Introduction: Active Vision Revisited

Y. Aloimonos.
University of Maryland

What is, is identical with the thoughts of the one who recognizes what it is.

–Parmenides

Abstract

This book is devoted to technical problems related to the design and analysis of intelligent systems possessing perception, like the existing biological organisms and the "seeing" machines of the future. Since the appearance of the first technical results on Active Vision [2, 5], researchers are beginning to realize that perception (and intelligence in general) is not transcendental and disembodied, as Parmenides noted more than 2,000 years ago. To be more precise, it is becoming clear that in our effort to build intelligent visual systems we must consider the fact that perception is intimately related to the physiology of the perceiver and the tasks that it performs. This viewpoint, known as Purposive, Qualitative or Animate Vision, is the natural evolution of the principles of Active Vision and in this introductory chapter some fundamental questions about vision are examined under the light of this new framework. A discussion of the ideas and research efforts that contributed to the development of the new paradigm of Active Vision, along with a short description of the rest of the book, is provided. Finally, the principles underlying purposive recognition are described and various topics including intentionality, functionality, behavior and visual categories are discussed in some detail.

1. What Is Vision?

"Rose is a rose is a rose is a rose," Gertrude Stein said [32]. In her later days she went around asking: What is the answer? Getting no reply, she then started asking:

This work was funded in part by ARPA, ONR and NSF (under a Presidential Young Investigator Award)

What is the question? Finding the question to ask is the most important problem in any intellectual endeavor. So, what is vision? Or, to be more precise, what are the questions we should ask in order to both understand the vision of living organisms and equip machines with visual capabilities? Let us first ask a few such questions [29]. When reference is made to a visual system, it could be a biological or machine system; no distinction is made.

What kind of information should a visual system derive from images? Should the information be described in some kind of language? Should this information be in a single, general purpose form, leaving it to other cognitive modules (planning, reasoning, memory, learning, language, etc.) to transform it to suit their needs, or could a visual system directly produce forms of information suited to specific cognitive processes? Are descriptions of 3-D location, shape and material properties the only descriptions that should be produced, or can a visual system directly produce "high level" information (about qualities like edibility, for example)? Are there sharply distinguished modules for vision, reasoning, learning, planning, memory, etc., or are the boundaries blurred and different subsystems closely integrated with one another?

In simpler terms, can we examine an intelligent system possessing vision and say that this part is engaged in visual processing, this other part is performing planning, this other part is doing reasoning, these cables (or connections) are for vision to communicate with planning, these cables for planning to communicate with learning, etc.? In still other words, is an intelligent visual system designed in a very clean modular manner where the different modules correspond to the different cognitive modalities, or is the design more complicated than that? Is visual perception a passive process or an active process in space-time? Are visual systems entities that just "see" and do nothing else, or are they designed in such a way that they use their vision to do something, i.e., take an action? An action is anything that changes the state of the system or the environment; it could be a motion or a decision or the building of a representation, for example.

2. What Is General Vision?

There are many researchers working in the "field of vision." There are psychologists, psychophysicists, neurobiologists, neuroanatomists, zoologists, neuroethologists, computer scientists, engineers, mathematicians, physicists and cognitive scientists. Are they all trying to answer the same question? Of course not [29]. There are those who ask *the empirical question* (what is), i.e., they are trying to find out how existing visual systems are designed; those who ask *the normative question* (what should be), i.e., they are trying to find out what classes of animals or robots would be desirable (good, best, optimal) for a set of tasks; and finally there are those who address the *theoretical question* (what could be), i.e., what range of possible mechanisms could exist in intelligent visual systems. It is obvious that these three questions, all of them very important, do not necessarily have the same answer, although they are related. Marr [23] dealt with both the first and the third question, but most of his work was on the theoretical question. He obviously mixed them up because he took the human visual system to be general. But a general vision system is a theoretical concept; it does not exist in nature and cannot be designed for many reasons, some of which are explained later.

Let us now return to the initial question. What is vision? To avoid philosophizing, let us take, as a working answer, Marr's answer: "Vision is the process that creates, given a set of images, a complete and accurate representation of the scene and its properties." We can now notice several things. This is a purposeless definition, i.e., it does not consider what is vision going to be used for. As a consequence, the extracted representation is as general as possible.[1] This is of course what researchers refer to, correctly, as general vision. However, general vision addresses the *theoretical question*, and it exists only in theory! The goal of research on the "theoretical questions" is not to create the most "general" observer, because the "general" observer exists only in theory; it is nothing but a concept. In nature, nothing is general. There is no general athlete, no general warrior, no general scientist. Research on the theoretical question will uncover general principles behind the miracle of perception. Such research will determine what is theoretically possible, or impossible, to derive using vision. It will provide the geometrical and physical constraints relating the data (images) to properties of the 3-D world. It will give the capabilities and limitations of general visual recovery techniques. For example, it will show that if models for the reflectance of surfaces or for their texture are available, then shading and texture may be used to infer properties about shape [16]. To give another example, the theoretical approach will prove that from a series of images taken by a moving camera we may be able to recover, in principle, the structure of the scene and the 3-D motion involved if we have some way of solving the correspondence problem [21]. It will also show that the correspondence problem cannot be solved, unless we employ a specific model of the scene.

