Promising Directions in Active Vision

Written by the attendees of the NSF Active Vision Workshop
University of Chicago, August 5-7, 1991

Edited by Michael J. Swain and Markus Stricker

Sponsored by the National Science Foundation,
IBM and Hughes Research Laboratories

University of Chicago Technical Report CS 91-27
November, 1991

For additional copies, write:

Department of Computer Science
University of Chicago
1100 E. 58th Street
Chicago, Illinois 60637
U.S.A.

Contributors:

A. Lynn Abbott	Virginia Polytechnic Institute
Narendra Ahuja	University of Illinois
Peter K. Allen	Columbia University
John (Yiannis) Aloimonos	University of Maryland
Minoru Asada	Osaka University
Ruzena Bajcsy	University of Pennsylvania
Dana H. Ballard	University of Rochester
Ben Bederson	Vision Applications Inc.
Ruud M. Bolle	IBM
Peter Bonasso	The MITRE Corporation
Christopher M. Brown	University of Rochester
Peter J. Burt	SRI/David Sarnoff Research Center
David Chapman	Teleos Research
James J. Clark	Harvard University
Jonathon H. Connell	IBM
Paul R. Cooper	Northwestern University
Jill D. Crisman	Northeastern University
James L. Crowley	LIFIA (France)
Michael J. Daily	Hughes Research Laboratories
Jan-Olaf Eklundh	Royal Institute of Technology, Sweden
Frank P. Ferrie	McGill University
R. James Firby	University of Chicago
Martin Herman	National Institute of Science and Technology
Philip Kahn	Advanced Decision Systems
Eric Krotkov	Carnegie Mellon University
Niels da Vitoria Lobo	University of Toronto
Howard Moraff	National Science Foundation
Randal C. Nelson	University of Rochester
H. Keith Nishihara	Teleos Research
Thomas J. Olson	University of Virginia
Daniel Raviv	Florida Atlantic University
Giulio Sandini	University of Genoa
Eric L. Schwartz	New York University / Vision Applications Inc.
Markus Stricker	University of Chicago
Michael J. Swain	University of Chicago
John K. Tsotsos	University of Toronto
Richard Wallace	Vision Applications Inc.

Summary

Active vision systems have mechanisms that can actively control camera parameters such as position, orientation, focus, zoom, aperture and vergence (in a two camera system) in response to the requirements of the task and external stimuli. They may also have features such as spatially variant (foveal) sensors. More broadly, active vision encompasses *attention*, selective sensing in space, resolution and time, whether it is achieved by modifying physical camera parameters or the way data is processed after leaving the camera.

In the active vision paradigm, the basic components of the visual system are visual behaviors tightly integrated with the actions they support; these behaviors may not require elaborate categorical representations of the 3-D world. Because the cost of generating and updating a complete, detailed model of our everyday environment is too high, this approach to vision is vital for achieving robust, real-time perception in the real world. In addition, active control of imaging parameters has been shown to simplify scene interpretation by eliminating the ambiguity present in single images.

This document describes promising directions for research in active vision and possible applications of this research. It also discusses progress in experimental equipment for supporting this research and potential applications. Important research areas in active vision include attention, foveal sensing, gaze control, eye-hand coordination, and integration with robot architectures:

Attention Selective processing of regions with restricted location, motion, or depth, is necessary for achieving real-time performance with limited resources. Although the design of the attention system is potentially highly complex, and will affect the design of visual processing at all levels, little is known in detail about what this system should look like.

Foveal Sensing A spatially-variant (foveal) sensor permits high resolution at the location of interest without the cost of uniformly high resolution. Issues that need exploring include the properties of unconventional "log-polar" coordinate systems specifically suited to foveal sensing, and the determination of visual features which

can be informative at low resolution to allow reliable selection of targets in the periphery.

Gaze Control The alteration of imaging parameters to aid in the performance of visual tasks, or *gaze control*, is useful for many tasks, including image stabilization, overcoming a limited field of view, figure-ground separation and range estimation. Gaze control is divided into two primary categories: Gaze stabilization and gaze change. The former consists of controlling the camera to maintain clear images of some world point that may be in motion with respect to the camera. The latter category is used to transfer stabilized gaze to new fixation points, similar to human saccadic eye movements.

Gaze Stabilization Timely extraction of the error signal is an important problem in this area. Several interdependent layers of control are needed to deal with the many different conditions under which stabilization or tracking is performed. Stabilization can involve inputs from different vision algorithms and sensors. The integration of these is an important, difficult problem. Work is also needed in the design of robust, adaptable domain and plant models. These models are needed to improve tracking of objects with various patterns of movement and robustness to errors in calibration of the camera-motor system. At a higher level, gaze stabilization must be integrated with methods of control changes in gaze.

Gaze Change Changes in gaze are often motivated by the task; learning the pattern of eye movements to accomplish a given task is an important but relatively unexplored problem. Tasks may have conflicting requirements, as in the case where different tasks would direct the fovea to different points in visual space. As systems with multiple behaviors are built, this scheduling problem will become increasingly paramount.

Eye-Hand Coordination Sensors on a manipulator can augment vision with additional information. Manipulation behavior also provides a performance metric for the vision system and geometric feedback.

Eye-hand systems offer the unique ability to determine material, geometric, and haptic characteristics of objects that are inaccessible to vision systems alone. We need representations and strategies for integrating visual information with hand sensory information.

We must build systems that function robustly without accurate calibration. This is particularly important for eye-hand systems, since both subsystems have distinct geometric parameters and the subsystems must work together. An eye-hand system offers a unique potential for self-calibration, by sending motion instructions to the manipulator and observing their effects.

We see the single biggest obstacle to progress in this area as the cost of hand systems with significant capabilities. We would like to see a cost effective hand system be developed so that it becomes more affordable for many laboratories.

Integration with Robot Architectures Active vision research should be conducted in the context of ongoing research in robot architectures. This is important for two reasons. First, many of the relevant issues are being addressed by people doing research in robot architectures, and active vision investigators can benefit from these experiences. Secondly, researchers developing visual routines need to maintain a view toward the purposes for which robot architectures are being investigated. Important issues to be researched in this context include:

- What set of primitive visual operations are necessary for robot tasks in general?
- What are the appropriate levels of primitive operations?
- What is the right combination of pre-attentive and attentive processing?

Within the document we have made suggestions for benchmarks within the areas of research listed above. More importantly, we believe that the task of developing meaningful benchmarks applicable to many different types of active vision systems is itself a difficult research problem deserving support.

Hardware Support Simulation is inadequate for verifying and demonstrating active vision algorithms. To facilitate experimental research in active vision, we recommend the funding of design, development, and distribution of an inexpensive integrated camera platform with controllable optics. Funding for novel hardware concepts should also be available. For example, emphasis needs to be placed on investigating novel actuator technologies to make eye platforms that are smaller, faster, and more accurate.

Advances in image processing hardware and sensor technology have assisted progress in the active vision field. Fast image acquisition and low-level image processing are currently within the financial reach of most vision labs. The design of special purpose accelerator boards designed by vision researchers is becoming feasible and should be encouraged. Research in foveal sensor arrays is of interest, as well as digital technology for real-time simulation of foveal arrays by remapping the output of a uniform sensor.

While image processing is often well described by SIMD models, MIMD architectures may well be necessary to support the necessary concurrency of cooperating high-level vision, control, reasoning and planning tasks. Therefore, programming languages and software environments that make programming heterogeneous SIMD/MIMD systems easier would facilitate active vision research.

Applications Active vision techniques can potentially enable large gains in cost and performance of computer vision systems in many application areas. In the document we give examples of these gains in the areas of unmanned vehicles, industrial processes, aids for the handicapped, and service robots.

Contents

1 Introduction **1**
1.1 Why Active Vision? 1
1.2 Outline 2

2 Research Directions **4**
2.1 Attention 5
2.2 Foveal Sensing 7
2.3 Gaze Control 9
2.4 Where to Look Next 16
2.5 Eye-Hand Coordination 22
2.6 Integration with Robot Architectures 27
2.7 Benchmarks 33

3 Hardware Support **38**
3.1 Mechanical Hardware 39
3.2 Electronic Hardware 45

4 Applications **49**
4.1 Introduction 49
4.2 Examples 50
4.3 Summary 54

5 Conclusion **56**

6 Bibliography **58**

1 Introduction

A growing number of researchers have found that the ability to actively control the camera parameters, and more generally the ability to combine vision with behavior, is vital to achieving robust, real-time perception for a robot interacting with a complex, dynamic world. In this paradigm called *active vision*, vision is most readily understood in the context of the visual behaviors that the system is engaged in; these behaviors may not require elaborate categorical representations of the 3-D world.

This document describes promising directions for research in active vision and possible applications of this research. It also discusses progress in experimental equipment for supporting this research. The document was written by the attendees of the Active Vision Workshop held at the University of Chicago, August 5-7, 1991. This workshop brought together an international collection of leaders within this emerging research community.

1.1 Why Active Vision?

Active vision systems have mechanisms that can actively control camera parameters such as orientation, focus, zoom, aperture and vergence (in a two camera system) in response to the requirements of the task and external stimuli. They may also have anthropomorphic features such as spatially variant (foveal) sensors. More broadly, active vision encompasses *attention*, selectively sensing in space, resolution and time, whether it is achieved by modifying physical camera parameters or the way data is processed after leaving the camera.

Active systems can dramatically simplify the computations of early vision; enabling areas of interest to be examined at the desired resolution without the costs of uniformly high resolution sensing, simplifying segmentation of an object of interest from its background, using controlled motions to disambiguate solutions that are otherwise underconstrained, examining unseen areas of the scene, and simplifying the transition from image to world coordinates.

