, the system would pick an answer from the distribution $p(a_0, b)$. Learning the distribution $p(a, b)$ without any prior knowledge in statistical terms amounts to model-free estimation, i.e. no parametric models are assumed. The estimation error in model-free estimation can be of two kinds, one related to bias and one to variance. A system biased towards a certain behavior from which it is able to depart only to a certain extent will mainly suffer from an error in bias. A more flexible system will be able to reduce the error in bias, but as a consequence it will have to be punished with a large error due to variance. Geman et al. [22] claim that learning of complex tasks is essentially impossible without carefully introducing systems bias. In other words, it is essential that we choose an appropriate stimulus representation.

To learn the map $m : A \rightarrow B$ amounts then to learning the distribution $p(a, b)$ of the stimulus-response pairs. There are at least two different ways of doing this. One is to let the system model the average mapping from stimuli to responses and then apply some distribution with this function as its mean value. The parameters of the distribution may vary with the stimuli so that the system can be more or less certain of whether the mean value is appropriate. Another approach is to let the system estimate the distribution of the stimulus-response pairs in an analytic form. In this way, since $p(a \mid b) = \frac{p(a,b)}{p(b)}$, the conditional distribution (a posteriori density) can be calculated once the stimulus is known.

The implementation of both of the approaches outlined above can be realized in two different ways: a currently popular way and one not yet so popular. The popular technique refers to neural networks.

There is a large amount of literature on neural networks that estimate the average mapping from stimuli to responses [39, 53]. As described before, the success of this approach depends largely on whether the distribution of the examples is appropriate, i.e. on whether the network is required to interpolate or extrapolate. In the literature, one can also find the use of neural networks for estimating the distribution $p(a, b)$, to a limited extent, in the spirit of the kernel methods that have been developed in Statistics for density estimation. $p(a, b)$ is estimated by a weighted summation of a number of kernel functions that can be thought of as bumps. These kernels are usually placed at the sampled observations of the distribution, and then their positions, their parameters, and the weights of the summation are estimated.

The other, not so popular way to implement the learning of maps is to invent data structures that will be used to store information about the distribution of stimulus-response pairs in a way that will allow easy accessibility using the stimulus vector. The components of this vector will act as a key, or address, to the data structure. This could be considered as a memory structure much like the one in today's digital machines, and thus the problem of learning a map could be considered as a problem of memory organization. The stimulus-response space must be quantized, since the memory addresses are discrete and the number of memory locations is finite. Different data structures will decompose the stimulus-response space in different ways, producing different organizations of the memory. The goal will be to discover structures that adapt themselves to the underlying distribution so as to support fast access with little effort [45].

The available evidence from our knowledge about human and animal learning in vision, along with complexity arguments, suggests that the more adaptive of the hierarchical look-up structures are a more promising approach than standard feed-forward networks. The reason is that neural networks maintain a complete model of their domain. The model is wrong initially but gets better as more data comes in. The net deals with all the data in the same way and has no representation about uncertain or missing input. The architecture is chosen before the data is presented and the processing in the early phases of the training is similar to the processing in the later ones. Animal and human learning, on the other hand, seems to proceed in a different manner [46]. When a system has only few experiences in a domain, every experience (stimulus-response pair) is critical. Individual experiences are remembered more or less in detail in the early phases, and new responses are formed by generalizing from these small numbers of stored experiences. Later on, when more data becomes available, more complex models are formed, and there is no longer a need for storing individual experiences. Instead, the focus is concentrated on discovering regularities. In the first phase, thus, look-up structures are used for storing the distribution of the experiences, while the second phase resembles parameter fitting.

In addition, networks with global receptive fields seem to be unnatural for learning. In such networks, all neurons must participate in the computation, even if only a few contribute to the output. The effort to process an input in a fully connected network capable of exact classification of n memories is $O(n^2)$ [26]. In a one-layer net, each neuron defines a hyperplane. The input is processed by letting each neuron determine on which side of the hyperplane the input lies. This results in a waste of computational power since many of the comparisons will be unnecessary. It is thus preferable to make use of a structure with local receptive fields so that divide and conquer techniques can be applied, i.e. every time a piece of information is derived about the data it is used to prune away unnecessary further computations. Using hierarchical look-up structures [45] the computational time can be reduced on the average to $O(\log n)$.