Research on the "theoretical question" of general vision will not tell us how to best build a visual system. It will give hints, but it will not give a solution, because this is not what its goal is. It is research on the normative question that, using results of general vision, will propose ways of building intelligent, sophisticated and flexible visual systems. Human visual systems are just one example in the spectrum of biological vision systems; they are not general. The thousands of visual tasks that they cannot perform—while other animals can—and the many celebrated illusions to which they are subject are just trivial proofs of this simple fact. Adrian Horridge [17], a vision researcher who has been studying bees for the past thirty years or

[1] Of course Marr followed the general AI methodology of his times. The main emphasis in AI research has been the finding of general purpose methodologies and general purpose representations that preserve as much information as possible. A three step approach for solving any problem has been taken for granted. The conversion of sensor data into an internal representation and vice versa (i.e., signals to actuators, or decisions, that is all together actions) has been separated from the phase of developing algorithms to perform computations on internal data. Most research has been devoted to processing the internal data and as a consequence different subfields such as planning, learning, reasoning, vision and many more have appeared. Therefore, it is not surprising that the first influential theory of vision mainly concentrated on the computational and representational aspects. Vision was described as a reconstruction process, that is a problem of creating representations for increasingly growing levels of abstraction leading from 2-D images through the primal sketch through the $2\frac{1}{2}$-D sketch to object descriptions ("from pixels to predicates" [24]).

so writes that humans and bees are not that different in their visual systems regarding visuomotor control tasks. Of course humans have very large brains, nearly the largest one can find on earth. As a result, they have large amounts of memory and various other cognitive capabilities (for example, they can play chess, which elephants cannot). The human visual system interacts with all these cognitive abilities, hence its performance may appear spectacular, amazing and general. Spectacular it is, amazing it is, but general it is not. The simple fact that the spectacular abilities of the bee or the fly seem, to many researchers, to have little to do with higher-level human abilities reflects how little we understand the fundamental processes involved in vision.

3. What Is Active Vision?

Marr's contributions set the foundations for vision as a scientific discipline. His theories about early vision led neurobiologists to find in the monkey cortex interesting retinotopic maps. However, Marr left out of his theory a very important issue: the fact that all existing visual systems, from insects and frogs, to fish, snakes, birds and humans, are active. Being active, they control the image acquisition process and thus introduce constraints that greatly facilitate the recovery of information about the 3-D scene. "I move, therefore I see" [13] is a fundamentally true statement. And if we manage to make the human eye stationary, humans start losing perception!

The first technical developments in active vision [2] considered the problem in the context of the general methodological paradigm at the time, namely the one that treats vision as the process of *general recovery* [23]. An active observer was defined as one who is capable of engaging in some kind of activity whose purpose is to control the geometric parameters of its sensory apparatus. The superiority of an active observer vs. a passive one was clearly established by the fact that an active observer can perform classical general recovery tasks (like shape from x) in a more efficient way than a passive observer can. General recovery problems that are ill posed and nonlinear for a passive observer become well posed and linear for an active observer. However, it slowly started becoming clear that not only are the observer's geometric parameters relevant, but a whole set of other visual parameters as well. The ability to manipulate them in a controlled manner, as both an action and a reaction, started building the concept of active vision. However, the way the manipulation is done, to what extent and what kind of manipulation appeared to be task-dependent. The typical property of passive vision is that the observer is not capable of choosing how to view the scene, but is instead limited to what is offered, determined by the preset visual parameters and environmental conditions, including time sampling. The active observer, on the other hand, utilizes its capability to change its visual parameters to acquire favorable data from the scene in solving the specific task it has at the time. A passive system has to extract all the information needed from the given images, possibly engaging in complicated reasoning and computations, but cannot acquire more data which could facilitate the interpretation of the scene in view in order to achieve the task it is engaged in. Researchers started realizing that an active vision system is a system able to manipulate its visual parameters in a controlled manner in order to extract useful data about the scene in space and time that would allow it to best perform a set of tasks. Since vision is trivially active, it became apparent that the design of a vision system depends critically on the tasks it has to perform. At the same time, the fallacy of general vision started becoming clearer. An important observation was that vision does not function in isolation, but as a part of a system that interacts with the world in highly specific ways. In addition, researchers started realizing that vision systems do not have to compute all things at all times, but only what they need. These observarions led to the next natural development or evolution of active vision, namely the paradigm of purposive [1] (or animate) vision [6].

4. Active Vision Leads to Purposive Vision

An active observer that wants to reconstruct an accurate and complete representation of its extrapersonal space needs an unrealistically large amount of computational power. To give an example from human vision: one of its special features is the fovea. Humans look at the world using a small window that they move around using a very elaborate gaze control system. If the resolution of the human eye were everywhere equal to its resolution near the optical axis, then humans would have a brain weighing approximately 30,000 pounds! [30]

For reasons such as this one and several others [1] related to the inherent difficulties of complete visual reconstruction, the platonic view of vision as a general reconstruction process started fading away. A perceptual system and the world it lives in cannot be separated. They need each other in order to be complete, in order to make sense. The perceptual system has a relationship with the world it lives in; thus the system itself is a particular embodiment of that relationship. To have a "general" relationship with the world that obeys the laws of physics is something beyond any comprehension.[2]

