ACTIVE PERCEPTION

"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 . . ." (1). 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 (2,3,3a,3b) 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 (4), 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 two-dimensional (2-D) images through the primal sketch and the $2\frac{1}{2}$-D sketch to object-centered descriptions ("from pixels to predicates") (5). Marr suggested that visual processes—or any perceptual or cognitive processes—are information-processing tasks and thus should be analyzed at three levels: (1) at the computational theoretic level (definition of the problem and its boundary conditions; formulation of theoretical access to the problem), (2) at the level of selection of algorithms and representations (specification of formal procedures for obtaining the solution), and (3) 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 (AI) and thus studied using the same methodology, influenced by the ideas and computational theories of the last decades (6–8).

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 has not been much work on "vision and learning") (9). 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 modeling of observer and world (as closed systems) but the modeling of observer and world in a synergistic manner will contribute to the understanding of perceptual information-processing systems (10). The question, of course, still remains how such a synergistic modeling 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?

WHERE ARE WE HEADING?

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 twentieth 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 modeling 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 (11)], information about its design can be deduced. Recent results from cognitive neurophysiology—the discipline that is concerned, among other topics, with the study of visual agnosia (a condition exhibited by patients with partially damaged brains) (12,13)—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, especially 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.

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. This is not the case: 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 twofold: 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 scientists what experiments to perform to find out how biological organisms can possibly function.

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.

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 representa-

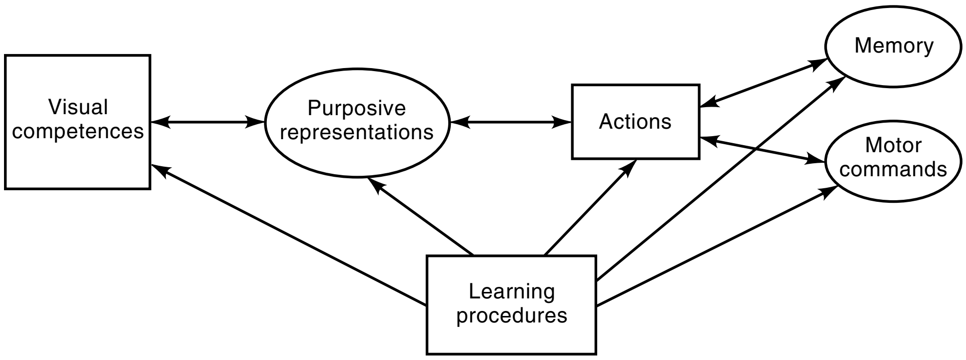

Figure 1. Working model: Basic components of a purposive vision system.

tions 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 that 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, that is, they initialize motor commands and thus provide the interface to the body. 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; that is, 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 Fig. 1, 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 model just described captures the study of perception and action in a synergistic way and address some of the questions posed at the beginning of the article: 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 that 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 that might either be visible externally or take place only internally in the form of parameter adaptation.

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 have not understood how we actually could develop visual competences for systems that work in environments as complex as our own, so we will not be able to obtain a global view of the overall architecture and functionality of vision systems. At this point in time it also would not 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 with 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 would not 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. We give here a short outline of the ideas behind this approach, which we discuss in detail in the remainder of the article. 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, and 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.

THE COMPETENCES

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 that 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 that 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; 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 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 gain increasingly insight into the process of vision, which is of high complexity. Capabilities that 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 that 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 that are applicable in general environments. Each such entity recognized constitutes a category relevant to the system. Some examples from navigation are as follows.

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 nonrigid flow fields for a moving observer partially knowing its motion (14–16). 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 3-D 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

(17). 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 (18). 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 (14,19).

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 derive directly the necessary parameters by recognizing a set of locations on this representation along with 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.

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, 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 has some similarities at the philosophical level with Brooks's proposal about understanding intelligent behavior through the construction of working mechanisms (20). In proposing the subsumption architecture, Brooks suggested a hierarchy of of competences such as avoiding contact with objects, exploring the world by seeing places, and reasoning about the world in terms of identifiable objects. 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.

