Active Perception

RUZENA BAJCSY, MEMBER, IEEE

Invited Paper

Active Perception (Active Vision specifically) is defined as a study of Modeling and Control strategies for perception. By modeling we mean models of sensors, processing modules and their interaction. We distinguish local models from global models by their extent of application in space and time. The local models represent procedures and parameters such as optical distortions of the lens, focal lens, spatial resolution, band-pass filter, etc. The global models on the other hand characterize the overall performance and make predictions on how the individual modules interact. The control strategies are formulated as a search of such sequence of steps that would minimize a loss function while one is seeking the most information. Examples are shown as the existence proof of the proposed theory on obtaining range from focus and stero/vergence on 2-D segmentation of an image and 3-D shape parametrization.

I. Introduction

Most past and present work in machine perception has involved extensive static analysis of passively sampled data. However, it should be axiomatic that perception is not passive, but active. Perceptual activity is exploratory, probing, searching; percepts do not simply fall onto sensors as rain falls onto ground. We do not just see, we look. And in the course, our pupils adjust to the level of illumination, our eyes bring the world into sharp focus, our eyes converge or diverge, we move our heads or change our position to get a better view of something, and sometimes we even put on spectacles. This adaptiveness is crucial for survival in an uncertain and generally unfriendly world, as millenia of experiments with different perceptual organizations have clearly demonstrated. Yet no adequate account or theory or example of active perception has been presented by machine perception research. This lack is the motivation for this paper.

Manuscript received November 23, 1987; revised March 21, 1988. This work was supported in part by NSF Grant DCR-8410771, Air Force Grant AFOSR F49620-85-K-0018, Army/DAAG-29-84-K-0061, NSF-CER/DCR82-19196 AO2, DARPA/ONR NIH Grant NS-10939-11 as part of Cerebo Vascular Research Center, NIH 1-RO1-NS-23636-01, NSF INT85-14199, NSF DMC85-17315, ARPA N0014-85-K-0807, NATO Grant 0224/85, and by DEC Corporation, IBM Corporation, and LORD Corporation.

The author is with the Computer and Information Science Department, University of Pennsylvania, Philadelphia, PA 19104, USA.

IEEE Log Number 8822793.

II. What is Active Sensing?

In the robotics and computer vision literature, the term "active sensor" generally refers to a sensor that transmits (generally electromagnetic radiation, e.g., radar, sonar, ultrasound, microwaves and collimated light) into the environment and receives and measures the reflected signals. We believe that the use of active sensors is not a necessary condition on active sensing, and that sensing can be performed with passive sensors (that only receive, and do not emit, information), employed actively. Here we use the term active not to denote a time-of-flight sensor, but to denote a passive sensor employed in an active fashion, purposefully changing the sensor's state parameters according to sensing strategies.

Hence the problem of Active Sensing can be stated as a problem of controlling strategies applied to the data acquisition process which will depend on the current state of the data interpretation and the goal or the task of the process. The question may be asked, "Is Active Sensing only an application of Control Theory?" Our answer is: "No, at least not in its simple version." Here is why:

1) The feedback is performed not only on sensory data but on complex processed sensory data, i.e., various extracted features, including relational features.
2) The feedback is dependent on a *priori* knowledge—models that are a mixture of numeric/parametric and symbolic information.

But one can say that Active Sensing is an application of intelligent control theory which includes reasoning, decision making, and control. This approach has been eloquently stated by Tenenbaum [1]: "Because of the inherent limitation of a single image, the acquisition of information should be treated as an integral part of the perceptual process . . . Accommodation attacks the fundamental limitation of image inadequacy rather than the secondary problems caused by it." Although he uses the term *accommodation* rather than *active sensing* the message is the same.

The implications of the active sensing approach are the following:

1) The necessity of models of sensors. This is to say, first, the model of the physics of sensors as well as the noise of the sensors. Second, the model of the signal processing and

0018-9219/88/0800-0996$01.00 © 1988 IEEE

data reduction mechanisms that are applied on the measured data. These processes produce parameters with a definite range of expected values plus some measure of uncertainties. These models shall be called Local Models.

2) The system (which mirrors the theory) is modular as dictated by good computer science practices and interactive, that is, it acquires data as needed. In order to be able to make predictions on the whole outcome, we need, in addition to models of each module (as described in 1) above), models for the whole process, including feedback. We shall refer to these as Global Models.

3) Explicit specification of the initial and final state/goal.

If the Active Vision theory is a theory, what is its predictive power? There are two components to our theory, each with certain predictions:

1) Local models. At each processing level, local models are characterized by certain internal parameters. Examples of local models can be: region growing algorithm with internal parameters, the local similarity and size of the local neighborhood. Another example is an edge detection algorithm with parameter of the width of the bandpass filter in which one is detecting the edge effect. These parameters predict a) the definite range of plausible values, and b) the noise and uncertainty which will determine the expected resolution, sensitivity/robustness of the output results from each module. Following the edge detection example, from the width of the bandpass filter, we can predict how close two objects can be before their boundaries will merge, i.e., what is the minimum separation distance. The same parameter will predict what details of the contour are detectable, and so on.

2) Global models characterize the overall performance and make predictions on how the individual modules will interact which in turn will determine how intermediate results are combined. The global models also embody the global/external parameters, the initial and final/global state of the system. The basic assumption of the Active Vision approach is the inclusion of feedback into the system and gathering data as needed. The global model represents all the explicit feedback connection, parameters, and the optimization criteria which guides the process.

Finally, all the predictions have been or will be tested experimentally. Although all our analysis applies to modalities of touch, we shall limit ourselves only to visual sensing, hence the name Active Vision (AV).

III. Previous Work

To our knowledge, very little work has been performed on Active Vision in the sense of the above definition. To that extent the state of the art in this domain is poorly defined. In this section, we will briefly describe the systems most pertinent to our design.