There is simply too much that can be known about the world for a vision system to construct a general-purpose complete description. The information contained in the visual signal is much more than the system actually needs or can cope with. Thus the fundamental problem of a vision system is to determine what information from the image should be used and what representation of it needs to be built so that the relationship of the system with its world can best be implemented. In other words, the system needs to recover partial information about the scene. Knowing what kind of partial recovery needs to be performed depends on the relationship of the system to its world; to put it differently, it depends on the tasks that the system has to carry out, i.e., on its purpose. Purposive vision does not consider vision in isolation, but as part of a complex system that interacts in specific ways with the world. It is important to note that if the visual system knows what kind of information is needed and what it will be used for, this permits the system to alter its interaction with the world dynamically in order to make this information more easily available. Finally, since the relationship of the visual system to the world consists of perceptual capabilities and actions, its implementation (i.e., the

[2] One might be able to achieve such a general relationship in Plato's world of ideas.

design of the visual system) can be achieved through a reduced instruction set of perceptions and actions (behaviors) that does not require an elaborate categorical representation of the world (qualitative vision). The visual categories a visual system uses, and consequently the algorithms it needs to develop or learn in order to derive them, and the meaning of the symbols representing these categories, depend totally on two things:

(a) the characteristics of the system itself (its physiology, its mobility—is it flying, crawling, walking, etc.—and its computational capacity); and
(b) the tasks it needs to accomplish.

To summarize, purposive vision calls for partial and opportunistic visual reconstruction and for the development of new, flexible representations related to action.

5. Theoretical, Normative and Empirical Vision

If we want to use images in order to accomplish a task, we obviously need to recover something from them. However, there are many dimensions along which we can move when we perform visual recovery. The recovery can be partial or complete, the problem can be general or specific, and the assumptions used can be general or specific. The kind of research we perform depends on our position in this 3–D space. For the readers' benefit, some dimensionality reduction is performed and two 2-D spaces are presented which convey the message and improve the clarity of the pictorial description. Tables 1 and 2 describe this, if one examines the labeling of the axes. The concept of complete vs. partial recovery is clear. How "partial" partial is depends on the task (or purpose). What is a general or specific problem? A general problem is a problem of general vision, a module of Marr's theory (like optic flow estimation, shape from X, and the like)—a problem whose goal is to obtain complete recovery. A specific problem is one that answers a specific question about the scene. Usually, specific problems are simple applications of general problems. An example here will help clarify matters: The module of structure from motion (from a series of dynamic images, recover the shape and 3-D motion of the observer relative to parts of the scene) is a general problem. The following problems are specific: Is there anything moving in the scene as seen by a moving observer? Is it moving closer to the observer or is it going away? Is it going to hit the observer? How should the observe move in order to intercept it? etc. Obviously, specific problems require partial recovery of the visible world. In the sequel the difference between general and specific assumptions is explained. First of all, assumptions are constraints on space-time, the visual system and their relationship. An assumption is general when it holds universally, under any circumstances. A specific assumption is one that holds only for a subset of the world (or for a subset of visual systems and their relationships to the world).

When we assume that the world is piecewise planar, that is a general assumption, because indeed the world satisfies it (it is actually a theorem in differential geometry). However, in order to utilize this assumption for visual recovery we must make an additional assumption regarding the number of planar patches; and then

Assumptions axis →

Recovery axis ↓	General	Specific
Complete	Not possible	Marr's theory
Partial	Purposive and qualitative vision	Existing machine (industrial) vision

Table 1

Assumptions axis →

Problem axis ↓	General	Specific
General	Not possible	Marr's theory
Specific	Purposive and qualitative vision	Existing industrial vision

Table 2

our assumption has become specific. Similarly, we may assume that the world is smooth between discontinuities; that is general. Again, however, in order to utilize this assumption we must make some assumption regarding the discontinuities, and then our assumption becomes specific. We may assume that an observer only translates. If indeed the physiology of the observer is such that it only translates, then we have made a general assumption. 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. If we assume that the noise in our system is Gaussian or uniform, again we have made a specific assumption.

Tables 1 and 2 show how general vision, purposive vision and industrial vision fit together in a continuum of completeness and partialness, generality and specificity. Purposive vision and general vision work together, with general vision providing hints and ideas and purposive vision designing the whole spectrum of visual systems. We can always formulate general vision problems as optimization problems. But global optimization is, in general, impossible and it requires specific assumptions in order to determine the coefficients involved in the global functionals (or to determine the priors in various general probabilistic approaches). Through such studies, however, we can learn aspects of the problem space that make it difficult and formulate ideas for robust partial recovery, under general assumptions. The fact that we can solve partial recovery problems under general assumptions should not be surprising; it is because this requires much less of the visual computation.

6. The Big Picture

Purposive vision is the vision of any visual system, including the human. Researchers are uncomfortable when they do not immediately see the big picture, how different cognitive mechanisms fit together, i.e., the brain. How can we view the brain as an intelligent, flexible, purposive perceptual system?

This problem is easier if we study vision in isolation [23]. The dominant view regarding the brain has been that it is modular. Each cognitive process is assigned a module; each of these "big" modules is broken down into submodules, and so on. It is becoming increasingly clear that this is not true. It is, of course, important to view the brain in a modular way. Modularity is a general principle that we follow when we design or analyze large, complex information systems. However, purposive vision (or purposive perception) suggests that these modules are behaviors!