The tight coupling between perception and action proposed in the active vision paradigm does not end with camera movements. In this paradigm, visual processing is tied closely with the activities it supports (navigation, manipulation, signaling danger or opportunity, etc.) allowing simplified control algorithms and scene representations, quick response time, and increased success at supporting the goals of the activities. In addition, integration provides new metrics to judge vision algorithms that are better tuned to applications of computer vision.

Current interest in active vision research has been made possible by the recent advent of lightweight cameras and off-the-shelf computer-controllable motors, and advances in real-time image processing equipment. The spread of the experimental equipment for active vision research is a good indicator of the growing popularity of the field. For example, the first computer-controlled binocular camera platform that could perform rapid changes in gaze was built at the University of Pennsylvania in the early 1980's. Now at least twenty other labs have acquired or will soon be acquiring such vision platforms, including: Harvard University, Hughes Research Laboratories, INRIA (France), LIFIA (France), McGill University, MIT, MITRE Corporation, New York University, Osaka University (Japan), Oxford University, Royal Institute of Technology (Sweden), Teleos Research, University of Chicago, University of Illinois, University of Rochester, University of Maryland, University of Massachusetts, University of Toronto, and York University (Canada).

1.2 Outline

Attention is discussed in Section 2.1. The concept of selective sensing, introduced here, motivates many of the topics discussed in later chapters. Foveal sensing is discussed in Section 2.2. It is essential with foveal sensor and often useful with a uniformly sampled sensor to maintain the camera axis pointing at the location of interest. This problem is called *gaze control*, and is the subject of Section 2.3. Many locations in the world need to be examined, either because of events which occur outside the agent, or because of the goals of the agent itself. How to shift attention to other areas of the scene and schedule the allocation of resources such as the fovea is the subject of Section 2.4.

One of the most important tasks that the human visual system engages in is eye-hand coordination. This task exhibits the locality that makes it a natural application of spatially-variant sensing. In addition, there are many other reasons why this is an ideal domain in which to study active vision. Because of the multitude of tasks that can be attempted with an eye-hand system, the domain is natural for investigating the computational advantages and robustness of task-based algorithms. Through the use of the manipulator, the agent can control more than just the camera parameters in order to simplify the interpretation of the world. Eye-hand coordination is discussed in Section 2.5.

In any system that combines perception and action, such as an eye-hand system or a mobile robot, the question of how to integrate sensing, planning and action arises. This is an especially important question in research in active vision because of the tight coupling proposed between perception and action. Section 2.6 discusses this question in the context of an *agent architecture*. Especially important here is the question of how a set of task-based vision/action modules or *routines* can be assembled to solve a wide variety of tasks.

The final section under the heading of Research Directions suggests benchmarks for evaluating progress in these areas of research and stresses the importance of having these measures.

Because active vision algorithms are tightly integrated with action, it has proven valuable to perform experiments on equipment in which the camera parameters (position, orientation, focus, zoom, aperture) are computer controllable and with which visual computations can be done in real-time. Fortunately, hardware such as lightweight motors, motor controller boards, cameras and real-time image processing equipment have become smaller, faster and more affordable, facilitating the construction of labs in which real-time experiments can be carried out. Within the community designs for computer-controllable binocular camera platforms have proliferated, and researchers are starting to build custom hardware for real-time execution of important visual modules that cannot be easily mapped onto standard image processing hardware. The potential for further advances in these areas is great. The state of the art and proposals to further it are described in two sections. Mechanical hardware, such as binocular heads, computer-controllable lenses, and manipulators for studying eye-hand coordination, is discussed in Section 3.1. Computer hardware for real-time image processing and foveal sensor arrays is discussed in Section 3.2.

The advantages of active vision point to real-world applications of active vision technology. Potential applications are outlined in Section 4.

2 Research Directions

2.1 Attention

The signal available to a system from a cameras contains information about pattern, motion, depth, color, etc. The source image signal has the following characteristics: (1) it contains far more data than can be analyzed by a practical vision system in real time; (2) it is highly redundant; and (3) information of importance to a given visual task is episodic within the source signal—it tends to be clumped in space (objects of interest) and time (events of interest).

It is useful to view the function of a foveated sensor as one of selective data reduction. Selection is in space, resolution, and time. The sensor can perform selection in other signal dimensions as well, such as in depth (vergence), motion (tracking), etc. Combining the sensor with early image processing provides further selective data reduction. Included in the sensor is the kind of processing that requires full signal data. This processing can include motion, stereo, certain types of feature extraction.

There are at least four mechanisms for data reduction in early visual processing: Signal selection, signal compression, signal abstraction, and property measurement. Each is described separately below.

Signal Selection Most data available in the image signal is irrelevant to a vision task. It is important, therefore, to focus analysis on the critical regions and events, that is to select critical segments of the signal for further processing while ignoring the majority of the signal that is not relevant. Signal selection can take place along various signal dimensions: Space, velocity, and distance.

- Focal Selection. For a given camera position, interest is spatially restricted to a small region within the camera's field of view. This (focal) region is observed at high spatial resolution, while the remaining (peripheral) field is observed at much reduced resolution. The fovea and periphery distinction can be implemented at the level of the sensor device and/or through subsequent signal processing.

The selection of a region of interest may be made independently of the location of the fovea; this is termed *covert attention*. In this case the distinction between the region of interest and the rest of the visual field is solely in how the information from the region is processed.

- Motion Selection. Selection can be achieved in motion by tracking an object moving at a particular velocity. The object tracked will be rendered stationary while objects or patterns moving at other velocities are non-stationary and perhaps blurred. Tracking is an effective means for isolating one component/pattern at a time in a complex world containing many differently moving patterns.

- Selection in Depth. A system can select objects at a particular distance from the camera using focus, stereo or motion parallax. Selection in stereo and motion can be realized by transformations that align regions of interest between images while leaving other regions unaligned. The 3-D locus of zero-disparity can be controlled flexibly as appropriate for the imaging conditions. Alignment can be achieved to very high precision, at high speed.

Data Compression The next step in data reduction after signal selection is compression. There remains considerable structure in the signal, and hence redundancy. In general terms, signal compression seeks to transform the way the signal is represented so that the same information is present, but the redundancy is removed. This is achieved by representing the signal in terms that mimic the pattern structure. For vision applications the representation should be well suited for subsequent processing.

Alerting While early processing selects a component of the signal for further analysis by the vision system, it must also provide some form of reduced representation of the signal as a whole. Only then can the system have the context it needs to determine where to look next. The reduced representation can be in the form of coarse feature maps. These representations encode the occurrence of critical features, such as change, but only represent their location at much reduced resolution.

Property Measurement (Signal Analysis) Certain basic measurements should also be computed prior to data reduction and abstraction. These are measures that are generic, and that need to be computed directly on sensor data. Examples include “time to impact” as well as component motion and depth.

While all these methods of data reduction are of interest to those attempting to build vision systems that work in real time, signal selection is a primary focus of active vision, and will be discussed in greater depth in subsequent sections.

At a higher level of visual processing, attention is necessary to employ limited computational resources to best advantage in the allowable time-frame. While visual and other forms of attention have been area of primary interest in cognitive neuroscience and it is known that the human system is complex, little research in this area has been done in computer vision.

Issues include: In which ways should processing be enhanced or suppressed with respect to a region of attention, which visual cues should be used to delimit the region of interest from the background, what number of regions of attention should be supported and how this should vary as viewing conditions change.

2.2 Foveal Sensing

Biological systems make extensive use of multiple resolution visual systems. In primates, there are two aspects to this space-variant architectural principle:

- At a given point in visual space (visual cortex) there are several octaves of different sized receptive fields present.

- Across the visual field, receptive fields vary by as much as 10,000 to 1 in solid angle (i.e., 100:1 in angular measure).

While most current machine vision architectures are uniform in spatial resolution, there are clear advantages to using a space-variant system. It can provide a wide angle of field with a high resolution in its center (the fovea) and a uniformly decreasing resolution towards its periphery. If we assume a sensor with space-variance roughly equivalent to the human retina, then the ratio of the sample points needed for this sensor to the sample points needed to provide uniformly high resolution is approximately 1 : 1000-10,000. A similar advantage accrues to a machine vision system which uses a space-variant architecture.

In order to realize these possibilities, several engineering and algorithmic research problems are presented to us:

- Space variant systems must be active in order to utilize the higher resolution parts of the sensor. Actuators must be provided to fixate the sensor on task specific regions of interest.

- Space variant systems must be able to effectively "boot-strap" features which fall on lower resolution areas of the sensor in order to make a rapid fixation to the center (fovea).

- Image processing and pattern recognition methods appropriate to space variant sensors must be developed. Even a simple act such as convolution is "different" in the space-variant context.

In summary, space-variant active vision is characterized by providing outstanding possibilities for small, cheap and high performance real-time vision. There is a very large industrial niche for machine vision systems of this type: If they are small and cheap, they can be useful even given the current mediocre state of machine vision! But, these advantages are acquired at the cost of developing actuator technologies, sensors, and control; and next-look, image processing, and classification methods appropriate to this specialized image architecture. Solutions to these problems are discussed in the sections that follow. It should be noted that while foveal sensors require mobile sensors and gaze control, these technologies promise to be valuable even with standard, uniform-resolution cameras.