The hand-eye system at Stanford used a pan-tilt head, a lens turret for controlling focal length, color, and neutral filters mounted on a "filter wheel," and a vidicon camera whose sensitivity is programmable. This system is limited in processing and I/O speed (it used a PDP-6 computer) and in resolution (four bits), but its design is well conceived. It was successfully used for adaptive edge following [2].

POPEYE is a grey level vision system developed at Carnegie Mellon University [3]. It is a loosely coupled multiprocessor system on a MULTIBUS, with a MC68000, a Matrix frame grabber and buffer, as an array processor, dedicated image processing units, and a programmable transform processor. Image positioning is achieved with a pan/tilt head and motorized zoom/focus lens. Although, this is a powerful and flexible system. Weiss [4] describes a model reference adaptive control feedback system that uses image features (areas, centroids) as feedback control signals.

Kuno *et al.* [5] report a stereo camera system whose interocular distance, yaw, and tilt are computer controlled. The cameras are mounted on a specially designed linkage. The distance between the two cameras is controlled: the larger the distance, the more precisely disparity information from stereo can be converted to absolute distance; the smaller the distance, the easier is the solution to the correspondence problem. From the article, it is not clear how this flexibility will be exploited, nor how the 3 degrees of freedom will be controlled.

Recently the Rochester group [6] has reported a mobile stereo camera system as well as Poggio at MIT [7]. Both these groups are interested in modeling the biological systems rather than the machine perception problems.

One approach to formulating sensing strategies is described by Bajcsy and Allen [8]. But in this paper, control is predicted on feedback from different stages of object recognition. Such feedback will not be available from the processes determining spatial layout, which do not perform recognition.

Recently a very relevant paper to our subject matter appeared at the First International Conference on Computer Vision by Aloimonos and Badyopadhyay [9] with the title, "Active Vision." They argue that: "problems that are ill-posed, nonlinear or unstable for a passive observer become well-posed, linear or stable for an active observer." They investigated five typical computer vision problems: shape from shading, shape from contour, shape from texture, structure from motion and optic flow (area based). The principal assumption that they make is that the active observer moves with a known motion, and of course, has available more than one view, i.e., more data. Hence, it is not surprising that with more measurements taken in a controlled fashion, ill-posed problems get converted into well-posed problems.

We differ, however, with Aloimonos and Bandyopadhyay in the emphasis of the Active Vision (or perception in general) as a scientific paradigm. For us the emphasis is in the study of *modeling and control strategies* for perception, i.e., modeling of the sensors, the objects, the environment, and the interaction between them for a given purpose, which can be manipulation, mobility, and recognition.

IV. Active Vision

As we stated earlier, Active Vision for us is modeling and a study of control strategies. It is very difficult to talk about the theory of Active Vision without describing the system which embodies this theory. So our strategy of exposing the theory is to layout the system in such a way that indeed predictions can be made and verified either formally or experimentally. Some of the predictions can be made specific only when one describes a concrete hardware/software setup. This is why we shall use some examples to support our arguments.

In Section II we have categorized models into local and global. The distinction is based on extent of the process in time and space. By and large the physical and geometric properties are encoded in local models, while the interaction among the local models is controlled by the global models. We shall discuss different models as we proceed in our presentation of different systems. Here we just mention a few related works on this subject. Models and estimation theory have been successfully applied by Zucker [10]. The problem that Zucker addressed is how to find curves from visual data. He decomposed the process into three steps:

1) Measurement step, implemented via series of convolutions.
2) The interpretation step. This has been realized via functional minimization applied on the results from convolution.
3) Finally the integration process in order to get the curve.

This decomposition in steps, with the parameters at each step explicit, allows Zucker to make clear predictions about where contours will be found and what the limitations are.

The very same flavor of models and predictions can be found in the papers of Leclerc and Zucker [11] and Nalwa and Binford [12] which are applied to edge detection and image discontinuities. Signal models have been investigated by several workers, examples are Haralick [13] and recently Pavlidis [14].

Now we shall discuss the problem of control. There are three distinct control stages proceeding in sequence:

- initialization,
- processing in midterm,
- completion of the task.

Strategies are divided with respect to the tradeoff between how much data measurement the system acquires (data driven, bottom-up) and how much *a priori* or acquired knowledge the system uses at a given stage (knowledge driven, top-down). Of course, there is that strategy which combines the two.

To eliminate possible ambiguities with the terms bottom-up and top-down, we define them here. *Bottom-up (data driven), in this discussion, is defined as a control strategy where no concrete semantic, context dependent model is available, as opposed to the top-down strategy where such knowledge is available.* While semantic, context dependent models are *not* relevant to bottom-up analysis, other types of models serve an important function. For example, bottom-up analysis in active vision requires sensor models.

Table 1 below outlines the multilayered system of an Active Vision system, with the final goal of 3-D object/shape recognition. The layers are enumerated from 0, 1, 2, $\cdots$ with respect to the goal (intermediate results) and feedback parameters. Note that the first three levels correspond to monocular processing only. Naturally the menu of extracted features from monocular images is far from exhaustive. The other 3–5 levels are based on binocular images. It is only the last level that is concerned with semantic interpretation. There is no clear delineation among the levels 0–5 with respect to the device control and software control. This of course is not surprising.

Table 1

Level	Feedback Parameters	Goal/Stopping Condition
0 i.e., control of the physical device	directly measured: current in the lighting system position of the camera open/close aperature simple computation: Gross focus	grossly focused scene camera optimally adjusted aperature
1 i.e., control of the physical device	directly measured: focus, zoom computed: contrast	focused one subject distance from focus
2 i.e., control of low level vision modules	computed only: threshold of magnitude of edge width of a filter similarity criterion on region growing	2-D segmentation, i.e., maximum number of edges max. and min. number of regions
3 i.e., control of binocular system (hardware and software)	directly measured: vergence angle computed: threshold of similarity criteria for correspondence purposes range of admissible depth values	Depth map, $2\frac{1}{2}$ sketch
4 i.e., control of intermediate geometric $2\frac{1}{2}$-D vision module	computed only: threshold of similarity criteria on surface growing and 3-D boundary detection	depth map segmentation into surface patches
5 i.e., control of a) several views b) integration process c) matching between data and the model	measured directly: the position, viewing angle of several views of the scene computed only: threshold on similarity criteria between consecutive views threshold on similarity criteria between the data and the model	3-D object description via volumetric categories
6 i.e., control of semantic interpretation	computed only: the similarity criterion on matching between the data and the model	3-D object description via the model. The model can be of various complexity, i.e., basic categories, hypercategories, and subcategories.