Biological Hierarchy

It remains to discuss the actual simple capabilities on which we should concentrate our first efforts. Other scientific disciplines give us some answers. 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 (21,22), 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 Ref. 23, 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 that use both static and dynamic information. Zeki (24), 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.

On the basis of anatomical evidence and behavioral studies (studies on patients with lesions of specific cortical areas) the hypothesis has been advanced (25) that there exist two visual pathways originating from V1: a dorsal one leading to the parietal cortex and a ventral one leading to the inferotemporal cortex. The dorsal path is concerned with either the computations concerned with "where" (object localization) or "how" [the visual guidance of movements (26)], 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 (24); one of the reasons is that this theory fails to

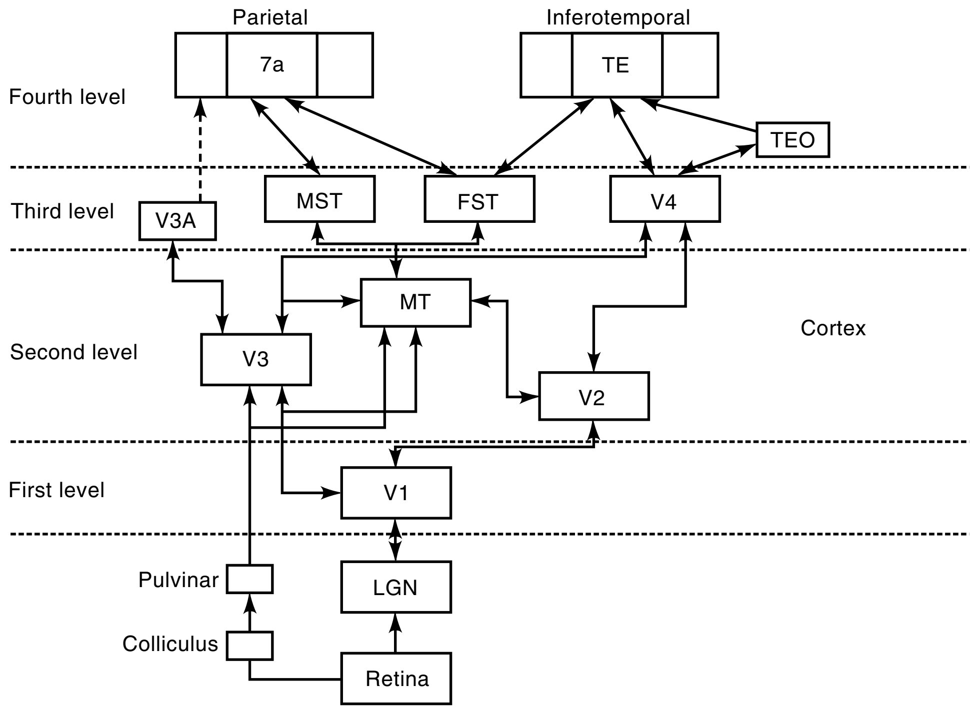

Figure 2. Diagram of the primate visual system indicating the subcortical structure as well as the four tentative levels of cortical visual processing (from Ref. 23).

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 (27).

Results from the brain sciences show us that there is not just one hierarchy of visual processes, but various different computations are performed in parallel. Also, it is not 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 3-D 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 3-D only locally. Therefore, it will be necessary to study new forms of shape representations. In nature too there is not 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 (28) 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 modeling such competences by means of more complicated motion models (nonrigid-motion models).

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 in which 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 (29). The idea was to recover the relative 3-D 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 simplest navigational competence, according to our definition, is the estimation of ego motion. The observer's sensory apparatus (eye or 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 (30). 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 ego motion 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 in which the observer is engaged, we argue that in many cases when an object is moving in an unrestricted manner (translation and rotation) in the 3-D world, we are only interested in the object's translational component, which can be extracted using dynamic fixation (31).