The problem we address next is what maps to learn. If we are interested in learning the map of a complex task, where the stimulus and response spaces are well defined, we should keep in mind that the map (task) in which we are interested might be the composition of a set of "simpler" maps which it could be easier to synthesize or learn from experience (task decomposition). Knowledge from the field of Neurobiology can provide inspiration for hypothesizing a set of simpler maps. On the other hand, geometry will provide us with the constraints and models for many problems. There obviously exist different kinds of maps, some which we provide the system with (or which the system is born with), and some that the system learns from experience. But it would not be of much help to try to learn maps that we can model, for example the map that recognizes egomotion, when we understand the geometry behind this process. Otherwise, we would end up trying to mimic evolution.

An intelligent system consists of a set of maps such as described above. With the exception of the learning programs and of some of the action routines, these maps relate different space-time representations. It is the geometry of space and the physics of light that together with the learning programs will contribute to our understanding of these maps, with learning taking over whenever the geometric or physical models required become too complicated. This is consistent with our principles behind the synthetic approach. We should develop the competences in the order of the complexity of the underlying models, starting from simple models and moving to more complex ones. When the models become too complex, we should let experience take over the building of the map, through learning.

5 Conclusions

The study of vision systems in a behavioral framework requires the modelling of observer and world in a synergistic way and the analysis of the interrelationship of action and perception. The role that vision plays in a system that interacts with its environment can be considered as the extraction of representations of the space-time in which the system exists and the establishing of relations between these representations and the system's actions. We define a vision system as consisting of a number of representations and processes, or on a more abstract level, as a set of maps which can be classified into three categories: the visual competences that map different representations of space-time (including the retinotopic ones) to each other, the action routines which map space-time representations to motor commands or representations of various kinds residing in memory, and the learning programs that are responsible for the development of any map. To design or analyze a vision system amounts to understanding the mappings involved. In this paper we have provided a framework for developing vision systems in a synthetic manner, and have discussed a number of problems concerning the development of competences, learning routines and the integration of action and perception. We have also described some of our technical work on the development of specific motion-related competences.

To achieve an understanding of vision will require efforts from various disciplines. We have described in this study work from a number of sciences, mainly empirical ones. Besides these, the general area of Information Processing also has to offer various ideas from which the design and analysis of vision systems can benefit. Areas that might be of interest include the realization of specific maps in hardware (VLSI chips or optical computing elements), the study of the complexity of visual tasks under the new framework, and information theoretic studies investigating the relationship between memory and task-specific perceptual information. We also have not discussed here the control mechanisms for behavioral systems. A promising framework for implementing purposive, behavioral systems that act and perceive over time is that of hybrid automata, which allow the modelling of sequences of continuous as well as discrete processes.

References

[1] J. Aloimonos and C. Brown. Direct processing of curvilinear sensor motion from a sequence of perspective images. In *Proc. Workshop on Computer Vision: Representation and Control*, pages 72–77, 1984.

[2] J. Aloimonos, I. Weiss, and A. Bandopadhay. Active vision. *International Journal of Computer Vision*, 2:333–356, 1988.

[3] J.Y. Aloimonos. Purposive and qualitative active vision. In *Proc. DARPA Image Understanding Workshop*, pages 816–828, 1990.

[4] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76:996–1005, 1988.

[5] D. Boussaud, L. Ungerleider, and R. DeSimone. Pathways for motion analysis: cortical connections of the medial superior temporal fundus of the superior temporal visual areas in the macaque monkey. *Journal of Comparative Neurology*, 296:462–495, 1990.

[6] M. Brady, J. Hollerbach, T. Johnson, T. Lozano-Perez, and M. Mason, editors. *Robot Motion.* MIT Press, Cambridge, MA, 1983.

[7] R. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2:14–23, 1986.

[8] T. Collett, E. Dillmann, A. Giger, and R. Wehner. Visual landmarks and route following in desert ants. *Journal of Comparative Physiology A*, 170:435–442, 1992.

[9] T. Collett, S. Fry, and R. Wehner. Sequence learning by honeybees. *Journal of Comparative Physiology A*, 172:693–706, 1993.

[10] T. Collett, U. Schwartz, and E. Sobel. The interaction of oculomotor cues and stimulus size in stereoscopic depth constancy. *Perception*, 20:733–754, 1991.

[11] C. Duffy and R. Wurtz. Sensitivity of MST neurons to optical flow stimuli I: a continuum of response selectivity to large field stimuli. *Journal of Neurophysiology*, 65:1329–1345, 1991.

[12] G. Ernst and A. Newell. *GPS: A Case Study in Generality and Problem Solving.* Academic Press, New York, 1969.