Several comments are in order:

1) Although we have presented the levels in a sequential order, we do not believe that is the only way of the flow of information through the system. The only significance in the order of levels is that the lower levels are somewhat more basic and necessary for the higher levels to function.
2) In fact, the choice of at which level one accesses the system very much depends on the given task and/or the goal.

This design is closest in spirit to that of Brooks [15]. In the remaining part of this chapter we shall present two separate scenarios (one bottom-up, the other top-down) which will use some of the control levels from Table 1.

V. The Bottom-Up Scenario

In this scenario we begin without any *a priori* given task or request. Hence the system must generate a task for itself as: Open your eyes and look around, perhaps measure how far are you from an object or segment the image. Naturally for the system to be able to do so, it must have some built-in models of the signal and geometry in order to be able to function at all. Examples of such models are: edge models (step functions, linear edges, etc.), region models (piecewise constant, piecewise linear, etc.), and topological models.

There are two test cases that we shall present: one is to obtain depth maps using range from focus and vergence/stereo; the other is 2-D image segmentation.

A. Obtaining Depth Maps

The objective of the first case was to study the control strategies in the data driven situation of an Agile Camera system and how to combine information from different low-level vision modules, such as focus and stereo ranging [16]. The Agile camera system is an 11 degrees of freedom system under computer control. The Camera system is shown in Fig. 1.

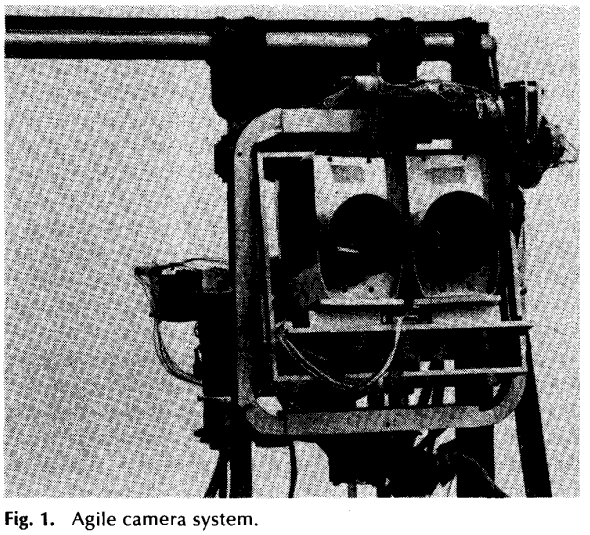


Fig. 1. Agile camera system.

Two cameras each with 340 by 480 spatial resolution, gray scale 256 levels. Each camera has a control of *Focus, Zoom*, and *Aperature*. There is a control of the vergence angle between the two cameras. There is a coupled control of *Pan/Tilt*, of *Up/Down* motion and of *Right/Left* motion. In addition there is available a controlled lighting system in order to simulate: lambertian source, one point source up to four point sources.

The initialization stage is comprised from two steps: Waking up the camera system and Gross focusing. After accomplishing this, we proceed to the intermediate stage which is carried out by orienting the cameras for stereo imaging, stereo ranging with verification by focusing, and focus ranging with verification by stereo. The final task is to obtain a depth map with a given accuracy.

1) Waking Up: Waking up the sensors involves opening the device controllers and setting them to default parameters. The lens controller zooms both lenses out, focuses them at a distance of 2 m, opens the aperatures, and verges to a parallel position. The platform controller positions the camera in the middle of the gantry, and orients them to look straight ahead. The light controller adaptively illuminates the lamps until the image intensities reach reasonable mean values, and good contrast. Now awake, the camera system is ready to acquire images for any application.

2) Gross Focusing: One of the cameras is the master, and the other is the slave. To bring the images into sharp focus and compute an initial, rough estimate of the range of objects in the scene, we grossly focus the master lens, beginning by zooming in the master camera. As analysed in Krotkov's work [16], the depth of field of the lens limits the precision of the computation of range from focusing. To maximize this precision, the depth of field should be as small as possible, which can be achieved by increasing the aperature diameter. For this, the process starts by turning off all the lights, and then opens the aperature as wide as possible without saturating. (An image saturates when more than a certain n S of pixels are at their maximum intensity; the value $S = 200$ is employed in the current implementation.) Next it adaptively illuminates the lamps until image saturation, and finally turns them down slightly. The gross focusing process determines the focusing distance bringing the scene into sharp focus by Fibonacci search over the entire field of view and the entire space of possible focus motor positions. The process records the best focus motor position and the range Z. The gross focusing terminates by zooming the master lens out, and servoing the slave focus motor position to that of the master lens.

3) Orienting the Cameras: The task now is to orient the cameras so that objects at distance Z will lie in the field of view and have zero disparity. Using the formula derived by Krotkov, one can compute the vergence angle from which by further calculation we get the vergence motor position corresponding to the vergence angle and finally servoing the vergence motor to this position. After verging, some objects may have drifted out of view. To reacquire these objects, a corrective pan by the amount $-\alpha/2$ is executed.

The reason we have gone into such details of explaining this initial stage is to make the reader aware of how much interaction and feedback takes place just in the process of initial adjustment of the camera system.

4) The Final Task—Depth Maps: The steps carry out the

computation of how to get range from focus and range from stereo. Here we shall not describe these follow-up steps in detail, but just to summarize the results of this portion:

1) It is a cooperative sensing operation in that the two modules mutually check each other, that is, stereo ranging is verified by focusing and focus ranging is verified by stereo ranging.
2) The result is the conservative measure of distance that has passed the error analysis of both the processes.
3) There is no feedback only feed forward in this portion.