The Log-Polar Transformation The output of a space variant sensor characterized by an exponential increase of the sampling distance (the distance between neighboring photoreceptors) with eccentricity takes the form of a Cartesian log-polar representation. This representation has a number of potentially interesting features. First of all, the mapping between the image projected on the sensor and the log-polar output of the sensor is conformal which means that all the local geometrical properties are preserved (e.g. angles) or, in other words, that all local visual processing (i.e. filtering, edge detection, optical flow etc.) are not affected by the mapping.

The second property is that linear scalings and rotations are transformed into shifts in the log-polar plane if the origin of the polar plane does not move (the mapping is shape-invariant). Behaviorally this means that if the sensor gets closer and closer to an object while fixating it, the shape of the image of the object does not change. Moreover the amount of information increases only near the fixation point (if the image of the object gets bigger it will fall over a lower resolution part of the sensor and therefore does not increase the overall amount of information).

A further, even more important point, is related to the processing of dynamic information and in particular to a kind of measure which is directly related to motor control: Time to impact. It has been demonstrated that the time to impact of a point in the environment only affects the radial component of the optical flow. This means that the time to crash computation becomes a 1-D process because one needs to compute only the component of the velocity along the log(r) axis of the representation (rate of dilation). This result holds, in the general case, if the sensor is actively tracking a point of the environment (or the object approaching the sensor).

Research Questions As far as the log-polar representation is concerned, some peculiar aspects have to be investigated. First, given the fact that the representation is shift-variant (the images of the same object in different positions are geometrically different) it will be important to study how to "merge" the information collected with different fixation points. Even the case of simple geometric features like "linearity" or "circularity" needs to be studied in log-polar space in order to derive invariants (a straight line is represented as a sort of parabola in log-polar space).

A different aspect is related to the computation of dynamic features in their different components in space (radial vs. angular) and/or time (smooth motion vs. irregular motion, object motion vs. ego-motion etc.).

2.3 Gaze Control

Gaze control in its most general form is the alteration of imaging parameters to aid in the performance of visual tasks. These include the six degrees of freedom for camera position, lens parameters such as aperture, focus, and zoom, plus relative parameter values for multicamera systems. For some applications, rates of change (such as camera velocity or acceleration) may also be important.

The primary goal of gaze control is active manipulation of the imaging system in order to acquire images that are directly suited to the tasks being performed. In an abstract sense, the ability to control imaging parameters permits an autonomous system to select a working domain within the multidimensional space of these parameters. This paradigm can lead to the simplification of many visual tasks.

The problem of gaze control can be partitioned in several ways. A common approach is to divide the problem up according to the biological mechanisms that drive eye movements: Vergence, smooth pursuit, vestibulo-ocular reflex, etc. If a computational system is based on such biological models, this leads to interesting, unsolved problems which concern the integration and coordination of these modules. In order to simplify the discussion, we consider a more abstract partition in which gaze control is divided into two primary categories: Gaze stabilization and gaze change. The former, known as "fixation," consists of controlling the available degrees of freedom to maintain clear images of some world point that may be stationary or in motion with respect to the camera. For moving targets, this typically involves target tracking. The latter category - gaze change - is perhaps more involved: In general, gaze changes may be directly motivated by the need to reduce the computational complexity of visual tasks. These changes may be used to transfer stabilized gaze to new fixation points, similar to human saccadic eye movements, to assist in solving low-level or higher-level visual tasks.

For fixating systems, gaze control is closely linked to the problem of selecting new fixation points; this problem ("where to look next") is very complex and may depend on many different levels of processing. It is our view that low-level motivations for gaze changes (such as those involving scene geometry or motion) should be considered as integral to the study of gaze control. Higher-level mechanisms (which may involve visual search or object recognition, for example) are discussed elsewhere in this report. As an example of the interdependence of these problems, consider that the difficulty of tracking a target varies with the type of target. A high-level process might designate a general area on an object to fixate, without having any preference as to the precise fixation point. Thus the exact choice of fixation points could be partly dictated by the specific needs of the gaze control system.

Gaze control research deals with the benefits of controlling gaze as well as the mechanisms required. These issues, as well as performance metrics, are described in the following sections. In each case, we will concentrate on the interactions of gaze control with low-level vision.

Motivation

Vision in an unrestricted environment must involve the control of imaging parameters. Imagine, for example, a person trying to thread a needle or pilot a vehicle without making any head or eye movements. With only "passive" imaging, these and many other tasks are extremely difficult. In contrast, an active observer will seek an advantageous configuration of camera parameters for a given situation. The ability to control imaging parameters in such a manner facilitates the close interaction of perception and action that is necessary for autonomous interaction with the environment.

For practical, real-world situations, the problems of gaze stabilization and gaze change can only be accomplished using active control mechanisms. These capabilities are important even within static environments. In order to perform 3-D reconstruction of a general scene, for example, the cameras must be moved so that different portions of the scene come into view and lens parameters must be tuned so that sharp, focused images are acquired with appropriate brightness values. For moving objects, the need to control gaze is even more compelling. The ability to stabilize the image of a moving object may be not only necessary for robust processing, but also can provide significant computational advantages over the passive case.

The following list describes several of the advantages of gaze control for both static and dynamic situations.

1. Image stabilization. For an object that moves relative to a camera, translational image blur may result unless the object is tracked. Because tracking prevents the object's image from translating, fixation permits simpler schemes for control and for processing of image sequences. Furthermore, tracking objects of interest keeps them in the field of view for a longer period of time, permitting higher-resolution cameras and additional analysis of object properties.

2. Overcome limited field of view. Any given camera system provides only a limited field of view. Gaze control is needed so that new portions of the scene can be brought into the image, and camera movements are often needed to overcome problems of occlusion. A special case of limited field of view involves sensors that have a central high-resolution region (fovea) and a lower-resolution periphery. Although the overall field of view is very wide, the foveal region is typically very small and must be aimed at locations of interest. Without the ability to fixate, these sensors are severely limited. To state this another way, gaze control makes possible all of the advantages offered by foveal sensors.

3. Reduce computational complexity. With an active vision system, it is possible to shift the operating point of the system so that imaging parameters are well matched to the object being fixated. For example, when an object lies near the optical axis of a camera, the image may be closely approximated by orthogonal projection. This simplifies many computations. An active vision system can also manipulate the level of detail in an image by moving a camera or changing focal length to control the level of detail present in the image.

4. Figure / background separation. If a moving object is tracked, the optical flow in the image for that object is reduced to very small values relative to the rest of the scene. The large differences in the flow fields can be used to segment the image, so that the object being tracked is extracted from the background.

5. Range calculations. If a stationary scene point is fixated by a camera mounted on a moving vehicle, the image flow for that point will have zero magnitude, allowing the use of gradient-based flow extraction methods. Directionality in the rest of the flow field gives relative range from the fixated point. Range data fusion is also possible; for example, a scene point predicted by stereo disparity can be examined by directing camera gaze toward the predicted location. Focus or vergence information can then be used to accept, reject, or improve the position estimate.

6. Depth-of-field control. A facility that permits control of aperture, zoom, and shutter speed (or lighting) can be used to manipulate the depth of field. Large depth of field is well suited for target acquisition, exploration, and stereo ranging; shallow depth of field increases the accuracy of focus-based ranging, and can assist in figure / ground segmentation through the selective control of focus blur.

Research Issues

In order to realize the advantages offered by gaze control, an active vision system needs several capabilities, as described in this section. Because active vision is a relatively new research area, many of these capabilities are open research issues. We first consider gaze change, and then issues of gaze stabilization.

Issues in Gaze Change As discussed earlier, part of the motivation for active vision comes from the desire to automatically and dynamically identify the changing requirements of a vision task, and then to accommodate those requirements through the active manipulation of imaging parameters. This manipulation may be characterized with regard to (i) permissible configurations of imaging parameters, and (ii) integrated use of the resulting visual information. Mechanisms for gaze change therefore facilitate the dynamic integration of different forms of visual information.

The following list contains several areas in which research is needed:

1. Dynamic integration of visual cues. Active vision seeks to exploit the relationship between actuation, sensing, and control. As a logical consequence of manipulating several imaging parameters independently, there is a need to develop schemes for optimally integrating the different sources of visual information that result. For example, different independent depth cues (including image blur, stereo disparity, and motion disparity) need to be combined in a consistent manner. These image features depend directly on camera vergence, lens focus, etc., and each of these cues may be more reliable for different depth ranges. For example, focus-based ranging is most reliable for nearby objects, and other cues such as camera motion or stereo disparity may be more useful for greater distances. Research is needed to develop new models for the control, integration, and resolution of conflict for these visual sources. Some of this research may be motivated by biological mechanisms.

2. Strategies for interleaving image acquisition and visual processing. For any visual task of moderate complexity, many different combinations of imaging parameters may need to be employed in order to examine different points in the scene. This implies that further study is needed in developing strategies for acquiring images and processing them. The resulting methods must depend on the information content of the images obtained and on the computational complexity of the visual task.

3. Computational complexity and resource allocation. How should the gaze selection be done so as to minimize total computational effort spent in analyzing a given scene? For any given fixation point, new locations need to be selected based on low-level criteria for the optimal exploration of unknown scene regions. There may be tradeoffs between selecting new gaze parameters compared with the reallocation of computational resources. Research is needed to exploit cooperation among processing and control tasks to provide more efficient and effective methods for gaze control.

4. Low-resolution vs. high-resolution information. Information at different levels of resolution may be available, and these should directly affect the strategies for controlling imaging parameters. For example, multiresolution sensors may require a significant restructuring of gaze-control strategies. Focus control can also lead to image information at different resolutions. Traditionally, the lack of perfect focus has been considered a deficiency of the imaging system. However, the ability to control the degree of image resolution may be used in some cases to increase robustness of the vision system.