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 should not discourage us and drive us to the conclusion that ego-motion estimation and independent-motion detection are chicken-and-egg problems: unless one of them has been solved, the other can not be addressed either. Have you ever experienced the illusion that you are sitting in front of a wall that covers most of your visual field, and suddenly this wall (which actually is not a wall) starts to move? You seem to experience 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 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 (15). If some upper bound on the observer's motion is known (maximal speed), it is possible to detect motion 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 ego motion, 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 ego motion 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.

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 that 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 ego-motion estimation is a prerequisite for obstacle avoidance mechanisms, and general independent-motion detection will require a model that is as complex as that used in ego-motion 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 ego motion 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.

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 previously mentioned competences and differentiate 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 3-D 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 nonrigidly. However, considering only a small patch in the image of a moving object, a rigid-motion approximation is legitimate. For the case of ego motion, 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 3-D 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 spatiotemporal 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 (32,33)].

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 that relate the spatiotemporal derivatives of the image intensity function globally to the 3-D 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 ($\boldsymbol{r}$, $\boldsymbol{s}$). These so-called copoint vectors ($\boldsymbol{r}$, $\boldsymbol{s}$) are vectors that are perpendicular to straight lines passing through the point ($\boldsymbol{r}$, $\boldsymbol{s}$). In addition, the normal-flow vectors of a class are distinguished as to whether their direction is counterclockwise or clockwise with respect to ($\boldsymbol{r}$, $\boldsymbol{s}$), in which case they are called positive or negative (see Fig. 3). Since any point ($\boldsymbol{r}$, $\boldsymbol{s}$) in the image can be chosen as a reference point, there exists an infinite number of such classifications.

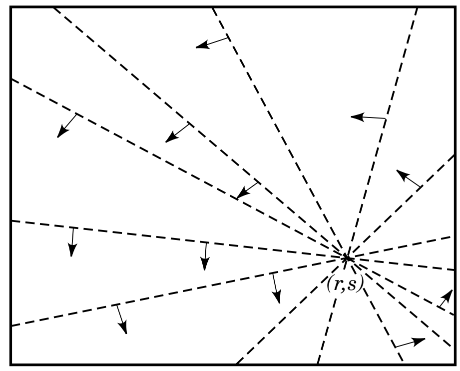

Figure 3. Positive ($\boldsymbol{r}$, $\boldsymbol{s}$) copoint vectors.

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 are positive, in the other the vectors are negative, and on the line they are zero [Fig. 4(a)]. Vectors due to rotation, on the other hand, are separated by a conic section into positive and negative ones [Fig. 4(b)]. Vectors of a general rigid motion (rotation and translation) thus obey the structure shown in Fig. 4(c). 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 normaflow vectors is called the copoint pattern. Similar patterns exist for other classifications (34,35).

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 (30).