In summary, Krotkov's work is an example of data driven initialization with the goal of obtaining depth map of a scene. An alternative approach to the intermediate goal of the Initialization stage can be 2-D segmentation of each view before or concurrently performed with the process for obtaining range from focus.

There are other researchers who have integrated different visual cues for obtaining range data, examples are Abbot and Ahuja [17], Hoff and Ahuja [18]. The novelty of Krotkov's work is in the thorough modeling of each module, the explicit parametrization which allowed him to predict a range of distance values in which the measurements are physically possible. Furthermore, he measured the *a posteriori* distribution of errors during the calibration phase which allowed him to implement a policy for integrating the computed ranges into maximum likelihood estimate.

B. Image Segmentation, Another Example of Bottom-Up Process

The group at the University of Massachusetts [19] have been advocating for some time that one must be goal directed in order to do low level image processing and/or segmentation. They argue that the failure of general segmentation technique can be traced to the following:

1) the image is too complex because of the physical situation from which the image was derived and/or the nature of the scene; or,
2) there is a problem of evaluating different region/line segmentations.

They say: "... it has been our experience that no low-level evaluation measure restricted to making measurements on the segmentation can provide a useful comparative metric. Rather the quality of segmentation can be measured only with respect to the goals of an interpretation ..."

We agree that the images are complex, but some of the image acquisition process can be modeled, and hence, one can account for the variability of the acquisition process as shown by Krotkov [16]. Of course, one cannot predict the spatial arrangement of objects, (their surface properties) but we can have models that are somewhat invariant to these variables. We also agree with the authors that there are no good segmentation evaluation functions. It is known that segmentation process is not unique given any number of parameters. But we wish to argue with the authors that the only thing that can determine the segmentation is the goal of its interpretation. If this would be so, then we cannot ever have a general (or even semi-general) purpose system which can bootstrap itself and adapt to an *a priori* unknown environment. In reality, they do present an intermediate system called GOLDIE which indeed has several syntactic (context-independent) evaluation functions. This is quite encouraging although it still must be tested in a variety of domains. (It was tested only in one domain, so far.)

Recently Anderson [20] has considered a modular, context independent approach to the problem of 2-D segmentation, denoted as the control level 2 in Table 1. The modules are edge and region formation modules, as shown in Fig. 2. It is a well-known fact that one can get very different segmentations from a picture by just changing the parameters.

The most important results of this work so far are the following:

1) definition of the task/goal and parameters for each module in terms of general, geometric goals rather than with respect to context and semantic information,
2) an extensive analysis was done on the relationship between the parameters and the false detection errors and the false dismissal errors of a true object boundary,
3) the detection of boundaries is an interdependent process between edge and region formation process,
4) the idea of feedback within a module and the interdependency between modules, implies multiple outputs and hence the need for fusion, i.e., combination rules.

Items 1) and 2) are issues of local models while items 3) and 4) are aspects of global models. The most popular approach to global models in the context of image segmentation is the cooperative network or relaxation approach [21]. Another such model is the Random Markov Fields model used by several researchers [22]–[24]. The principle of this model is that the effects of members of the field upon each other are limited to local interaction as defined by the neighborhood. This is also the weakness of this model because it assumes *a priori* spatial arrangement of objects and their projection on the image. This assumption is too strong, and applicable only in a few, highly controlled experiments. The work of Anderson shows that for image segmentation the global models should represent topological and integral (size and number of regions/edges) properties which are positionally invariant rather than neighborhood dependent (Fig. 4). The following predictions have been verified: the minimum separability between two adjacent objects, and the amount of detectable details on the boundary, (this follows from the width of the filter that is used with the edge detector) the homogeneity criterion for regions depends on the number of expected segments in the image. In Fig. 3 we show the original image, then after edge detection with the width of the bandpass filter 3 pixels. Note that the pennies separated less than 3 pixels are merged together, as predicted. The last portion of this figure, labeled "Segmentation" is the result of a combination of edge detection and region growing which operates on local similarity in the neighborhood of one pixel. This small neighborhood explains the fact that one detects the touching boundary. The magnitude of similarity criteria eliminates the single lines. The magnitude of the similarity criterion is a parameter that is adaptively determined by external criterion of the number of desired regions. We have investigated this relationship between the number of

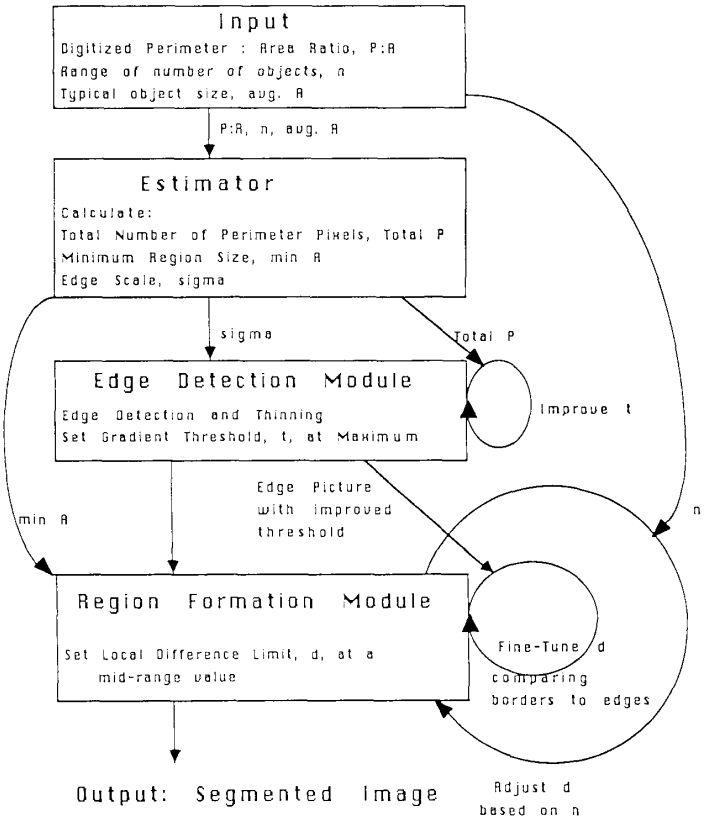


Fig. 2. Segmentation system.