Issues in Gaze Stabilization Gaze stabilization, like any control problem, consists of an error estimator, a plant, and a control law that tries to drive the error to zero. As such, this research builds on decades of work in traditional control theory. Gaze stabilization involves advanced control theory topics, such as nonlinear, optimal, multi-input, multi-output, and hierarchical control. There exist many unsolved problems for gaze stabilization that depend fundamentally on visual processing. Some of these directions are current topics in control theory and others seem unique to visual tasks. The following list of open research issues is not necessarily complete but indicates how gaze stabilization involves nonstandard aspects of control theory.

1. Extraction of the error signal. The error signal for gaze stabilization (or for fast gaze-changing operations) must be derived from image sequences using computer vision techniques. Gaze stabilization therefore must exploit progress in dynamic computer vision algorithms. Segmentation (target extraction, figure/ground differentiation) is a basic problem. Further, models of noise, sensitivity, stability, and performance must be appropriate for sampled spatio-temporal signals, perhaps at multiple scales of spatial resolution.

2. Gaze stabilization is a hierarchical, multi-input, multi-output, multi-rate problem. Several interdependent, nested layers of control are needed to deal with the many different types and levels of visual tasks. For example, simple centroid tracking may involve only low-level control, but tracking a person who moves through a crowd also involves processes at higher levels. Gaze control can involve inputs from different vision algorithms (eg. optic flow, focus blur, position estimators) and different sensors (eg. accelerometers, position encoders, images, efferent copies of control signals). The output must drive multiple cooperating effectors governing intrinsic and extrinsic camera parameters. Inputs often arise from sensors that have different sampling rates or different computational bottlenecks.

3. Dynamic Creation and Adaptation of Domain and Plant models. To achieve performance robust against noise, dropout, and delays, predictive capabilities arising from the creation and exploitation of models of the plant, the signal, and of the feedback control process are required. Adaptation occurs at several levels and at several timescales. For example, acquiring a moving target and tracking it require two different types of error signals: To acquire the target, image flow must be detected and fixated; to track the same target, image flow now represents background and is ignored. Changing target behavior (maneuvering, for instance) affects control algorithms. Control should also adapt to changes in knowledge about target or plant acquired through time.

Metrics for Gaze Control Evaluation

The performance of systems that implement gaze control may be evaluated from two points of view. The first is based on traditional domain-independent approaches. However, gaze control is not an end in itself, but must ultimately support applications. Its efficacy should therefore be measured by the amount by which the performance of the associated application is improved.

Still, it is useful to consider several traditional domain-independent metrics to characterize control performance, such as frequency response and error sensitivity. For fast gaze shifts, the following indications seem most useful:

– rise time

– overshoot / accuracy

Research in physiological vision could also benefit from the development of additional metrics in gaze stabilization. For example, biologists are often unable to characterize the quality of eye-movement performance by traditional control measures. Part of the reason for this is the sophistication of biological gaze control systems; humans are extremely adept at exploiting predictability in a target stimulus (such as sinusoidal motion). Consequently, biologists sometimes describe the quality of smooth pursuit by the number of catch-up saccades used by the subject in a given period of time.

In general, non-traditional metrics can include performance measures for tasks that are not possible without gaze control. These include the following:

– depth recovery for very wide visual fields and over large ranges of depth

– motion estimation over large ranges of velocity

Such non-traditional metrics can also be used to quantify the improvement in performance of visual tasks using active vision, which, although feasible, are done less efficiently using passive techniques. Some examples of this are improvements in

– estimation of motion parameters and surface reconstruction for moving objects

– accuracy for depth estimation, tracking, etc.

– segmentation of dynamic scenes

Conclusion and Recommendations

Active vision entails automatic, context-dependent manipulation of imaging parameters. Such control is needed for the performance of many real-world tasks, and improves the reliability of performance through the integration of different sources of visual information. As a consequence of using these different sources, there is a need for optimally and dynamically combining the strengths of different visual cues. Research in this area is needed to characterize the performance of a system that can control gaze changes, and must involve both theoretical and empirical analysis of the parameters involved.

For gaze stabilization, reasonable proposals have begun to emerge, but stabilization and tracking continue to be very difficult problems. Most current systems rely on very simple image features such as region centroids; these are rarely accurate or robust except for very specialized applications. Better stabilization methods should be a high priority for research.

At a higher level, gaze stabilization must be integrated with methods of control changes in gaze. When scenes are dynamic, containing independently moving objects, the gaze control system must dynamically switch between objects and concurrently analyze the object areas. Research is needed that directly addresses strategies for integrating gaze change with gaze stabilization.

2.4 Where to Look Next

Real time vision requires a system that purposefully gathers information from the visual world. Information gathering is a dynamic process that responds at once to events in the visual world, to the system's evolving understanding of that world, and to changing requirements of the vision task.

Vision must be understood as a spatio-temporal process. This is true, first of all, because events in the world are distributed in time as well as space. It is also true because cost and complexity considerations require that system resources be focused on restricted regions of the scene. Hence a sequence of such focal probes is needed to explore the visual world. Finally, it is true because sequential processing provides for high efficiency through directed analysis: Results of each step direct subsequent steps to the most important regions of a scene.

One way to interpret the need for sequential, problem-dependent eye movements is as a suggestion that the general problem of associating many models to many parts of the image simultaneously is too hard. In order to make it computationally tractable within a single fixation, it has to be simplified, either into one of location (one internal model) or identification (one world object). Table 2.1 summarizes this view. This table suggests that a visual task can be solved by partitioning the task into two different kinds of subtasks[1]. We model the visual organization more crudely into a center and a surround. The center is *where-I'm-looking* and the surround is a source of new gaze points. A location task is to find the image coordinates of a single model in the presence of many alternatives. In this task the image periphery must be searched; one can assume that the model has been chosen a priori. An identification task is to associate the foveated part of the image with one of many possible models. In this task one can assume that the location of the material to be established is at the fixation point (see Table 2.1).

In order for the suggestions made in Table 2.1 to be reasonable, there has to

[1] A major feature of the gross organization of the primate visual brain is the specialization of the temporal and parietal lobes of visual cortex into WHAT/WHERE functions. The parietal cortex seems to be subserving the management of locations in space, whereas the temporal cortex seems to be subserving the identification of objects in the case where location is not the issue.

		Object to Match Against	
		One	Many
Image Portions	One		Identification: Trying to identify an object whose location can be fixated
	Many	Location: Trying to find a known object	Image interpretation: Too hard?

Table 2.1: The biological organization of cortex into WHAT/WHERE modules may have a basis in computational complexity. Trying to match a large number of image segments to a large number of models at once may be too difficult.

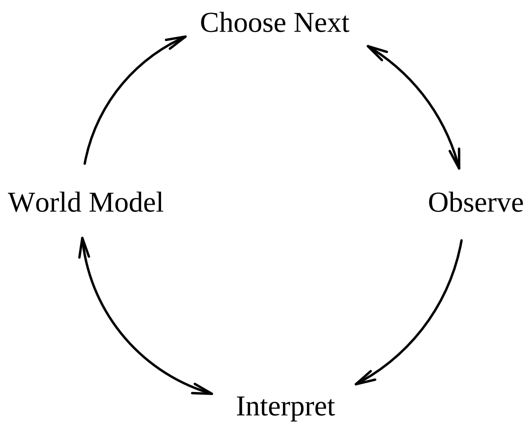

Figure 2.1: The cycle of perception.

be some way of changing gaze to fixate a specific object. One difficulty is that it is unreasonable to assume that the location of the object is known precisely, at least on the initial gaze change. Another is that whatever visual mechanism is posited, it must only require very low spatial resolution, since just prior to the gaze change, the target object is typically at the periphery of the visual field. Features that work under these circumstances include motion and color.

A System Model A system that addresses the problems characterized in the previous section can be formulated in terms of four functional modules, as shown in Figure 2.1: Observe, Interpret, World Model, and Choose Next.

There are two inputs to the system: The sensors (cameras) provide an input to the Observe module[2]. The task requirements are provided by a larger external system to the Choose Next module. There are two outputs from the system: The

[2]The Observe module can also direct attention within an image or other internal representation.

Observe module directs the sensors and results of the vision process are provided to the larger system from the World Model module. Within the system information flows in a continuous feedback loop. The primary direction of flow is clockwise.

In addition there are two feedback paths: Status information from the World Model to the Interpret module, and alerting/orienting (abstracted signal information) from the Observe module to the Choose Next module. The selected task from the Choose Next module is fed to the Interpret module to provide context for interpretation.

According to this model real time vision in a changing world can be organized as an unending sequence of observation steps. Each step begins with the Choose Next module. Based on the current task requirements, the current world model, and alerting/orienting signals (from the Observe module) the system decides where to direct its next observation, where to "look next". A corresponding request is sent to the observation module. An observation is a measurement performed on image data. The Observation module directs the sensor and performs required early processing to obtain this measurement. Results of the measurement (pattern data, estimated motion, depth, etc.) is passed to the Interpret module. There it is understood in terms of the system's current world model. Finally the interpretation results are passed to the World Model module where they are used to update that module. The updated model is made available to processes outside the vision system that require results. This completes one step of an observation sequence. On completion the cycle begins again.