[13] M. Farah. *Visual Agnosia: Disorders of Object Recognition and What They Tell us about Normal Vision.* MIT Press, Cambridge, MA, 1990.

[14] O. Faugeras. *Three Dimensional Computer Vision.* MIT Press, Cambridge, MA, 1992.

[15] C. Fermüller. *Basic Visual Capabilities*. PhD thesis, Institute for Automation, University of Technology, Vienna, Austria, available as Technical Report CAR-TR-668, Center for Automation Research, University of Maryland, 1993.

[16] C. Fermüller. Navigational preliminaries. In Y. Aloimonos, editor, *Active Perception*, Advances in Computer Vision. Lawrence Erlbaum, Hillsdale, NJ, 1993.

[17] C. Fermüller and Y. Aloimonos. Tracking facilitates 3-d motion estimation. *Biological Cybernetics*, 67:259–268, 1992.

[18] C. Fermüller and Y. Aloimonos. The role of fixation in visual motion analysis. *International Journal of Computer Vision*, Special Issue on Active Vision, M. Swain (Ed.), 11:165–186, 1993.

[19] K. Frisch. Die Polarisation des Himmelslichts als orientierender Faktor bei den Tänzen der Bienen. *Experientia*, 5:142–148, 1949.

[20] D. Gabor. Theory of communication. *Journal of the IEE*, 93 (part III):429–457, 1946.

[21] H. Gelernter. Realization of a geometry theorem-proving machine. In *Information Processing: Proceedings of the International Conference on Information Processing*, UNESCO, 1959.

[22] E.B.S. Geman and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

[23] M. Goodale, A. Milner, L. Jacobson, and D. Carey. A neurological dissociation between perceiving objects and grasping them. *Nature*, 349:154–156, 1991.

[24] R. Gregory. Distortion of visual space as inappropriate constancy scaling. *Nature*, 119:678, 1963.

[25] J. Hervé. *Navigational Vision*. PhD thesis, University of Maryland, Computer Vision Laboratory, Center for Automation Research, University of Maryland, 1993.

[26] J. Hopfield. Neural networks and physical systems with emergent collective computational capabilities. *Proceedings of the National Academy of Sciences*, 79:2554–2558, 1982.

[27] B. Horn and E. Weldon. Computationally efficient methods for recovering translational motion. In *Proc. International Conference on Computer Vision*, pages 2–11, 1987.

[28] G. Horridge. The evolution of visual processing and the construction of seeing systems. *Proceedings of the Royal Society, London B*, 230:279–292, 1987.

[29] G. Horridge. Evolution of visual processing. In J. Cronly-Dillon and R. Gregory, editors, *Vision and Visual Dysfunction*. MacMillan, New York, 1991.

[30] G. Humphreys and M. Riddoch. *To See But Not To See: A Case Study of Visual Agnosia*. Lawrence Erlbaum, Hillsdale, New Jersey, 1992.

[31] D. Jacobs. Space efficient 3d model indexing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 439–444, 1992.

[32] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[33] E. Johnston. Systematic distortions of shape from stereopsis. *Vision Research*, 31:1351–1360, 1991.

[34] J. Koenderink and A. van Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32:211–216, 1979.

[35] J. Koenderink and A. van Doorn. Affine structure from motion. *Journal of the Optical Society of America*, 8:377–385, 1991.

[36] J. Lubbock. *On the Senses, Instincts, and Intelligence of Animals with Special Reference to Insects*. K. Paul Trench, London, 1889.

[37] D. Marr. *Vision*. W.H. Freeman, San Francisco, 1982.

[38] J. Maunsell and D.V. Essen. Functional properties of neurons in middle temporal visual area of the macaque monkey I. Selectivity for stimulus direction, speed and orientation. *Journal of Neurophysiology*, 49:1127–1147, 1983.

[39] J. Moody and C. Darken. Fast learning in networks of locally tuned processing units. *Neural Computation*, 1:281–293, 1989.

[40] A. Movshon. Visual processing of moving images. In H. Barlow, C. Blakemore, and M. Weston-Smith, editors, *Images and Understanding*, pages 122–137. Cambridge University Press, 1990.

[41] R. Nelson. Qualitative detection of motion by a moving observer. *International Journal of Computer Vision*, 7:33–46, 1991.

[42] R. Nelson and J. Aloimonos. Finding motion parameters from spherical flow fields (or the advantage of having eyes in the back of your head). *Biological Cybernetics*, 58:261–273, 1988.