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,

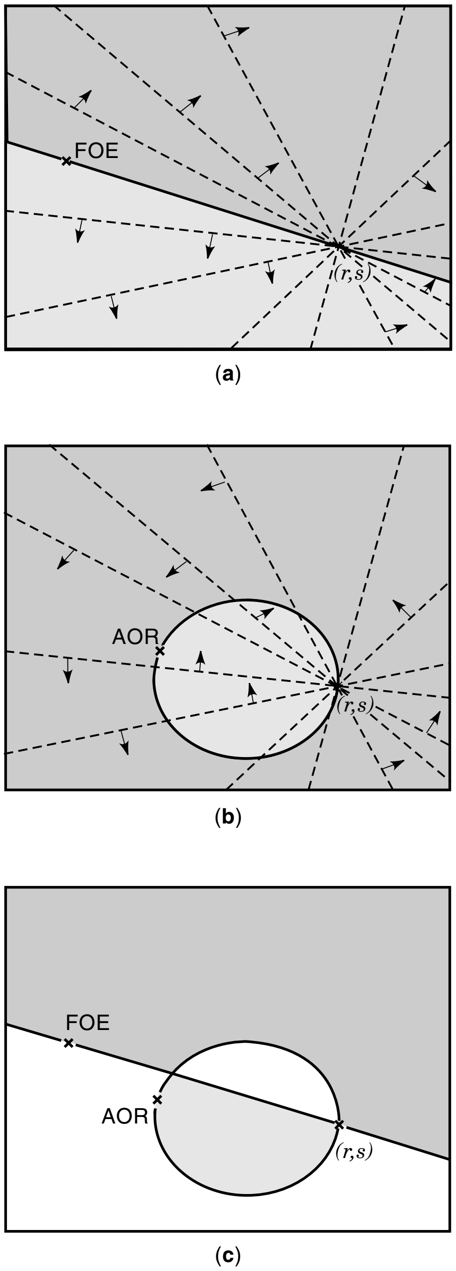

Figure 4. (a) The translational ($\boldsymbol{r}$, $\boldsymbol{s}$) copoint vectors are separated by a line that passes through the FOE (the point that denotes the direction of translation); in one half-plane all vectors have positive values (light gray), in the other half-plane negative values (dark gray). (b) The rotational ($\boldsymbol{r}$, $\boldsymbol{s}$) 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 ($\boldsymbol{r}$, $\boldsymbol{s}$) copoint vectors into an area of negative vectors, an area of positive vectors, and an area that may contain vectors of any value (white).

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 (36). Several recent results have strengthened this framework (37–48).

A Look at the Motion Pathway

There is a very large amount of literature (49–52) on the properties of neurons involved in motion analysis. The modules that 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 modeling 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-stage-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 Ref. 53) shows an outline of the process to be explained that 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 [Fig. 5(a)]. 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) [Fig. 5(b)]. 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 [Fig. 5(c)]. 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 that in V1 [Fig. 5(d)]. In MST the size of the receptive fields of neurons becomes even larger, ranging from 30δ to 100δ, each responding to particular 3-D motion configurations (23,49,51).

One can easily envision an architecture that, using neurons with the properties previously listed, 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 flow).

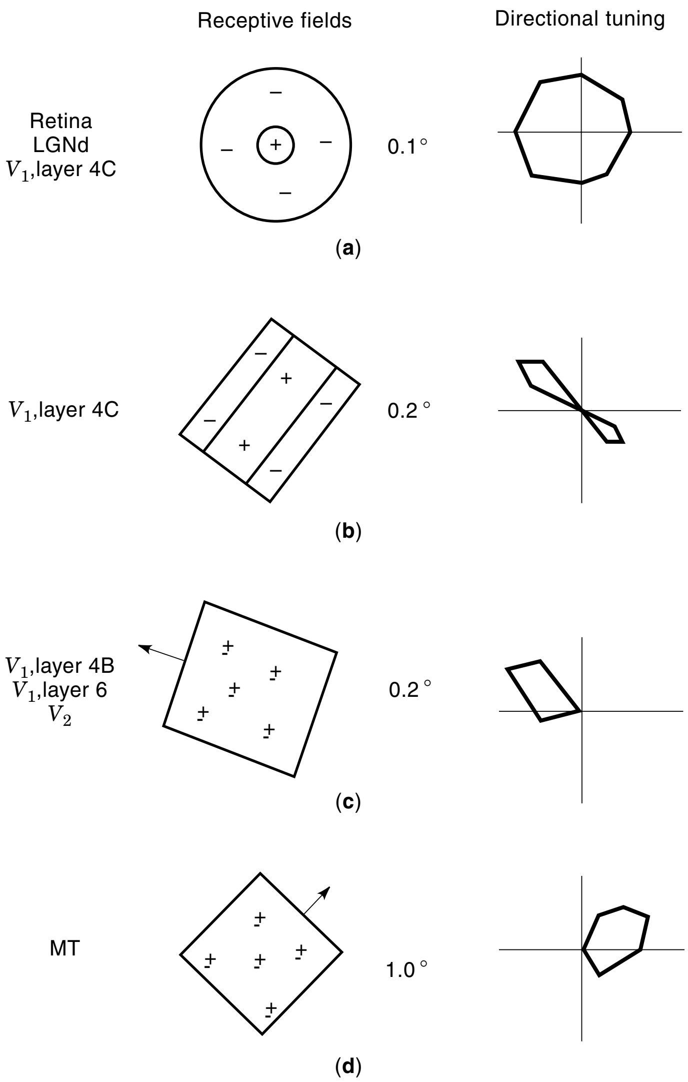

Figure 5. The spatial structure of visual receptive fields and their directional selectivity at different levels of the motion pathway, from Ref. 53. The spatial scales of the receptive fields (0.1°, etc.) listed here are for neurons at the center of gaze; in the periphery these dimensions would be larger. 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.