Three Inputs to Attention Within the proposed framework "Next Look" can be:

- Task driven
- Feature driven
- Context driven

Task driven control of visual attention can be performed independently of the visual processing. For example during walking visual attention shifts between a distant point (to control the general direction) and a point just in front and on the ground (to detect obstacles). If "nothing strange" happens this strategy does not depend on the results of visual processing. The selection of the fixation point can be done merely on the basis of geometric measures (e.g. the angle between the optical axis and the ground plane) and does not vary with the visual appearance of the environment.

Feature related control of visual attention, on the other hand, is based on the result of local visual processing (features). An important distinction can be made between static and dynamic features. Static features can be, for example, regions with a peculiar color, or with a very high contrast or with peculiar geometric structures (corners, blobs, etc.). Dynamic features are, for example, regions of the images with a very high (or anomalous) velocity. This second strategy is based upon very simple (meaning "low-level") visual processing and may have the only purpose of selecting where to focus or verge while "wandering around" or to react to unexpected events or, finally, to track a moving "target" to stabilize its image on the fovea.

Context related strategies are those based upon the results of "high level" processing (e.g. if this is a square I should find a corner at a specific point in space). These strategies are those described in the famous works of Yarbus and are usually driven by recognition purposes. In general the major difference is the fact that the choice of the next point to look depends on the result of the visual processing performed before and on the position of the current fixation point.

Implementing Next Look One model for implementing a next look module seems to have been independently developed by a number of different labs. It is based on the notion of an "interest map". In one of the existing implementations, the interest map is taken to be embedded in a 2-dimensional manifold of eye position coordinates, and the "interest" projection is provided from a high-dimensional feature space to the 2-D interest, or "motor" space. Peaks in the interest map determine the next eye position, and familiar pattern recognition and image processing techniques (Parzen windows, lateral inhibition, winner-take-all, etc) can be applied to these interest maps. This allows familiar hardware, and software, from other image processing applications, to be applied to the next look problem.

This model is reminiscent of the vertebrate superior colliculus, in which electrical stimulation of a location on the surface of the colliculus causes an eye movement to the position in space corresponding to the topographic map of space represented in the colliculus. The colliculus is thought to control eye position by some form of spatial averaging of "featural" activity in the colliculus. Thus, although some of us have not specifically sought to mimic colliculus function, while others have, there seems to be a natural convergence towards the type of solution which is likely to operate in the vertebrate mid-brain.

Research Questions

Competing Tasks $\Rightarrow$ (scheduling problem) A useful analogy to make in describing an attentional system is that of a computer operating system. The operating system serves to allocate scarce resources. Likewise the attention mechanism in vision allocates the scarce resources of the fovea and allied computing machinery to multiple visual tasks. These tasks may have conflicting requirements as in the case where different tasks would direct the fovea to different points in visual space.

Shifts in visual attention can be thought of as arising from a scheduling system which is giving control of the position of the fovea over to different tasks. The scheduler determines whether a currently active visual task has been completed or whether an alarm or interrupt (from a peripheral process, for example) has occurred and decide which visual task to process next. The activity of the scheduler will depend on the current state of the system, the current estimate of the state of the world, and higher level goals or priorities.

Current active vision systems have concentrated on demonstrating essential components of the system or individual behaviors. As systems with multiple behaviors are built, the scheduling problem will become increasingly paramount. The operating system paradigm should not be thought of as limited to visual tasks. A general robotic system would have multi-tasking to handle conflicting requests for its limited motor resources. The active vision scheduler could form a part of the more general system (see section 2.6).

Problem-Dependent Features ⇒ (learning problem) An early goal in computational models of vision was to reconstruct the relationship between the photometric data in the image and internal models with sufficient fidelity to support cognitive requirements. However, the problem is that the cognitive demands of a system tend to be special purpose and demand a tailored processing of the image. For this reason the image cannot be effectively analyzed independent of its goals. Work in active vision spans a wide variety of processing styles, but it may be that the processing of an image is task dependent in the extreme. That could mean that individual frames or saccades have to be interpreted at run-time without recourse to extensive internal models.

If general purpose processing is to be abandoned for special purpose processing, an important research question is the automated discovery of the structure of special purpose behaviors. One promising route is to develop learning algorithms that can discover appropriate special purpose features through the presentation of examples. Work on learning algorithms has been given an important boost with recent developments under the heading "neural networks" and we expect these developments to be very useful here.

Metric For Evaluating Attention Models Attention models can be based on i) stimulus response, ii) models of structures in the environment (context), and iii) task dependencies. In each of these cases a metric is required to evaluate the different strategies involved. A number of different cost functions are possible, but they usually fall into on of the following classes: Time required (evaluated over task duration), accuracy, and robustness.

It would be desirable to be able to formulate metrics independent of specific tasks. For example, certain features are more reliable than others (e.g. corners in edge-based tracking); entropy measures can be formulated given specific models independently of how they are employed; and so on. However, the study of canonical tasks is likely to play an important role into gaining the necessary insight into what constitutes a good metric.

Feature Extraction As discussed above, a number of different groups have arrived at a common architecture for the final stage of attention, namely that part that converts information about an object's visual features into motor commands. Owing to its similarity with biological structures, we have called this the superior colliculus model. This model isolates a portion of the image in retinal coordinates and interprets the result as a motor error signal. The portion of the image is isolated by projecting the location of relevant features onto the retinal coordinates.

Within this common architecture, the central problem is the projection process, which selects the features that govern the projection. To understand the questions involved, consider just the aspect of this problem related to resolution. In a spatially-variant sensor the features at the periphery are represented at reduced resolution, yet they must have sufficient fidelity to direct the sensor. If everything about the peripheral image was known, then there would be no reason to look there. On the other hand if the resolution is reduced then the essential components of it that are associated with an object may be lost. Thus a very important research question is the determination of visual features which can be informative at low resolution.

Minimizing Uncertainty As described earlier, "Next Look" can be task, feature, or context driven. One aspect of context is related to the role of active vision in minimizing the uncertainty of interpretation at different levels of representation. Consider the case of recognition. A view of a scene might not provide sufficient information from which to infer the identity of an object. In this case a new vantage point (Next Look) needs to be chosen such that additional evidence can be gained from which to make a more certain interpretation. The process proceeds in an iterative fashion until the level of confidence is such that a final interpretation can be made.

This scenario has a number of implications: Uncertainty must be represented and reasoned about explicitly, enabling the computation of Next Look from these representations together with a criteria for knowing when a valid interpretation can be drawn. These implications result in a number of important tasks for active vision research. Representations for uncertainty and algorithms for reasoning explicitly about uncertainty need to be developed. Traditional approaches, e.g., confidence intervals in parameter spaces, do not always have an intuitive meaning in the visual world. This becomes particularly important when reasoning about how to compute Next Look. Representations for uncertainty need relevance to the visual world so that perceptual strategies can be derived for minimizing uncertainty. Finally, there has to be a way of knowing when to stop, i.e., when the level of uncertainty is low enough to make some inference or assessment of the scene. This is presently an open question.

Data Representation / Features The issues of data representation and features useful for active vision need to be studied together. Particular emphasis should be put on those global features which may be useful to control the position of the fovea or to guide the motion of the sensor (behavioral features). Examples are the computation of the focus of expansion, the estimation of the aiming direction during navigation, and the information necessary to control vergence. A particularly critical issue is tracking, either the tracking of a moving object or the tracking of a static object during motion (or both). The peculiarities of these features and their feature extraction process in relation to space-variant vision need to be studied to demonstrate which approach (uniform vs. non-uniform sensors) is best suited to the active vision paradigm.

As far as representations are concerned, it would be important to study which are the best and simplest representations in terms of behavior control (i.e., how to represent "behavioral features" in a useful and simple way). For example how far can you get by using only disparity and not depth (the premium by using disparity is that camera calibration becomes much easier and, to some extent, not necessary). The same issue applies to time to crash as opposed to depth. One probably does not need to compute depth to avoid obstacles. To control a moving vehicle it may be sufficient to use disparity maps and time-to-impact maps (i.e. features you can derive directly from the incoming images).

2.5 Eye-Hand Coordination

This section discusses active vision issues specifically related to tasks that exhibit coordinated visual and motor action, particularly the control of manipulators and robot hands.

Motivation

Vision is incomplete in two ways. First, as a sense, it can only detect geometric and reflective information, imperfectly. For example, properties such as elasticity and functionality cannot be sensed directly by vision. Secondly, vision alone lacks well-specified performance criteria. Researchers have had to evaluate the performance of their visual algorithms by how the results "look". The robotic hand offers a solution to both problems. The sensors on the hand augments vision with additional information. Manipulation behavior also provides a performance metric for the vision system and geometric feedback.

Manipulators are traditionally thought of as effectors or actors, but manipulators can be used to perceive as well. Hands are often equipped with tactile and force sensors that can give information about the shape and weight of the object. Furthermore, through the kinematics of the manipulator, we can get positional information. We can also imagine equipping manipulators with temperature or other sensors that may be suited to a particular task. We shall refer to the these senses as hand senses.

In short, a hand is both a perceiving and an acting thing that both augments vision and closes the perception/action loop.

Research Issues

Eye-hand systems offer the unique ability to determine material, geometric, and haptic characteristics of objects that are inaccessible to vision systems alone. We can also distinguish between liquids or solids and measure the viscosity of liquids and measure the weight of solids. We can verify visual positional measurements using the kinematics of the manipulator and hand. We can also determine if an object is rigid, deformable, attached, movable, and removable. To explore these characteristics, we need representations and strategies for integrating visual information with hand sensory information. In this section we explore the representations, strategies, and learning of representations and strategies, and calibration as research issues in eye-hand systems.