A behavior is a sequence of cognitive events and actions, a set of visual, planning, memory, and reasoning processes working in a cooperative manner and "acting" on the system itself or its environment.

Is there a formal theory of behaviors? How do we describe them in a formal manner? How do we learn new ones and how do we build complex behaviors using a synthesis of existing, simpler ones? Is there a behavioral calculus? Several chapters in this book shed some light on these issues, but these problems are very deep and complex.

Regarding an architecture supporting the integration of behaviors, it is clear that Brooks' [7] layered architecture is not satisfactory because low- and high-level processes should cooperate directly and because the order of visual systems is not complete, but partial. This simply means that flexible and sophisticated vision systems may not have capabilities that simpler (apparently) systems possess. In addition, flexible and sophisticated systems need to be able to represent knowledge, and Brooks' viewpoint that the world can be used as a repository of information can give rise, at best, to some simple reactive behaviors. On the other hand, formalisms such as discrete event dynamic systems, I/O automata, Petri nets, Bayesian networks, and the like [18, 28, 20, 22, 12, 26], appropriately modified, are good candidates for formalizing behavior[3], i.e., for serving as mechanisms for knowledge representation inside a behavioral process. For lack of a better word and slightly modifying Sloman's [29] word, this architecture has been called labyrinthic [1], in order to stress

[3] There is a large amount of literature in theoretical computer science that basically models behaviors.

the interaction among the different cognitive modalities and action. The brain is thus a set of behaviors, and it makes a lot of sense to study it as such.

7. Building Visual Systems

The main reason why the approach to vision suggested by Marr has not led to the development of successful artificial systems is that vision was studied in a vacuum, i.e., the utilization of vision was completely ignored. However, the vision of no system is purposeless and thus it necessarily needs to be studied in conjuction with the task the system is involved in. From this viewpoint, understanding vision means understanding a system possessing visual capabilities.

In general, if our goal is to study (or more precisely formulated, analyze in order to design) a system, we are advised by engineering considerations to follow some common principles in system design and address a set of basic questions: What is the functionality of the system? What are the autonomous subsystems (modules) the system is divided in? What is the relationship of the modules to each other? What is the representation of information within the subsystems, and how are the modules communicating with each other? Finally, we have to ask: What is the most efficient and effective way for designing the single modules?

Having these questions in mind, C. Fermüller [11] proposed a new approach for studying vision and developing artificial vision systems. This approach takes it for granted that the observer (the system) possesses an active visual apparatus. Since, furthermore, it is inspired by evolutionary, neuroethological considerations, it is called synthetic (evolutionary) approach. This approach constitutes a philosophy for how to systematically study visual systems which live in environments as complex and multifarious as those of human beings. It is substantiated by basically two principles. The first principle is related to the overall structure of the system and how it is modularized, what are the problems to be solved and in which order they ought to be addressed. The second principle is concerned with the way the single modules should be realized.

As basis for its necessary computations a system has to utilize mathematical models, which serve as abstractions of the representations employed. The first principle of the synthetic approach states that the study of visual systems should be performed in an hierarchical order according to 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. The synthetic approach calls for first studying capabilities whose development relies on only simple models and then go on and study capabilities requiring more complex models. Simple models do not refer to environment or situation specific models which are of use in only a limited amount of situations. On the contrary, 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. In other words, the employed assumptions have to be general with regard to the environment. The motivation for this approach is to gain increasingly insight in the process of vision, which is of such high complexity. Therefore the capabilities that require more complex models should be based on "simpler" already developed capabilities. The complexity of a capability is thus given by the complexity of its assumption; what has been considered a simple capability, might require complex models and vice versa. For example, the celebrated capability of egomotion estimation does not require as has been believed complex models about the geometry of the scene in view or the time evolution of the motion, but only a simple rigid motion model, while the detection of independent motion, which has been considered as primitive, requires more elaborate non-rigid motion models.

The second principle, motivated by the need for robustness, is the quest for algorithms which are qualitative in nature [1]. The synthetic approach does not have as its goal the reconstruction of the scene in view, but the development of a class of capabilities that recognize aspects of objective reality which are necessary to perform a set of tasks. The function of every module in the system constitutes an act of recognizing specific situations by means of primitives which are applicable in general environments. For example, a system, in order to avoid obstacles, does not have to reconstruct the depth of the scene in view. It merely has to recognize that the distance to a close-by object is decreasing at a rate beyond some threshold given by the system's reaction time. Recognition, of course, is much easier than general reconstruction of the scene, simply because the information necessary to perform a specific task can be represented in a space having only a few degrees of freedom. Moreover, in order to speak of an algorithm as qualitative, the primitives to be computed don't have to rely on explicit quantitative models. Qualitativeness can be achieved due to a number of reasons: 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 can be proved stable and robust in all possible cases.