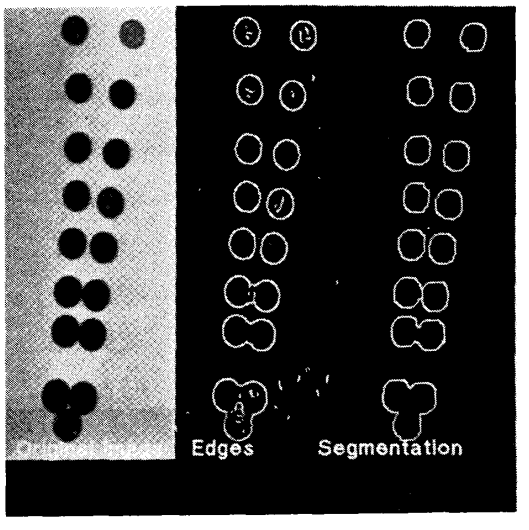


Fig. 3. Image segmentation of pennies.

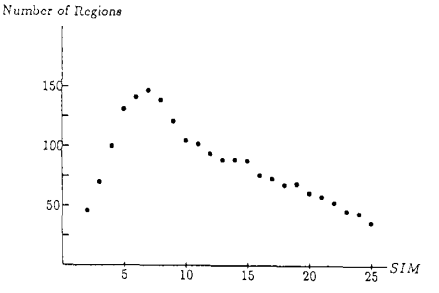


Fig. 4. Relationship of the number of regions and the similarity magnitude.

regions and the similarity magnitude, and it is shown in Fig. 4. This curve shows that up to a point there is almost a linear relationship between the similarity magnitude and the number of regions. As one goes beyond a certain value (in this graph around 10) the number of regions increases because one has detected more noise than true segmented regions.

Although we have shown that one can have a theory with predictions on sensitivity and robustness of 2-D segmentation and obtaining 3-D range data, there are still many open questions left in the bottom-up scenario. Some of them are the following:

- parallel explanation versus sequential,
- partial explanation versus total, i.e.,
  - Should one say: "As far as I can see this is it . . ." or "From this view this is . . . , from another view this is . . ." etc.

VI. Top-Down Scenario

We shall consider two cases of the top-down initialization process: Task driven and query driven. While in the bottom-up mode the strategy is to start with the data, analyze it, identify some gestalt property, such as a closed contour, or the largest or brightest object; in the top-down mode the strategy is to start with the *Data Base* (entered either via query or by the Task Description), which should suggest what perceptual properties to look for. The task-driven mode differs from the query-driven mode only in the fact that in the task-driven mode the Active Vision System not only has to identify the queried object but also must monitor the work space, measure the changes that occur until the task is accomplished. But most importantly, in the task-driven mode, the system by interacting with its environment may change it; while when you are a reporter, you have no direct influence on events that you are reporting about. In the Query Driven mode the system acts as *an observer*, while the Task-Driven mode, the system is a participant. An interesting example is the DARPA project, Autonomous Land Vehicle paradigm. Within this paradigm, let us suppose that the task of an autonomous vehicle is to get from place *A* to place *B*, stay on the road, and avoid obstacles if there are any. Then we can ask further: is the Active Vision system acting as an Observer or a Participant? At some level of abstraction, it acts as an Observer, but on the sensor-action level it behaves as a Participant. Consider for example that the wheels leave tracks, and hence, change the surface of the road which will require a change in processing strategy. The top-down, model-driven recognition systems are quite popular in the vision community. There are several review articles, an example is Binford's work [25]. Their advantage to the bottom-up scenario is that the expectations, models can help in the recognition process even in the case of rather noisy and incomplete data. Naturally, the danger in this mode is that one may see anything that one wishes to see and not detect unexpected objects. "How does the Active Vision approach fit into the top-down scenario?" Clearly, if the expectations fit the data there is no need for further control. But this is hardly ever the reality. In any other case, i.e., when decisions must be made during the intermediate stage, the Active Vision approach becomes a necessity. In spite of all the past work in this mode there are still open issues, such as the following:

- What is a satisfactory answer to a given query?
- Should one include into the answer a degree of uncertainty?
- Should one answer via another question?
- If any doubts what to do?
- In the task driven mode should one generate another subtask?
- Can one talk about convergence to an answer?

VII. The Intermediate Stage

After the initial processing discussed in Sections IV and V, we have either: a) a segmented scene or at least labeled entities such as two-dimensional regions, or lines or 3-D points (after a bottom-up initialization), in other words, some geometric description; or b) a set of perceptual properties to look for (after the top-down initialization).

In the case of a) we go into the *Knowledge* Base to search for some *hypothesis-concepts*. In the case of b) we go into the *sensory data* to search for the expected *perceptual* properties, which are formed by the low level models. The common fact for both of these cases a) and b) is that both have a number N: the number of concepts-hypothesis or the number of objects with given perceptual properties. Now let us study this number N: If $N = 0$ then this means in case a) there is no hypothesis that would match the measurements; in case b) there is no object in the data that matches the expectations. If $N = 1$ then there is a unique interpretation of the data. If $N > 1$ then in case a) we have multiple hypothesis; in case b) we have multiple objects in the data with expected perceptual properties. This number N can be a feedback signal for further processing. For example, if $N = 1$ we are done, unless we are specifically looking for more than one object in the data. In the case of interpretation we are done unequivocally. If $N = 0$ we are also done, at least for the moment. This fact tells us that in the current data the searched object does not exist. If $N > 1$, then the question is: "What should be the Nmax?" In the case of a) the Nmax corresponds to a short term conceptual buffer, which contains the hypothesis; in the case of b) the Nmax corresponds to the short term visual memory.