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 preceding 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; that is, 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 (MT and MST) 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 ego motion 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 ego motion from normal flow, information about 3-D 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).

Form-Based Competences

Since computer vision was considered to have as a goal the construction of 3-D descriptions of the world, a lot of effort was spent on developing techniques for computing metric shape and depth descriptions from 2-D 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, an exact, quantitative 3-D 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 (54,55) suggests that human performance in tasks involving metric structure from binocular disparities is very poor. Also, other cues do not seem to allow humans to extract the kind of depth information that has usually been considered. In their experiments, Todd and Reichel (56) had subjects estimate the depths of points on a drapelike 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 5° 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 one-quarter 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 before give rise to the following conclusions:

1. Shape or depth should not be computed in metric form, but only relative depth measurements (ordered depth) should be computed.
2. A complete shape or depth map relating every pixel to every other pixel should not be computed globally but only for parts of the image. Then the information derived for different parts 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 3-D shape model, obtained by exactly putting ("gluing") 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 2, 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 spatiotemporal imagery does not have to be greatly violated. The measurements that 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 nonlinear in the unknown parameters.

However, concerning argument 1, we do not just 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 article, 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

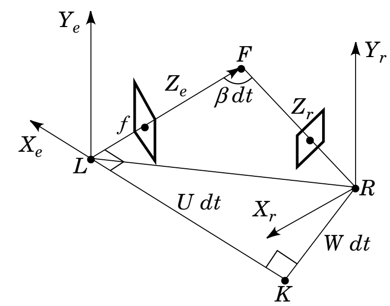

Figure 6. The translation along the X_l axis is $LK = U\,dt$. The translation along the Z_l axis is $KR = W\,dt$. The angle denoting rotation around the Y_l axis is $LFR = \beta\,dt$. L, K, R, F belong to the fixation plane. dt is a hypothetical small time interval during which the motion bringing $X_lY_lZ_l$ to $X_rY_rZ_r$ takes place.

deriving functions of the depth measurement Z of the form $f(Z) = (1/Z)a + b$, $f(Z) = aZ + b$, $f(Z) = e^{aZ} + b$, etc., where a and b are unknown constants] (57,58). We argue that one could try to compute even less informative features than metric depth or shape information by aiming at deriving more involved depth functions.

An example is given here from binocular vision. Given a fixated stereo pair, we can choose normal disparities (projections of disparity vectors on the orientation of the local image gradient) in such a way that the values of the normal disparities are sufficient for ordering the depth values. Consider an active binocular observer capable of fixating on an environmental point. The geometry of the system can be described as a constrained rigid motion between the left and right eye. If we fix a coordinate on the left eye with the z axis aligned with its optical axis and the y axis perpendicular to the fixation plane, then the transformation relating the right eye to the left is a rotation around the y axis and a translation in the xz plane (Fig. 6). At the fixation point the disparity measurements are zero and in a neighborhood around it relatively small. Thus, it is legitimate to approximate the disparity measurements through a continuous velocity field. This amounts to the small baseline approximation that has been used in the literature (58).

Denoting, as usual, by U and W the translation along the x and z axes and by β the rotation around the y axis, and setting x_0 equal to $(U/W)f$ (the x coordinate of the focus of expansion, where f is the focal length), the component u_n of the disparity vector $(\boldsymbol{u}, \boldsymbol{v})$ along the gradient direction (n_x, n_y) is

$$u_n = \frac{W}{Z}(-x_0 n_x + x n_x + y n_y) - \beta\left(f n_x + \frac{x^2}{f} n_x + \frac{xy}{f} n_y\right) \quad (1)$$

The exact geometry of the stereo configuration cannot be assumed and we do not wish to attempt the usual two-step approach of first computing it from the available information in order to utilize it and derive in a second step the depth estimates. The reason is that small errors in the parameter estimation of the extrinsic geometry can result in large errors in the depth or shape estimates.