8. Scanning the Book

The seven chapters of this book are devoted to various aspects of active perception, ranging from general principles and methodological matters to technical issues related to navigation, manipulation, recognition, learning, planning, reasoning and topics related to the neurophysiology of intelligent systems. In the first chapter, Pahlavan et al. address active vision as a methodology and elucidate its methodological superiority over passive vision; they also discuss issues of system design, purpose dependency and problems related to control. The second chapter treats the problem of navigation from the perspective of active, purposive vision and advances a methodology borrowed from the field of programming languages for formalizing the behaviors of an agent in such a way that we can reason about them (i.e., prove their properties). In Chapter 3, Fermüller discusses the problem of 3-D motion for an active observer and shows that the problems of estimating egomotion and the motion of an object, although mathematically equivalent, are perceptually different and should be addressed through the development of different behaviors. She develops novel geometric constraints that serve as the basis of a set of algorithms for 3-D motion estimation that do not rely on correspondence or optic flow computation as a preprocessing step. The fourth chapter demonstrates that the simultaneous and coordinated operation of vision and action can be used not only to simplify some traditional visual tasks, but also to extend the overall scope of vision to important new areas. The concepts of purposiveness, closed-loop control and concurrency of motion and sensing are dominant throughout the chapter and they synergistically contribute to novel solutions in problems of navigation, manipulation and gaze control. In Chapter 5, Raviv and Herman treat the problem of visual servoing for an active and purposive vision system in a novel manner. They present a solution to the problem of visually controlling a vehicle which is very different from the traditional approaches to navigation. In Chapter 6, A. Blake presents the beginnings of a theory for the computational modelling of hand–eye coordination, a topic of increasing scientific interest. Finally, Ballard and Brown present a set of ideas around cooperative gaze-control behaviors that reduces the need for explicit representation postulated in the perceiver. These ideas take the form of principles governing active vision systems and they represent a framework for sequential decision making and visual learning. There exists a good reason why most of the chapters are devoted to problems related to navigation and manipulation. This is because the most basic visual capabilities found in living systems are based on motion [17]. In addition, it is not very hard to classify categories in the visual environment in a way which is related in a purposive manner to visuomotor coordination tasks of an agent. Later, a high-level analysis of purposive recognition is given.

9. Final Thoughts

Marr was influenced by Warrington [31]. She described the capacities and limitations of patients who had suffered left or right parietal lesions. She noticed the existence of two classes of patients: The members of the first class could recognize an object provided that its view was conventional; the members of the second class could not name the object or its purpose but could still perceive its shape. To Marr these results suggested that vision alone could deliver a description of the shape of an object even when the object was not recognized in a conventional sense; Marr took this to be the main (central) purpose of vision. After many years of research and after looking at many patients, Martha Farah does not seem to be finding the same results. Her findings agree much more with a labyrinthic picture of the brain than with a Marrian modular point of view (see Appendix 1 for a very short summary of Farah's work).

It is very hard to understand a purposive approach to recognition because this problem is equivalent to the embodiment of categories. To shed some light on this complex issue, a few more appendices have been added: on intentionality and behavior (Appendix 2); on the recognition process (Appendix 3); and on functional categories (Appendix 4).

What does the future hold? Researchers will realize that general vision is a chimera. Although general vision will continue to be studied, it will become clear that it does not make much sense to insist on developing systems possessing general vision. They will develop basic visual capabilities that, in the framework of purposive, qualitative and active vision, will contribute in a synthetic manner to the development of flexible vision systems. We will see, not general, but specific invariance techniques appearing; we will find out how to achieve the most while spending the least effort. We will discover the computational capabilities of uncalibrated (or partially calibrated) vision systems. Most important, flexible vision systems will be constructed in several laboratories (see [3] for some existing ones); we will observe them and experiment with them. In the past, many discoveries have resulted from unexplained engineering observations rather than from the development of a successful theory.

Vision has a purpose, and that purpose is action. Action can be practical (motor control), theoretical (creation of a purposive representation, a decision or an internal change of state) or aesthetic [29]. Flexible and sophisticated vision without action is meaningless. Purposive vision bridges the gap between theory (general recovery) and vision systems. In this framework, we can advance the field to a post-paradigmatic stage, in the sense of Kuhn. In simple terms, we can integrate the different cognitive modalities - perception, planning, reasoning, learning - into intelligent beings. We should also note that learning can be done much more successfully in the purposive paradigm, because we learn well defined things—behaviors, instead of general-purpose representations whose beauty is only of theoretical importance. Working on general vision has discouraged the integration of learning and visual processes, something that has just started [4, 25]. In addition, the problem of photointerpretation—analysis of static images—a problem receiving increasing attention lately under the name "Image Databases" and which humans are very good at–is trivial when the researcher is allowed to collect the perceptual images and to ask the appropriate questions, and is currently impossible when general questions are asked.

References

1. Y. Aloimonos, "Purposive and qualitative active vision," In *Proc. Image Understanding Workshop*, 1990, 816–828.

2. Y. Aloimonos, I. Weiss and A. Bandopadhay, "Active vision," *Int'l. J. Comp. Vision* **7**, 1988, 333–356.

3. Y. Aloimonos (Ed.), Special Issue on Purposive and Qualitative Active Vision, *CVGIP B: Image Understanding* **56**, 1992.

4. Y. Aloimonos, R. Michalski, P. Pachowitz and A. Rosenfeld, "Report on the NSF/DARPA Workshop on Vision and Learning," October 1992, Harpers Ferry, VA.