Representations

Representations are dependent on the capabilities of the agent and should be dependent on the task to be performed. Traditionally artificial intelligence (AI) representations have ignored the sensors or manipulators that are going to be used for the task. Representations should be developed to meet perception/action goals. Specifically, we need to explicitly represent the agent performing the task (including the manipulator, hand, and sensors), the environment, as well as the task. Moreover, the different types of representations have to be integrated to efficiently and effectively perform the task.

For example, object representations used by many vision researchers have no explicit representation of the functionality or the elasticity of the object. This information cannot be tested or derived solely from visual information. Therefore there has been little need in standard computer vision to represent these characteristics. These features are only accessible through interaction with the world and need to be modeled explicitly by a eye-hand system. The functionality of the object is one feature that is clearly suited to the task to be performed.

Since both the visual and manipulation systems contain intrinsic uncertainties, the explicit representation of uncertainty will be important in eye-hand systems. For example, the manipulation task can provide a rationale for deciding when the vision system has "too much" uncertainty. This relates to the where-to-look-next problem as well.

As we build new representations that can model a richer set of object characteristics we must ask "what are the implicit assumptions?", and "how can we trade-off accuracies between the different representations?". What are the data reduction mechanisms? Consider the trade-offs between accuracies of representations: How accurate must vision be, what is the manipulation accuracy required by the task, and what resolution of information can we get from the hand sensors?

Strategies for Hand-Eye Sensing and Manipulation

One strategy is to use vision to make predictions about the world and have the hand verify. This is most useful for redundant geometric information or sensing positioning strategies. For example, the vision system could guess about the shape and position of an object and the manipulation can verify by testing point contacts. For determining full three-dimensional object models or resolving recognition ambiguities, the manipulation system can be used to move objects that the vision system can look at all faces. In this case, the vision system is controlling the manipulator to meet visual perception goals. Positioning the object with respect to the vision system is also important for occlusions and concavities. In this case, the manipulation can further help by moving object away from occlusions or testing hypotheses concerning concavities. These features of objects would be impossible to determine with a vision system alone.

When both the vision and the hand system are being used for perception, it would be better that they work together. For example, for determining the rigidity or elasticity of objects, the manipulator can squeeze the object under the close monitoring of the vision system to determine if the shape of the object is changing. In assembly tasks, the demands of the manipulation task may specify to the vision system the viewing position that aids most in the manipulation task.

With either of these strategies, robust error recovery strategies are required. What would the system do if it dropped the object? What if too much force builds up while testing deformability? Errors can occur in all types of imaginable and unimaginable situations and the system must have some idea when it has succeeded in its task and what is should do if it fails. It should also know when to abandon the current task to proceed to the next task.

There are many design considerations in matching visual and actuating systems. One of the major problems in this area is the mismatch of effective sampling rates. Most manipulators and hands run at servo rates around 1KHz. On the other hand, most vision systems run at 60Hz. This can cause serious control problems. Faster image processing can help to support better coordination between mechanical positioning and visual inputs.

Learning

Learning is important to eye-hand systems. Object properties can be learned, perceptual strategies for eye-hand coordination can be learned, and the functionality of objects can be learned, as well.

First we must understand what knowledge is innate and what is learned. Innate knowledge or strategies are those that are implicitly encoded in the rules or representations of the system. The agent can only explicitly learn information or strategies that the representations permit. Biological or psychological experiments can provide a model as to what animals or humans innately know and what they learn. It is essential to choose representations meaningful for the information to be learned.

Problem solving strategies for hand-eye coordination can be learned. Picking an object up and turning it around can be extremely useful for determining a three-dimensional model of objects. Learning strategies for assembly by first disassembling are possible with eye-hand systems.

Learning about the functionality of objects requires a eye-hand agent. For example, the motion of a hinge can not be inferred from vision alone (unless one has a huge database of specific object models). The motion of a hinge can be learned by experimenting with the object and noticing changes in shape by the visual system. Deformability of objects must be tested by a manipulation system perhaps after being hypothesized by a visual system.

Calibration

We must build systems that function robustly without accurate calibration. This is particularly important for eye-hand systems, since both subsystems have distinct geometric parameters and the subsystems must work together. To achieve the calibration that is required, it would be helpful if the system could self-calibrate and adapt the calibration as the system changes. This becomes even more critical since the complexity of calibration for an adaptive vision system has a much higher dimension because there are more camera parameters being actively controlled. An eye-hand system offers a unique potential for self-calibration, by sending motion instructions to the manipulator and watching for the visual correspondence for motions. The system can be further calibrated for all variable camera parameters. Also there should be an adaptive calibration mechanism that continually test the expected visual motion with the manipulator and camera parameters. The calibration could then be altered slowly as conditions change.

Lab Requirements

Eye-hand research requires a significant investment in a laboratory; expensive equipment and substantial human resources are required over extended time periods.

At present, we see the single biggest obstacle to progress in this area as the cost of hand systems with significant capabilities.

Research in eye-hand systems requires laboratory equipment and personnel. The minimum equipment required for eye-hand research includes cameras, computers, manipulator(s) or hand systems, and interface devices (digitizing boards and motor control boards). The computer and controller boards must be fast enough to perform real-time computations. While not requiring that everyone have an articulate "human-like" hand, we suggest a minimum of a single finger (no joints) probe with tactile and force feedback. The development and integration effort necessary to build eye-hand systems is enormous. Typical personnel requirements are at least one professor and three graduate students with support for at least 3 years.

To achieve eye-hand coordination in real-time requires both a real-time vision system and a real-time control system for a manipulator. It is more difficult to finesse the requirement for real-time vision in an eye-hand coordinated task. This suggests that a full-fledged frame-rate vision system is essential for eye-hand tasks.

Robot manipulator systems are complex and costly as well. Simple manipulators with simple grippers primitive control with position, contact, and force feedback are the minimum that we recommend. Representative of the other extreme currently might be the MIT/Utah hand and supporting hardware and software. In the intermediate ground are manipulators with more capable control systems and force feedback, such as the Zebra or Puma arm.

We would also like to see a cost effective hand system be developed so that it becomes more affordable for many laboratories. Most labs do not have hands let alone perform research in eye-hand coordination. Future research in this area will be severely limited unless a cost effective hand system can be developed.

Conclusion

The advantages of eye-hand systems are that hand sensors complement the visual sense and that the system is closed loop and has a natural performance criterion. We feel that research should focus on the representations, strategies, learning, and calibration of eye-hand systems. We cannot learn appropriate internal representations of objects, materials, contexts, environments, and goals without a hand-eye system. While hand-eye research is expensive, we have outlined the basic equipment and personnel requirements needed to perform this research. We feel that this area of active vision is important and provides the biggest possibility of advancement due to the wide variety of tasks that this task-driven research can provide.

2.6 Integration with Robot Architectures

Robotics can be viewed as the intelligent connection of perception and action. One view of a robot architecture is a programming framework for expressing perception and action routines, and the algorithms, data structures and techniques/methodology for combining them in a robot agent to achieve a task. Active vision can be viewed as a subset of these routines used to exploit vision in the carrying out of robot tasks.

This description implies that active vision research should be conducted in the context of ongoing research in robot architectures. This is important for two reasons. First, many if not all of the relevant issues are being addressed by people doing research in robot architectures, and active vision investigators can benefit from these experiences. Secondly, researchers developing visual routines, need to maintain a view toward the purposes for which robot architectures are being investigated.

The primary idea in integrating active vision with robot architectures is that perception is used to support robot tasks. In any real robot system, sensing resources will be limited and will typically have to be deployed selectively to answer questions only after they have been determined to be relevant. This is in keeping with the active vision tenet that vision needs to be directed, focused, and told what to look for. Integration questions of both a hardware and software nature will revolve around the inherent limitations on the amount of sensing that can be done at any one time, and the visual routines that are available to carry out that sensing.

Hardware Architectures Much real-time hardware is currently available for achieving active vision experiments. These include real-time image processing equipment (e.g., PIPE, Datacube), real-time general purpose processors (e.g., VME multiprocessors), and special purpose hardware for specific algorithms (e.g., Teleos' PRISM). These kinds of vision systems draw large currents ($\sim$ 100 amps), which most robot platforms (e.g., Cybermotion, Denning, TRC, etc.) aren't designed to provide. As well, the addition of an active vision system to a robot hardware environment greatly increases the weight and therefore the mechanical capability required to traverse an environment.

But specialized vision hardware is still in its infancy and we believe that power consumption and size will decrease dramatically with a small increase in demand. We take as evidence for this view the new release of the MaxVideo 20 board by DataCube which incorporates most of the functionality of about 8 of its other boards at 1/8 the size and power consumption.

Research into algorithms for active vision should not necessarily follow currently available real-time hardware, because that hardware can drastically limit the kinds of algorithms that might be developed. Instead, we need to develop algorithms that can be potentially executed in real-time, and use today's hardware to implement those algorithms that happen to fit well, or a simpler version of those algorithms which do not execute well on today's hardware.

Software Architectures There seem to be two widely differing approaches to intelligent agents architectures – the historical, disembodied, deliberative approach and the newer, imbedded, situated approach. The former is characterized by a commitment to a world model which captures all or many levels of information all the time from retinotopic images to three-dimensional objects in time. As well, they tend to represent all aspects of the world and the task being carried out in a propositional form. Finally, they tend to carry out all actions through a centralized, deliberating (e.g., deductive) control system which is known to be computationally intensive.