The open question is: "How large should this buffer be?" For the conceptual buffer, we would like to argue that it should be: $7 +$ or $- 2$. This number is based on psychological studies [26]. In any case, whatever the Nmax is the question remains: "What do you do next?"

The intermediate stage is structured and subdivided in our Table 1 into levels 3 through 6. The issue here is what are suitable models for grouping purposes, which would lead to useful geometric descriptions.

A very interesting Global model in this sense has been inspired by the influence of the Gestalt psychological models implemented in Reynolds and Beveridge [27]. The global model is a geometric grouping system based on similarity and spatial proximity of geometric tokens. Examples for lines are colinearity, rectangularity, and so on. We think that this is a promising approach but what is missing is evaluation criterion and robustness studies. This cannot be avoided especially with such vague notions as similarity and spatial (or other) proximity relations.

Another way of explaining the 3-D data is to choose volumetric models that can be estimated and fit to the data. The difficult task here is the choice of such representation which naturally maps into shape categories. There are several possibilities as reviewed thoroughly by Besl and Jain [28]. One such model is the superquadric function [29], characterized by parameters: position and orientation in space, size and roundness/squareness of edges, amount of tapering along the major axis, amount of bending along the major axis, and the size of cavities (its depth and steepness). Given these parameters, it is not difficult to see that one can make classification of objects with respect to their shapes into categories: squash, mellon, pear, etc.

Fig. 5 displays an example of a squash. From top-left to bottom-right, the sequence shows first the gray scale image of the object, then the range data, and then the sequence of fits to the data starting with an ellipsoid and finishing with the best fit measured by the residuum. As the side product, one gets the above mentioned parameters. Naturally, there are common characteristics but also discernible features to each of these categories. Those distinguish-

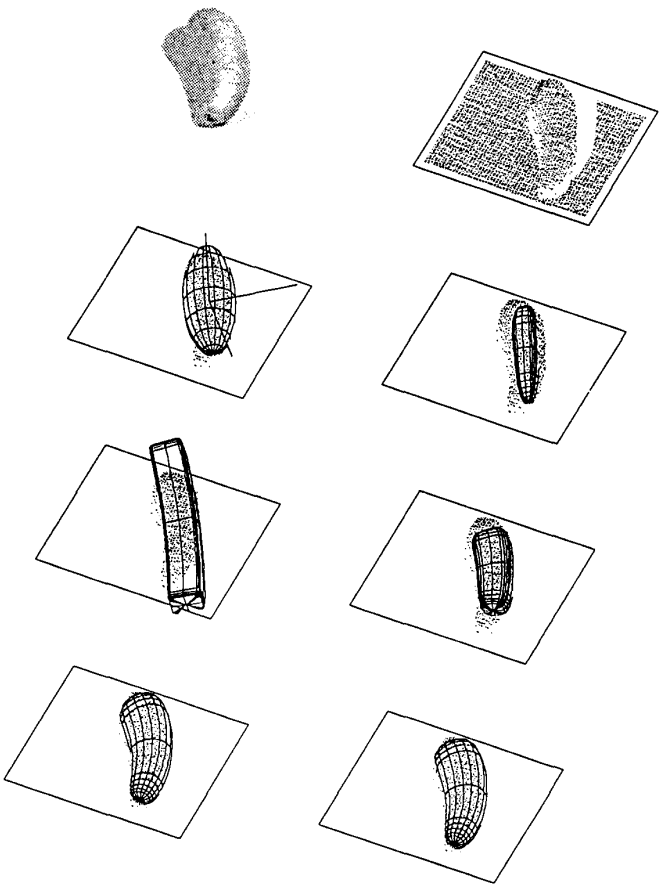


Fig. 5. Model recovery of a tapered and bent object—a squash.

ing properties will reduce the number N, i.e., the number of possible interpretations.

VIII. Search for Information

The fundamental ingredient to the Active Vision theory is the mechanism for decision making, in general, how to choose data information with respect to the perceptual goal. This involves both the method of estimation and the choices of general strategy for control. Following the work of Hager and Mintz [30], we formalize a controllable measurement device as a mathematical system of the form:

$$z_i = H(u_i, p) + V(u_i, p)$$

where u_i is m-dimensional control vector, p is s-dimensional quantity we are attempting to estimate, $V(\cdot, \cdot)$ is additive noise, and z_i is the observation.

The problem is to optimize, by choice of some sequence $\boldsymbol{u} = [u_1, u_2, \cdots, u_n]$ the performance of an estimation procedure $\delta_n(\cdot)$ estimating p from $\boldsymbol{z} = [z_1, z_2, \cdots, z_n]$. In order to choose a particular estimation procedure, we must pick a criterion or loss by which to judge the merit of a decision rule. A commonly used loss criterion is the mean square error loss. The optimal estimation procedure is that which given an *a priori* probability distribution on p and the conditional distribution on observations, minimizes the quantity

$$E\|\delta(z) - p\|^2.$$

This adds the Bayes solution to our estimation problem [31].

A control sequence is to be evaluated relative to its expected utility. This utility can be thought of in two parts: the performance of the estimation procedure for that choice of strategy and the cost of implementing that strategy. We can write a general loss for the combined estimation/control problem as

$$l(p, \hat{p}, n, \boldsymbol{u}) = l^d(p, \hat{p}) + c(n, \boldsymbol{u})$$

l^d represents the loss attributed to the estimation procedure δ, and c represents the cost of taking n samples via control strategy $\boldsymbol{u}$. The choice of actual forms for c and l^d reflects the desired behavior of the system.