5. R. Bajcsy, "Active perception," *Proc. IEEE 76* **8**, 1988, 996–1005.

6. D. Ballard, "Animate vision," *Artificial Intelligence* **48**, 1991, 57–86.

7. R.A. Brooks, "Achieving Artificial Intelligence Through Building Robots," A.I. Memo 899, MIT, Cambridge, MA.

8. B. Chandrasekaran, "Design problem solving: A task analysis," *Artificial Intelligence* **11**, 1990, 59–72.

9. M. Dennis, "Dissociated naming and locating of body parts after left anterior temporal lobe resection: An experimental case study," *Brain and Language* **3**, 1976, 147–163.

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

11. C. Fermüller, *Basic Visual Capabilities*, Ph.D. Thesis, Technical University of Vienna, 1993. (also, Technical Report, Computer Vision Laboratory, University of Maryland, 1993).

12. S. Owicki and D. Gries, "An axiomatic proof technique for parallel programs," *Acta Informatica* **6**, 319-440.

13. T. Hamada, "Active vision," In *Proc. Int'l. Neuroethology Congress*, 1992.

14. J. Hart, R.S. Berndt and A. Caramazza, "Category-specific naming deficit following cerebral infraction," *Nature* **316**, 1985, 439–440.

15. J. Hopcroft, J. Schwartz and M. Sharir, "On the complexity of motion planning for multiple independent objects: PSPACE hardness of the warehouseman's problem," *Int'l. J. Robotics Research* **3**, 1984, 76–88.

16. B.K.P. Horn, *Robot Vision*, McGraw Hill, New York, 1986.

17. G.A. Horridge, "The evolution of visual processing and the construction of seeing systems," *Proc. Royal Soc. London B* **230**, 1987, 279–292.

18. Y. Ho, "Dynamics of discrete event systems," *Proc. IEEE* **77**, 1989, 3–6.

19. J. De Kleer and S. Brown, "A qualitative physics based on confluences," in D.G. Bobrow (Ed.), *Qualitative Reasoning about Physical Systems*, MIT Press, Cambridge, MA, 1985.

20. F. Lin and W.M. Wonham, "Decentralized control and coordination of discrete event systems," In *Proc. IEEE Conf. on Decision and Control*, 1988, 1125–1130.

21. H.C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature* **293**, 1981, 133–135.

22. N. Lynch and M. Tuttle, "Hierarchical correctness proofs for distributed algorithms," In *Proc. ACM Symposium on Principles of Distributed Computing*, 1987.

23. D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, Freeman, San Francisco, 1982.

24. A. Pentland (Ed.), *From Pixels to Predicates: Recent Advances in Computational and Robot Vision*, Norwood, NJ; Ablex, 1986.

25. T. Poggio and S. Edelman, "A network that learns to recognize 3-D objects,", *Nature*, **343**, 1990, 263–266.

26. P.J. Ramadge and W.M. Wonham, "The control of discrete event systems," *Proc. IEEE* **77**, 81–98.

27. J.R. Searle, *Intentionality*, Cambridge University Press, New York, 1983.

28. S.S. Lam and A.U. Shankar, "A relational notation for state transition systems," *IEEE Trans. Software Engineering*, 1991, 12–25.

29. A. Sloman, "On designing a visual system," *J. Experimental Theoretical Artificial Intelligence* **1**, 1989, 289–337.

30. E. Schwartz, Personal Communication, *Workshop on Active Vision*, University of Chicago, 1991.

31. E.K. Warrington and T. Shallice, "Category specific semantic impairments," *Brain* **107**, 1984, 829–854.

32. R. Wehner, "Matched filters: Neural models of the external world," *J. Comparative Physiology A* **161**, 1987, 511–531.

33. G. Wilfong, "Motion planning in the presence of movable obstacles," In *Proc. ACM Symposium on Computational Geometry*, 1988, 279–288.

34. A. Yamadori and M.L. Albert, "Word category aphasia," *Cortex* **9**, 1973, 112–115.

Appendix 1: Human visual agnosia and object recognition

The condition of visual agnosia provides interesting evidence as to how object recognition might be done by humans. Visual agnosia refers to a condition in which the patient fails to recognize objects (by sight) due to some brain damage, despite relatively well preserved sensory capacities. A common distinction is between apperceptive agnosia, in which recognition fails because of an impairment in visual perception (patients do not see objects normally, and hence cannot recognize them), and associative agnosia, in which perception seems adequate to allow recognition, and yet recognition cannot take place.

Farah [10] summarizes associative agnosia cases. Especially revealing is the categorization of objects agnostic patients fail to recognize. Farah describes patients studied by Warrington [31] in which knowledge of stimuli roughly corresponding to living things is relatively impaired compared to knowledge of most nonliving things, or vice versa. Under the definition of living things she included animals, plants and foods. Nonliving things were exemplified by small, man-made objects. Farah adds cases like Dennis' [9], in which deficits were confined to body parts; Yamadori's [34], objects typically found indoors; Hart's [14], fruits and vegetables; and prosopagnosia, which is the inability to recognize faces.