[43] R. Nelson and R. Polana. Qualitative recognition of motion using temporal texture. *CVGIP: Image Understanding*, Special Issue on Purposive, Qualitative, Active Vision, Y. Aloimonos (Ed.), 56:78–89, 1992.

[44] N. Nilsson. *Principles of Artificial Intelligence.* Tioga Publishing Co., Palo Alto, CA, 1980.

[45] S. Omohundro. Bumptrees for efficient function, constraint and classification learning. Technical report, International Computer Science Institute, Berkeley, CA, 1991.

[46] S. Omohundro. Best-first model merging for dynamic learning and recognition. Technical report, International Computer Science Institute, Berkeley, CA, 1992.

[47] G. Orban. The analysis of motion signals and the nature of processing in the primate visual system. In G. Orban and H.-H. Nagel, editors, *Artificial and Biological Vision Systems*, ESPRIT Basic Research Series, pages 24–57. Springer-Verlag, 1992.

[48] A. Pentland, editor. *From Pixels to Predicates: Recent Advances in Computational and Robot Vision.* Ablex, Norwood, NJ, 1986.

[49] A. Pentland, B. Horowitz, and S. Sclaroff. Non-rigid motion and structure from contour. In *Proc. IEEE Workshop on Visual Motion*, pages 288–293, 1991.

[50] D. Perrett, M. Harries, A. Mistlin, and A. Chitty. Three stages in the classification of body movements by visual neurons. In H. Barlow, C. Blakemore, and M. Weston-Smith, editors, *Images and Understanding*, pages 94–107. Cambridge University Press, 1990.

[51] D. Perrett, A. Mistlin, and M.H.A. Chitty. Vision and action: The control of grasping. Ablex, Norwood, NJ, 1988.

[52] T. Poggio, S. Edelman, and M. Fahle. Learning of visual modules from examples: A framework for understanding adaptive visual performance. *CVGIP: Image Understanding*, Special Issue on Purposive, Qualitative, Active Vision, Y. Aloimonos (Ed.), 56:22–30, 1992.

[53] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78:113–125, 1990.

[54] R. Polana and R. Nelson. Detecting activities. In *Proc. IEEE Image Understanding Workshop*, pages 569–574, 1993.

[55] R. Qian and T. Huang. Motion analysis of articulated objects. In *Proc. International Conference on Pattern Recognition*, pages A220–223, 1992.

[56] G. Sandini, F. Gandolfo, E. Grosso, and M. Tistarelli. Vision during action. In Y. Aloimonos, editor, *Active Perception*, pages 151–190. Lawrence Erlbaum, Hillsdale, NJ, 1993.

[57] E. Shavit and A. Jepson. Motion using qualitative dynamics. In *Proc. IEEE Workshop on Qualitative Vision*, 1993.

[58] G. Sommer. Architektur und Funktion visueller Systeme. Künstliche Intelligenz, 12. Frühjahrsschule, March 1994.

[59] M. Srinivasan, M. Lehrer, S. Zhang, and G. Horridge. How honeybees measure their distance from objects of unknown size. *Journal of Comparative Physiology A*, 165:605–613, 1989.

[60] K. Tanaka and H. Saito. Analysis of motion of the visual field by direction, expansion/contraction, and rotation cells illustrated in the dorsal part of the Medial Superior Temporal area of the macaque monkey. *Journal of Neurophysiology*, 62:626–641, 1989.

[61] W. Thompson and T.-C. Pong. Detecting moving objects. *International Journal of Computer Vision*, 4:39–57, 1990.

[62] J. Todd and Reichel. Ordinal structure in the visual perception and cognition of smoothly curved surfaces. *Psychological Review*, 96:643–657, 1989.

[63] S. Ullman. *The Interpretation of Visual Motion.* MIT Press, Cambridge, MA, 1979.

[64] S. Ullman and R. Basri. Recognition by linear combination of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1006, 1991.

[65] L. Ungerleider and R. DeSimone. Cortical connections of visual area MT in the macaque. *Journal of Comparative Neurology*, 248:190–222, 1986.

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

[67] R. Wehner. Homing in arthropods. In F. Papi, editor, *Animal Homing*, pages 45–144. Chapman and Hall, London, 1992.

[68] P. Winston. Learning structural descriptions from examples. In P. Winston, editor, *The Psychology of Computer Vision*. McGraw-Hill, New York, 1975.

[69] S. Zeki. The visual image in mind and brain. *Scientific American*, 267(3):69–76, 1992.

[70] S. Zeki. *A Vision of the Brain*. Blackwell Scientific Publications, 1993.