Since sensors are to be used for reducing uncertainty, an appropriate quality measure is a tolerance, ϵ. Results must be computed within some time constraints balanced against the sensor's resource budget. Finally, the results returned by the sensor should indicate the degree to which the set task was found achievable. Based on this discussion, an appropriate evaluation criterion for the performance of an estimation procedure δ_n in the context of control is the 0-1 loss

$$l^d(p, \delta_n(\boldsymbol{z})) = \begin{cases} 0, & \text{if } |\delta_n(\boldsymbol{z}) - p| < \epsilon \\ 1, & \text{Otherwise.} \end{cases}$$

By the previous definition of the loss function we can compute the Bayes decision risk of an estimation δ_n as

$$r(n, \boldsymbol{u}, \delta) = E_p E_{z|p}[L^d(p, \delta_n(z)) + c(n, \boldsymbol{u})]$$
$$= r^d(\pi, \delta_n) + c(n, \boldsymbol{u})$$

If we disregard $\boldsymbol{u}$, then we can consider the problem of finding the n which minimizes the risk function for fixed $\boldsymbol{u}$. This procedure is called *batch* procedure. On the other hand we could also solve this problem by evaluating the risk conditioned on the observed data, and derive the *stopping rule* which says when enough data had been taken. This is of course a sequential procedure in nature. In addition to determining batch size or stopping rule, we must choose a plan of control to minimize the risk of the final outcome. That we do by minimizing the risk given δ_n:

$$\min_n \min_{\boldsymbol{u}} r(n, \boldsymbol{u}, \delta_n).$$

Recently a great interest has been spurred in the computer vision and robotics community on the integration of multisensory information. A prime example of this is the workshop on spatial reasoning and uncertainty, St. Charles, IL, October 1987, organized by A. Kak. The question is how to integrate multisensory information. For systems which are linear both in state and control parameters, the Kalman filter is the optimal linear estimation procedure. When the system is static, the Kalman filter becomes an iterative implementation of Baysian estimation procedure. There are no such general solutions for nonlinear measurement systems, though there are a number of approximation techniques [32]. One approximation is to linearize a given nonlinear measurement system about a nominal trajectory and apply linear estimation technique. In reference to Kalman filter, this approximation is called the Extended Kalman Filter (EKF).

Durant-Whyte [33] used a version of this technique to solve the problem of updating location estimates from observations. Smith *et al.* [34] have applied this method to a mobile robot estimating its position. Ayache and Faugeras [35] have looked at several problems in stereo ranging by building an EKF a general constraint equation. However, it is important to remember that EKF is only an approximation. The accuracy of this approximation depends on having a relatively good prior estimate to linearize about and the approximation is only good in a small neighborhood about this point. Another method of nonlinear estimation which does not rely on linearization is Stochastic Approximation [36]. This technique is similar to Newton's method (iterative method) adapted to work in the presence of noise. The choice of gain sequences is crucial to convergence. There are no results for the small-sample behavior of this estimator. The next question is the control of nonlinear systems.

While there are some results [31] in Bayesian sequential and batch decision problems, there is no general theory. Hager and Mintz [30] have analyzed theoretically and in simulation the EKF approach to the active perception task and have made the following conclusions: The Active perception setting has the following attributes:

1) the system is nonlinear both in state and control,
2) the measurement noise depends on the control of measurement system,
3) the control criterion is a direct function of the information returned by the estimation procedure, and
4) the information is limited by sensor scope.

Given this setting, the EKF fails on the robustness issue. This is so because the EKF is, like stochastic approximation, a differential correction technique and unless one has a good prior estimate it will not converge. One can ask: "Can we guarantee convergence by an appropriate gain?" Hager and Mintz have shown via simulation that as the interval of uncertainty widens, the error terms will drastically underrepresent the error estimation. For large enough intervals the filter fails to converge. Next they investigate the finite Bayes approximation and show that this technique is more robust and gives more predictable estimations. The important result of the work of Hager and Mintz is the critical analysis of the linearization techniques that are quite popular currently in robotics and active perception. It is an important methodology to examine which methods/tools are appropriate and what are the consequences of the approximations made at each step to a given problem. It also points to open problems that require new mathematical and formal tools that should be a challenge to theoreticians.

IX. Conclusions

In conclusion we have defined active perception as a problem of an intelligent data acquisition process. For that, one needs to define and measure parameters and errors from the scene which in turn can be fed back to control the data acquisition process. This is a difficult though important problem. Why? The difficulty is in the fact that many of the feedback parameters are context and scene dependent. The precise definition of these parameters depends on thorough understanding of the data acquisition devices (camera parameters, illumination and reflectance parameters), algorithms (edge detectors, region growers, 3-D recovery procedures) as well as the goal of the visual processing. The importance however of this understanding is that one does not spend time on processing and artificially improving imperfect data but rather on accepting imperfect, noisy data as a matter of fact and incorporating it into the overall processing strategy.

Why has it not been pursued earlier? The usual answers one gets are: lack of understanding of static images, the need to solve simpler problems first, less data, etc. This of course is a misconception. One view lacks information which is laboriously recovered when more measurements, i.e., more views can resolve the problem easier. More views add a new dimension—time—which requires new understanding, new techniques, and new paradigms.

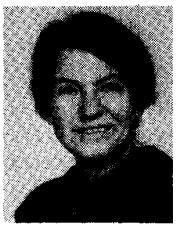
Acknowledgment

The authors would like to thank Patricia Grosse for her editorial assistance on this document.