Under apperceptive agnosia, Farah describes a case of a patient who had adequate elementary visual functions and general cognitive ability, yet was dramatically impaired in the simplest forms of shape discrimination. This patient could detect differences in luminance, color, size, as well as respond to small movements of presented objects, but could not distinguish between two objects of the same size, color, etc. when the only difference between them was shape. It is important, in our opinion, to emphasize that this patient could identify and name objects from tactile, olfactory, or auditory cues without any problem. This situation is reported for other patients as well. It is interesting to note the role played in the recognition process, as well as the help given by motion. Tracing was used for contour following, orientation judgement, or search for the best view.

Another interesting disorder is dorsal simultanagnosia. Accepting Luria's definition [10], it is a specific type of perceptual deficit in which only one object, or part of an object, can be seen at one time. Luria suggests that "objecthood" is not simply a matter of size or of visual complexity. For example, a face can be viewed as an object, a collection of other objects such as eyes, nose, etc., or as a part of a larger object, the human body. Luria's patient could perceive a face, but when a detail was perceived, its relation to the whole was lost. It seems that whatever cognitive process is disrupted in dorsal simultanagnosia is sensitive to shifts in what the visual system takes to be an object.

Appendix 2: Intentionality and behavior

Vision systems that operate in different environments and perform different visual tasks do not necessarily recognize objects using similar algorithms. A vision system that recognizes ten types of objects does not necessarily work in the same way as a system that needs to recognize one type or a hundred types. A system that serves a rapidly moving agent is not necessarily built in the same way as a system for a stationary agent. Object recognition should be studied by taking into account not only the objects that have to be recognized but also the agent that has to perform the recognition. Since different agents, working with different purposes in different environments, do not recognize visually in the same manner, we should not seek a general, universal theory of object recognition. Instead, we should concentrate on developing a methodology that, given an agent will suggest how to perform particular recognition tasks.

An agent is a system (robot) that has visual (and other) sensing capabilities and is able to carry out a set of behaviors. These behaviors are direct results of a set of purposes or intentions that the agent has. A behavior is identified as anything that changes the internal state of the agent and its relationship to the environment. Carrying out a behavior calls for the performance of various recognition tasks. By performing partial recovery of attributes of an object, we can find out if the object is suitable for the desired purpose. In general, an object can be used for many purposes. The agent must recognize the one needed to carry out its behavior.

Perception is a causal and intentional transaction between the mind and the world. The intentional content of our visual perception is termed [27] "the visual experience." When we see a table there are two elements in the perceptual situation: the visual experience and the table. The two are not independent. The visual experience has the presence and features of the table as conditions of satisfaction. The content of the visual experience is self-referential in the sense that it requires that the state of affairs in the world must cause the visual experience which is the realization of the intentional content.

When we visually perceive an object we have a visual experience. This visual experience is an experience of the object. It may be that the conditions of satisfaction are not fulfilled. This is the case for illusions, hallucinations, etc. The visual experience, and not the world, is at fault. The visual experience that we have, in this case, is indistinguishable from the visual experience we would have if we actually saw the real object. The intentional content of the visual experience determines its conditions of satisfaction. A visual experience in that sense is a mental phenomenon which is intrinsically intentional.

An agent is defined as a set of intentions, $I_1, I_2, \ldots, I_n$. Each intention I_k is translated into a set of behaviors, $B_{k1}, B_{k2}, \ldots, B_{km}$. Each behavior B_{ki} calls for the completion of recognition tasks $T_{ki1}, T_{ki2}, \ldots, T_{kij}$. The agent acts in behavior B_{ki} under intention I_k. The behavior calls for the completion of recognition tasks $T_{ki1}, \ldots, T_{kin}$. The behavior sets parameters for the recognition tasks. Note that the same object can answer positively to several recognition tasks. Under one behavior a chair will answer yes to some recognition task that is asking for obstacles, under another behavior it will answer yes to a recognition task that is asking for a sitting place, and under another it will answer yes to a task that is asking for an assault weapon.

We can view the recognition process along the axis intention, behavior, recognition task. For a theory of purposive object recognition we should be able to make two basic transformations: first from the desired intention to the set of behaviors that achieve it, second from a specific behavior to some needed recognition task(s). In the following paragraphs we will show that the intention-to-behaviors (or task decomposition) problem with a finite number of behaviors is undecidable by reducing it the halting problem. We believe that the transformation from behaviors to recognition tasks is also hard.

Let the state of the world S_n be defined by the two tuples $\langle O_n, R_n \rangle$, where O_n is a set of objects $(O_{n1}, O_{n2}, \ldots, O_{nk})$ and R_n is a set of relations $(R_{n1}, R_{n2}, \ldots, R_{nl})$ between the different objects. A behavior is a transition between two states. An intention is a desired state of the world. The intention-to-behaviors problem is: Given the triplet $\langle S_0, S_n, (B_1, \ldots, B_k) \rangle$ (i.e., a start state, an intention, a set of behaviors), find a sequence of behaviors that leads from the start state to the desired intention.

Assume we are given a Turing machine M and an input sequence of symbols X. We describe the transformation from $\langle M, X \rangle$ to $\langle S_0, S_n, (B_1, \ldots, B_k) \rangle$ informally. To encode a completely blank tape containing the input X, with the indication that M's head is pointing to the leftmost symbol of X and that the state is M's start state, we write the start state as $\langle (\#, X, \#), (R_1, R_2) \rangle$ where R_1 indicates M's state and R_2 the head position. The state-transition diagram is represented by the set of behaviors. The halting state is represented by S_n (the intention). Activating a sequence of behaviors from the start state corresponds directly to a computation of M on X. Consequently, if the intention-to-behaviors problem were decidable the halting problem would be decidable too.