As researchers began to realize that the deliberate approach produced robot plans which failed in field environments, a number of new approaches to robot control emerged under the general heading of situated reasoning. Situated reasoning stresses a tight coupling of sensing and action to deal with changing environments and data uncertainty. The situated approach is characterized by layered levels of perception-action behaviors that execute in parallel while arbitrating conflicting motor commands (be they to wheels or pan-tilt heads), or demands on limited sensing resources at runtime. Generally only a partial world model is used, with a set of indexical representations (sometimes only implied in hardware wiring) for the objects involved with the task at hand. Much emphasis is given to creating dynamic behaviors (e.g., obstacle avoidance, tracking) in the face of a changing environment, and decision-making cycles execute in fractions of seconds. If deliberation is used, the results serve as a strategy to be followed rather than as a program to be executed.

The situated approach to intelligent robotics has produced robots which function well in the real world because these robots can cope with environments that change either naturally (moving agents) or because of the nature of the task (catching a spinning satellite). The situated approach is thus designed to include active vision as an integral part of its architectures.

Issues

A fundamental need for integrating active vision into situated robot architectures is the development of a set of visual routines which can be used as the building blocks for vision-based tasks in situated robots. Developing this vocabulary of routines (and building integrated planning/acting systems that make use of them) should be of central interest to researchers both in robot architectures and in active vision, since such a vocabulary defines what the sensing system can do. Robot planners need to tell vision people what information is needed for acting in the world and vision people need to tell planners what kind of information an active vision system can extract from the world in real-time.

What set of primitive visual operations are necessary for robot tasks in general? The actions that need to be done define the types of sensory processes that are needed. The activity in active vision can arise from directives of the sensing system or due to the task the robot is carrying out. For instance, locating the table may be a required step in a perception activity to find a coffee cup, but it also may be the result of looking for a clear space on which the robot may place a book.

We expect that there will be some specialized routines for certain ubiquitous tasks such as navigation (fixation for stabilization) and grasping (visual servoing), but others will be of more general use (e.g., gaze control, homing on an object using color cues). Some active vision routines may involve controlling the physical parameters of the sensor, e.g., tracking, and others may involve simply combining more traditional techniques, e.g., optical flow, in new ways. Thus the output from these routines may simply provide useful data structures, or instead perform a complete subtask for the robot.

An example of the use of some sets of these routines to form more complex functions would be the use of real-time stereo on a mobile robot in support of find-and-fetch tasks. A routine could be developed to track the range to the ground in front of the robot as the robot moves, building up an expectation of the ground level slope. Any large deviations that were subsequently detected would indicate dangerous topography or an object of interest. A second routine could use vergence at several points in the field of view to detect vertical surfaces (e.g., walls) that could be used for relative navigation.

There are two sub-issues which concern visual primitives and the combining of primitives into visual routines.

- What are the appropriate levels of primitive operations? This issue involves the selection of a general set of primitive operations which can be used as building blocks for more complete visual routines. It is desirable to modularize the active vision primitives in such a way as to have them be neither hardware dependent (e.g., programming the motors) nor too task specific (e.g., color histogram tracking of a plaid pattern).

  There is an underlying assumption in our discussions that there will be a certain amount of ongoing bottom-up, early vision, pre-attentive processing accommodated in the robot architecture. Visual primitives could be used to support intermediate vision, their purpose being to derive more compact information from the non-compact results of the early vision processing. Criteria for selecting these type of routines might be that they can be implemented in biologically plausible hardware, are of general use, are neurophysiologically and psychologically plausible, and are relatively straightforward to implement. Examples of such operators include markers and specialized detectors (e.g., a detector of blue horizontal edges).

- How are these visual operators best combined? This question is not simply one of finding combinations of primitives which result in routines useful for given tasks. It is also a question of combining the primitives so as to simultaneously satisfy the timing and response constraints of the task, and the often conflicting demands on the vision apparatus. One might develop a routine to extract license plate characters from moving vehicles using a space-variant sensor, only to find that it works poorly when integrated on a mobile robot platform which must simultaneously avoid intermittent moving obstacles. This kind of effort is central to the ongoing research in situated robot control architectures, from which the active vision investigator can benefit. This issue also motivates the development of teams of active vision and situated reasoning robot researchers.

What is the right combination of pre-attentive and attentive processing? The architecture that incorporates active vision could consider two modes of processing: The pre-attentive bottom-up mode, where a similar type of processing is done everywhere in the image in parallel, and the attentive task-driven purposive mode, where the current task, state of the system, and state of the world determine how to process information from the image. Note that sometimes the relevant information is obtained from the state of the world relative to the system. Generally pre-attentive processing would be followed by attentive processing.

Pre-attentive processing is necessary so that the system can deal with an uncertain, unstructured, changing environment. Low level features of the image are extracted very quickly and provide the basis for interpreting the visual scene. Attentive processing is necessary in order to limit the amount of overall processing done on the image information. Attentive processing is task-driven and purposive, and involves using knowledge of the environment gained at the previous instant in time to aid in interpreting the image at the current instance in time.

Much of the answer to this question will turn on what falls out from defining the visual routine vocabulary discussed above. Yet some pre-attentive processing may be required without being motivated by the robot's specific repertoire of visual routines.

Which coordinate system should be used to simplify achieving a given task? This issue seeks to distinguish the usefulness of geocentric, viewer (robot)-centric, and/or fixation point reference frames as a function of the task to be achieved. Related to this question is that of the use of relative versus global computations for a given task. In conventional vision (e.g., the Marr school), the motivation for the particular reference frame was the need to build a complete 3-D model of the scene. In active vision approaches, the motivations derive from the task which dictates the kind of visual routines which in turn dictate the frames of reference.

If a combination of reference frames are needed, then computational methods for efficiently mapping between these are needed. Is a set of retinotopic maps with some sort of sensor position information the best way to go? Or should one carry out the mapping from viewer to world coordinates every time a new image is obtained? Related to this is the well known problem of maintaining a stable world representation (i.e., one that does not change drastically as the observer moves).

What representations should be used to simplify achieving a given task? This issue is at the center of the paradigm-shift from classical planning of robots to the situated reasoning approach. Extreme cases range from almost no representation to a homogeneous production rule memory exist. Many middle ground approaches seek to maintain combinations of local detailed representations (e.g., metric maps) and globally more abstract representations (e.g., containment models of relationships of objects and rooms). One aspect of analyzing the usefulness of a visual operator may be how well it supports the representations needed in the reasoning parts of the robot architecture used for task achievement. This is again another area where researchers in both fields need to be working closely together.

What is the role of data fusion in these architectures? Active vision is distinguished by its use of all three forms of visual information: Pictorial, spatial image understanding using image structure and content in the static domain; sensory (photogrammetric), using aspects of the camera geometry and other imaging properties to infer structure; and kinetic, the use of movement to infer structure. Integrating across these different visual information sources may provide a richer set of possible visual cues to support robot tasks.

But there are at least two schools of thought on this issue. One school, deriving from a reconstructionist approach says that every piece of information about the environment that one obtains as the sensors move about should be used. Bad information should be weighted less than good information, so measures of the reliability of information should be a standard output from vision algorithms. The information is combined to give an increasingly more complete picture of of the world.

The other view is that active vision requires something closer to data fission. The information from a module of one form of visual information can affect what another module does (e.g., when a suddenly detected motion alters the next movement in a series of saccades). This type of data fusion called strongly coupled data fusion involves switching among sensory modules rather than fusing the data structures of the modules.

How are active vision routines best used with force-torque routines, acoustic, haptic (touch) procedures, other non-optic sensors? Most active perception research concentrates on the modalities of touch and vision. Much active touch research follows from Gibson's observation that when we touch something with our fingers, the skin stimulation depends on the movements that we make. Similarly, recent active vision research reflects Gibson's view of fixation, accommodation, convergence, and pursuit as dynamic visual activities.

Sensory modalities other than vision can benefit from an active approach. A prime example is tactile sensing, where the process is fundamentally active. Active vision researchers should look to non-visual modalities for examples of useful active strategies and routines.

Here perhaps we may also find forms of combination more akin to traditional sensor fusion. For example, in a task to pick up a white cue ball, the procedure of squeezing two white balls of equal size would add malleability information to the color information.

2.7 Benchmarks

The design and use of benchmarks for testing and comparing systems is a difficult topic. For testing active vision systems we have divided benchmarks into three groups. A good single measurement benchmark should isolate some feature of a system and accurately measure how well the system performs with respect to that feature. A simple behavior benchmark typically measures the performance of a loop that couples a single sensed feature with some form of motion. Finally, an integrated system benchmark tries to capture the performance of a more complex behavior. Balancing the need for rigorous comparison and the need to foster innovative systems requires careful attention to benchmark results and their appropriateness.

Benchmarks should not be confused with applications, which are discussed in Section 4. A benchmark is a specific task designed to measure an active vision/robotic systems performance and thus enable us to compare different systems. Benchmarks should be applicable to many systems designed for different things and should be shared by many researchers. An application is an area or task in which active vision might be useful; every researcher is likely to have his or her own favorite. We feel that the task of developing meaningful benchmarks applicable to many different types of active vision system is itself a difficult research problem deserving support.

Designing a Good Benchmark

It seems desirable that a single benchmark test only one module in a system. In many cases this is not possible or meaningful due to the high level of interaction between the modules in an active vision/robot system.; e.g., it is hard to separate the performance of "where-to-look-next" and gaze shift. Consequently, each benchmark should come with a clear description of which modules it tests and how the results are measured. Furthermore, benchmark tasks should have varying levels of difficulty. In order to obtain a measure of reliability/repeatability, a benchmark might have to be run several times. If so, one must specify how the results from different runs can be combined to obtain some kind of performance index (e.g., average time to failure, number of false positives/negatives).