References

[1] J. M. Tenenbaum, "Accommodation in computer vision," Ph.D. Thesis, Stanford University, Nov. 1970.
[2] ——, "A laboratory for hand-eye research," *IFIPS*, pp. 206–210, 1971.
[3] R. Bracho, J. F. Schlag, and A. C. Sanderson, "POPEYE: A gray-

level vision system for robotics applications," *CMU-RI-TR-83-6*, May 1983.
[4] L. E. Weiss, "Dynamic visual servo control of robots: An adaptive image-based approach," *CMU-RI-TR-84-16*, 1984.
[5] Y. Kuno, H. Numagami, M. Ishikawa, H. Hoshino, and M. Kidode, "Three-dimensional vision techniques for an advanced robot system," *IEEE Conf. on Robotics and Automation*, (St. Louis, MO), pp. 11-16, Mar. 1985.
[6] D. H. Ballard, "Eye movements and spatial cognition," Computer Science Department, University of Rochester, TR-218, Nov. 1987.
[7] T. Poggio, "MIT progress in understanding images," in *Proceedings of DARPA IU Workshop*, (Los Angeles, CA), pp. 41-54, Feb. 1987.
[8] P. Allen and R. Bajcsy, "Converging disparate sensory data," in *Proceedings of the 2nd Int. Symp. on Robotics Research*, H. Hanafusa and H. Inoue, Eds. Cambridge, MA: MIT Press, 1985, pp. 81-87.
[9] J. Aloimonos and A. Badyopadhyay, "Active vision," in *IEEE 1st Int. Conf. on Computer Vision*, pp. 35-54, June 1987.
[10] S. W. Zucker, "Early orientation selection: Tangent fields and the dimensionality of their support," Computer Vision & Robotics Laboratory, Department of Electrical Engineering, McGill University, Montreal, Quebec, Canada, TR-85-13-R, May 1985.
[11] Y. G. Leclerc and S. W. Zucker, "The local structure of image discontinuities in one-dimension," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, no. 3, pp. 341-355, May 1987.
[12] V. S. Nalwa and T. O. Binford, "On detecting edges," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8, no. 6, pp. 699-714, Nov. 1986.
[13] R. M. Haralick and L. G. Shapiro, "Image segmentation techniques," *Computer, Vision, Graphics, and Image Processing*, vol. 29, no. 1, pp. 100-133, Jan. 1985.
[14] R. Pavlidis and Y. T. Liou, "Integrating region growing and edge detection," submitted to *ICVPR*, 1988.
[15] R. A. Brooks, "A hardware retargetable distributed layered architecture for mobile robot control," in *Proceedings of IEEE Int. Conf. on Robotics*, vol. 1, pp. 106-110, 1987.
[16] E. Krotkov, "Exploratory visual sensing for determining spatial layout with an agile stereo camera system," University of Pennsylvania Ph.D. Dissertation also available as a Tech. Rep. MS-CIS-87-29, Apr. 1987.
[17] L. Abbott and N. Ahuja, "Surface reconstruction by integrating focus and stereo," to be published in proceedings of *CVPR*, Ann Arbor, MI, 1988.
[18] W. Hoff and N. Ahuja, "Surfaces from stereo," in *Proceedings Image Understanding Workshop*, pp. 98-106, Dec. 1985.
[19] C. A. Kohl, A. R. Hanson, and E. M. Riseman, "Goal directed control of low-level processes for image interpretation," in *IU Proceedings DARPA*, vol. 2, pp. 538-551, Feb. 1987.
[20] H. Anderson, "Edge-detection for object recognition in aerial photographs," University of Pennsylvania, Grasp Laboratory, Tech. Rep. MS-CIS-87-96, 1987.
[21] A. Rosenfeld, R. A. Hummel, and S. U. Zucker, "Scene labelling by relaxation operations," *IEEE Trans. Syst. Man Cybern.*, pp. 420-433, 1976.
[22] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributing and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, pp. 721-741, Nov. 1984.
[23] H. Derin and C. S. Won, "A parallel image segmentation algorithm using relaxation with varying N neighborhoods and its mappings to array processors," *CVGIP*, vol. 40, no. 1, pp. 54-78, Oct. 1987.
[24] G. R. Cross and A. K. Jain, "Markow random field texture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 5, pp. 25-39, 1983.
[25] T. O. Binford, "Survey of model based image analysis systems," *Int. J. Robotics Res.*, vol. 11, (Spring), pp. 18-64, 1982.
[26] G. Miller, "The magic number seven, plus or minus two," *Psychological Review*, vol. 63, no. 2, pp. 81-97, 1956.
[27] G. Reynolds and J. R. Beveridge, "Searching for geometric structure in image of natural scenes," in *Image Understanding Proceedings DARPA*, vol. I, pp. 257-271, Feb. 1987.
[28] P. Besl and R. Jain, "Three-dimensional object recognition," *Computing Surveys*, vol. 17, no. 1, Mar. 1985.
[29] F. Solina, "Shape recovery and segmentation with deformable part model," University of Pennsylvania Ph.D. Dissertation also available as a Tech. Rep. TR-MS-CIS-87-111, Dec. 1987.
[30] G. Hager and M. Mintz, "Estimation procedures for robust sensor control," University of Pennsylvania Tech. Rep. MS-CIS-87-109, Feb. 1987.
[31] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York, NY: Springer-Verlag, 1985.
[32] A. Gelb, Ed., *Applied Optimal Estimation*. Cambridge, MA MIT Press, 1974.
[33] H. Durrant-Whyte, "Integration and coordination of multisensor robot systems," University of Pennsylvania Ph.D. Dissertation, Aug. 1986.
[34] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," in *Proceedings of Uncertainty in Artificial Intelligence*, (Philadelphia, PA), pp. 267-289, Aug. 1986.
[35] N. Ayache and O. D. Faugeras, "Building, registrating and fusing noisy visual maps," submitted for publication in *Robotics Research*, 1986.
[36] A. E. Albert and L. A. Gardner, *Stochastic Approximation and Nonlinear Regression*, vol. 42 of Research Monograph. Cambridge, MA: MIT Press, 1967.



Ruzena Bajcsy (Member, IEEE) received the M.S.E.E. degree in mathematics and the Ph.D. degree in electrical engineering from Slovak Technical University, in 1957 and 1967, respectively. She received a second Ph.D. degree in computer science from Stanford University in 1972.

She joined the Computer and Information Science Department at the University of Pennsylvania first as an Assistant Professor in 1972, and now serves as Professor and Chairman of the department as well as Director of the GRASP Laboratory (General Robotics Active Sensory Perception). She has authored numerous book chapters and journal publications and has served as editor and associate editor of several journals including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. Her research interests include multi-sensor integration, robotics, and computer vision.