If we add constraints to our definition of the problem we can move from undecidability to intractability. For example, by constraining ourselves to a constant set of objects we can show a PSPACE-hard lower bound. This can be shown by reducing our problem, for example, to that of motion planning for an object in the presence of movable obstacles, where the final positions of the obstacles are specified as part of the goal of the motion. The complexity of this problem is discussed in [33, 15]. The reduction is straightforward. The set of objects contains the moving objects and the obstacles (the polygonality can be given in the relation set or as part of the objects' definition). The positions of the objects are part of the relation set. The intention encodes the final state. Grasping, pushing and moving are the behaviors. Solving the intention-to-behaviors problem gives a solution to this problem.

Appendix 3: The recognition process

Object utilization is not the same problem as that of naming an object. Under this framework an agent acts in behavior B_{ki} under intention I_k. The behavior calls for the completion of recognition tasks $T_{ki1}, \ldots, T_{kin}$. The behavior sets parameters for the recognition tasks. Each recognition task activates a different collection of basic perceptual modules. Each module qualitatively finds a generic object property which is a result of one or a combination of direct low-level computations on some sensory data (possibly done by other modules). The result of a module's operation is given as a qualitative value. Each module has its own neighboring open intervals which are parameter-specific. The i^{th} module can take one of $q_{i1}, \ldots, q_{in}$ qualitative values.

The state of our recognition system, denoted by Q_i, is a tuple of all the qualitative values of our modules $(q_1, \ldots, q_m)$ under recognition task T_{kij}. Each recognition task T_{kij} defines a system state that will constitute a positive answer to that recognition task. Recognition is done when we complete our task, which means a stable answer from our modules. At this point we want to remark that a common recognition task can be defined as a new module and unexpected object recognition could be developed along these lines. The conditions for this kind of decision are not considered here and probably should take into account some utility measures (frequency of appearance, network complexity, etc.).

Under this framework learning can be defined as the process of matching the "correct" system state with the recognition task needed by a certain behavior. This process is actually the reverse of recognition. A behavior creates a need for an object. An object is segmented by some low-level modules, and a system state is achieved. The object is tested and a satisfied result for a needed behavior starts the creation or definition of a recognition task.

When we need to perform a given recognition task T_{kij} under behavior B_{ki} and intention I_k, we may assume that some parameter setting is done by the intention and the behavior. These parameters fix the setting for the task, which includes the required system state (some of the modules might be in the don't care position) and possibly some additional "common knowledge" parameters, such as environmental parameters (outdoor, indoor), predator, size, etc. From this point of view the recognition process is using high-level information.

Appendix 4: Functional categories

The relationships of an organism or robot to objects in its environment are functional relations: Objects can be obstacles, predators, prey, etc. Such relations are intention-dependent. Objects are related to actions from the utilitarian point of view. When we want to get a coconut from a tree, we must search for an object that provides the necessary functions—e.g., graspable, movable, rigid, elongated, long. We thus need to define transformations from these functions into the needed sensory data. Objects have observable characteristics from which we can infer which functional category an object belongs to. For example, does the object appear to be immobile or mobile? (It can be momentarily stationary.) Is the object graspable? Does the object appear to be organic or inorganic (animal, vegetable, mineral)?

This transformation is close, in some senses, to a design process. Design involves mapping from behavior to structure. A designer tries to specify an artifact that delivers some functions and satisfies some constraints [8]. It is interesting to note that general algorithms for design are computationally intractable [8], and that common methods for design use compiled knowledge, solved design cases, and the like. We believe that the general solution to this problem of finding the transformation from function to structure is hard.[4] In order to build working systems, we can use a set of common, useful, specialized functional translations.

We can define object categories like animate, inanimate, prey, predator, obstacle, etc., that belong to hierarchical structures. The hierarchies are functional and have sensory relevance, i.e., they must have perceivable characteristics that make them discriminable.

For example, consider a class of functional categories that relate to how your motion is constrained by the things in the category. Obviously the definitions of these classes depend on your motion ability: Are you a tortoise or a vehicle with wheels? Depending on the nature of your mobility and on your size, the world subdivides into different functional classes of objects in terms of their possible roles as obstacles. We can have movable obstacles, obstacles that can be climbed over, different surface classes that can be described in terms of how they impede motion (depending on their orientation), etc. We can have different taxonomies for agents that move on the ground (or the surfaces) and for agents that fly. This is only one taxonomy with which a large number of objects in the environment can be labeled with respect to how they can affect the agent's mobility. Other functional taxonomies can be built based on concepts such as prey, food, etc.

[4] We can see another aspect of the complexity by looking at the inverse problem of deriving function from structure. The general translation is not one to one (the same is true for our direction). In order to solve the problem there is need for many limiting assumptions [19]. With these assumptions the present solution relies on a library of generic components with their allowable paths of interaction and a set of boundary conditions which constrain the device's behavior. Even under these constraints the system (ENVISION [19]) may produce a set of behaviors, one of which corresponds to the actual behavior of the real device.