We show below how to obtain ordinal depth from one fixation. Additional fixations provide more information that can be fused into a single representation.

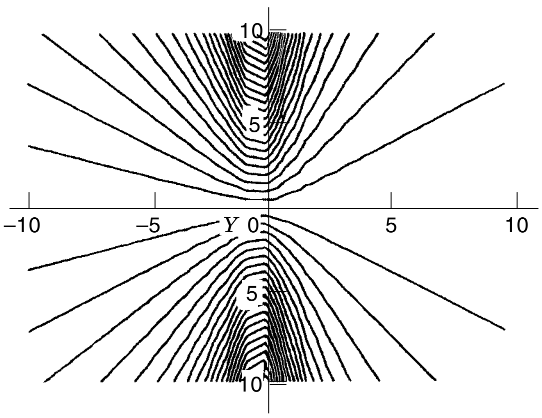

Figure 7. Shape computation from one pair of images taken by a binocular fixating vision system: A partial ordering of the depth values can be obtained for all points with edges tangential (or normal disparity measurements perpendicular) to the curves of a family.

An active binocular stereo system capable of changing its geometric parameters in a controlled way should be aware of the pose of its eyes with regard to some head-frame-centered coordinate system. Thus it should know the angle the optical axis makes with the baseline, which amounts to knowing the parameter x_0. If for a particular system this knowledge is not available, utilizing the constraints described in the section entitled "Motion-Based Competences," the direction of the translation x_0 can be derived from the patterns of the normal disparity field, utilizing only the sign of the disparity measurements.

We do not know, however, the amount of rotation β and we also do not have to know the distance between the two eyes. Using Eq. (1) it is possible to obtain an ordinal depth representation for the scene whose image points lie on families of curves. Dividing Eq. (1) by $-n_x$, we obtain

$$-\frac{u_\mathrm{n}}{n_x} = \frac{W}{Z}\left(x_0 - x - y\,\frac{n_y}{n_x}\right) + \beta\left(f + \frac{x^2}{f} + \frac{xy}{f}\frac{n_y}{n_x}\right) \qquad (2)$$

We define a classification in the gradient directions n_y/n_x in such a way that for each single class the ratio of the coefficients of W/Z and β in Eq. (2) is a constant C everywhere in the image plane. Consequently, for the vectors of each class an appropriately normalized value of the normal disparity u_n can be written as a linear function in the inverse depth with unknown coefficients. However, this allows the estimation of ordinal depth. To give a geometric interpretation to the selection of classes, the normal disparity vectors in each class are perpendicular to edges in the image that can be derived from a differential equation. Figure 7 shows the integral curves for one class of a particular stereo configuration ($f = 1$, $x_0 = 1$, $C = 0.2$).

From one stereo pair we can obtain partial ordinal depth maps. Additional fixations, since they are only separated by rotations, allow the comparison of depth values corresponding to different classes that are to be found in the same class in some other fixated disparity pair. This way, merging classes and building one ordinal depth map becomes possible (42).

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 that is less than an ordered one by itself does not seem to be sufficient for 3-D 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 a form and shape description will definitely be of importance.

It should be noted that whereas shape and depth measurements are equivalent for a metric 3-D 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 do not 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 3-D 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 a well-defined input. For example, it would not 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 those of scaled depth. If we had exact 2-D 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 the kind of information that could possibly be computed from a certain input allowing a defined class of operations. The study of Koenderink and van Doorn (59) 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.

Space Understanding

Since in the past the actions of the observer were not considered as an integral part of perceptual investigations, computational modeling, and in particular AI research, has dealt with space only at a symbolic level. For example, some early systems (60) dealt with the spatial relationship of objects in a block 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