Single Measurement Benchmarks

Single measurement benchmarks are the type that people most commonly think of when they hear the term benchmark. Such a benchmark defines a single feature and a criterion for comparing measurements of that feature. For example, measuring the color of an object might be the benchmark and the comparison criterion might be accuracy or speed.

An advantage of single measurement benchmarks is that the experiments are well-defined and can be repeated. When used alone, the problem with single measurement benchmarks is that they are often not appropriate for comparing systems, even those that make use of the feature the benchmark is designed to measure. For example, a system designed to tell objects apart based on their color might use a color-based discrimination tree in which rough color estimates are all that is really required. While such a system might work very well at classifying objects, it will score poorly on a benchmark test of color measurement accuracy.

Distance to a Point A classic example of a single measurement benchmark is the accuracy with which a stereo vision system can measure the distance to an object. The criteria for comparing stereo distance measurements are typically: Accuracy of distance, speed of computation, or repeatability.

Simple Behavior Benchmarks

Simple behavior benchmarks try to define measurements on systems designed to extract more than a single piece of information from the world. Simple behavior systems typically couple a single sensed feature with some form of motion. The interesting behavior of the system is not the preciseness of the sensed feature but the robustness of the combined sensor/actuator loop.

As simple behaviors are typically building blocks of an active vision/robotic system, the results of such benchmarks should be checked carefully for appropriateness when they are used to compare different systems: a system may do poorly with respect to the benchmark criterion yet perform well at its particular task. However, simple behavior benchmarks have the advantage that the benchmark task typically specifies an overall system goal rather than a specific feature or method of achieving that goal. Therefore, the benchmark can often be applied fruitfully to a larger variety of systems that purport to use the behavior.

Unfortunately, the flexibility of behavior benchmarks is a two-edged sword. When the task is to carry out a behavior it is often difficult to specify a useful criterion for comparison. For example, suppose the benchmark task is to visually track a moving object (see below). Comparison criteria might be: Average reported position error of the target over time, the speed with which the target can move and still be successfully tracked, or the number of times in a trial that the target is lost. All of these criteria seem reasonable, but are they useful things to measure? The problem is to select a few useful criteria that cover a wide range of systems.

Object Tracking Object tracking is the classic example of a simple behavior. For example, reasonable benchmarks might be tracking any of the following:

- a small highly maneuverable object
- a large object that possibly does not fit in the field of view
- a distributed target
- an object that is intermittently visible
- an object with complex geometric properties
- an object subject to physical or geometrical constraints
- an object subject to significant occlusion
- a set of objects in a cluttered environment

Moving to Touch Another example of a simple behavior benchmark might be moving an arm until a probe just touches an object in a specific spot. This benchmark seems simpler than tracking because the criteria for comparison are more straightforward. One might measure the distance the actual contact varies from the desired contact point, or the speed with which contact can be achieved. The benchmark can be made more complex by allowing clutter and occlusions in the environment.

Integrated System Benchmarks

Integrated system benchmarks attempt to capture the performance of the most ambitious behaviors that a system using vision might exhibit. Such systems usually sense many features and use a variety of mechanical effectors to perform complex activities in the world. Benchmarks for evaluating this type of system usually describe a task to be accomplished and the criterion for comparison is "how well" the system carries out the task. The individual features sensed and manipulations performed during the task are not generally part of the benchmark.

Integrated system benchmarks measure the ability to accomplish actual goals in the world without reference to the way the system chooses to carry them out. A good integrated system benchmark should specify a task for the system to do and not a behavior that seems like it ought to be part of achieving that task. The criterion for comparison then becomes a measure of how well the task was performed.

Almost by definition, single measurement and simple behavior benchmarks are designed to test integrated system building blocks and thus suffer from a tendency toward artificial comparison criteria. Integrated system benchmarks use criteria based on ultimate use. However, by specifying comparison criteria related to a system's purpose, integrated system benchmarks slide into the realm of applications. Almost every researcher has a different purpose in mind when building a system and we run the risk of having as many benchmarks as researchers.

Eye-hand coordination An example of an integrated system benchmark would be an assembly/disassembly task that includes recognition, coordination, and choice of manipulator action. Such a task might be taking a lid off of a jar and replacing it where there are no a priori object models. The simplest scenario is a pull-off lid where the lid and the base are easily distinguishable in the visual image. The difficulty of the task could then be increased by making the lid screw-on or making the color of the base match the lid. This task involves picking up the unmodeled object, perhaps constructing a complete three-dimensional object model, and then learning by exploration how to remove the lid.

Classification of Haptic Qualities Another integrated system benchmark that involves both visual and manipulator actions would be classifying materials based on visual and haptic qualities. A simple scenario would involve labeling objects as: Rigid, deformable, or single-hinged objects. More complex demonstrations would involve occluding parts or recognition of movable and removable parts. This task involves manipulating the unmodeled object by compression or by applying forces at various location of the object and watching the visual results. Eventually, strategies for discovering haptic qualities might be learned by the system.

Conclusion

A useful benchmark must consist of a specific task and a criteria for comparing the results of different systems executing that task. Tasks chosen for current benchmarks range from the precision of single measurements to the ability to robustly carry out very complex tasks.

There is a strong desire to make benchmark tasks measure a single system feature because their are relatively few reasonable criteria for comparing a single feature can they be made relatively precise. However, the more specific that a benchmark gets, the more it relies on a system's architecture for legitimacy. Since the ultimate performance that we seek is in real world applications, the measurement of single features and simple behaviors implicitly assumes that those features and behaviors will be used in the application. Furthermore, it is assumed that better performance of the features and behaviors, as measured by their benchmark criteria, will translate into better overall system performance on the application as well. Unfortunately, it is not clear at this time what sort of system will be best for anything but the simplest application.

The more goal oriented that the benchmark becomes, i.e the closer that it gets to describing an actual application, the more likely it is to measure something that will ultimately by useful. However, that would suggest a proliferation of benchmarks designed to capture the idiosyncrasies of the many applications that might benefit from active vision. It also means that systems being tested must be complex enough to actually achieve that application. Currently, active vision work is still young and systems are simple.

In summary, benchmark tasks are likely to be limited to the single measurement and simple behavior varieties in the near term. As a result, considerable care must be taken when deciding whether a given benchmark actually tests something interesting about a system. It will be impossible to know whether a particular benchmark task and its comparison criterion capture a property that will ultimately be useful in applications until large systems are built and integrated benchmark tests can be made.

3 Hardware Support

3.1 Mechanical Hardware

Active vision systems do not just sense; they act. The most fundamental property of all active vision systems is their interaction with the world during sensing. Systems look in a particular direction, change their focus, move to another location, or reach out and touch something. Hardware beyond the visual sensor itself is required to make these motions: There must be a pan/tilt head to point the camera, the camera focus must be motor controlled, the camera head must fit on a mobile robot to move from place to place, and an arm is needed to reach out physically.

Simulation is inadequate for verifying and demonstrating active vision algorithms. Simulations cannot help but ignore many of the details and uncertainties that make active vision tasks complex (indeed that makes active vision ideas superior to more classical alternatives). For instance, uncertainties in camera positioning can be overcome by actively tracking an object and making small relative corrections as required. Simulations that do attempt to incorporate enough detail to model realistic situations are more complex, more difficult to deal with, and more expensive than the hardware they are meant to replace.

Active vision systems involve moving things and the mechanics of moving cannot be ignored. In fact, researchers in active vision believe that vision and mechanical systems working together often produce information more accurately and more quickly than either can working alone. The human eye system can be admired both as physics-based photon receptor and as an incredible servo control system; each part can be studied in isolation, but the package is synergy in action.

Categories of mechanical hardware

Three major categories of mechanical hardware are useful in active vision research. They are listed here in order of the frequency researchers currently use them:

– Computer controlled camera platforms

  camera pan and tilt

  camera heads (vergence, neck, etc.)

– Manipulators and mobile platforms

  robot arms

  mobile robot platforms

  robot manipulators (hands, grippers)

– Non-visual sensors complimentary to visual information

  tactile sensors

  inertial sensors

Each of these is discussed below along with the implications for their use in future active vision research.

Camera Platforms The single most important piece of mechanical hardware for current active vision research is the computer controlled camera platform. Pointing a camera at features of interest in the world is at the heart of virtually all active vision research. Hence, virtually every active vision research lab needs camera platforms.

The basic mechanical properties of a camera platform are identified in the list below. An ideal platform would make all of these parameters computer controlled so they could be actively adjusted. This list is designed as a guideline for those buying or constructing platforms.

For each camera:

– Pan

– Tilt

– Roll (about optic axis)

– Focus

– Aperture

– Zoom

– Optic Axis Calibration Adjustments

For multi-head platforms:

– Vergence (if appropriate)

– Baseline

From this list, the five characteristics considered most important to have under computer control are: Pan, tilt, aperture, focus and vergence.

In addition to the mechanical features listed above, an ideal camera platform would have the following characteristics:

– Simple geometry: This simplifies the conversion between visual and robot centered coordinate systems.

– Simple programming interface: This makes platforms easy to use and install.

– Low voltage and power consumption: These requirements must be met if the platform is to mounted on a mobile robot or small robot arm.

Commercial vs. Custom Hardware At the current time, the consensus is that commercial mechanical and optical hardware suitable for active vision camera platforms is quite difficult to come by and expensive.

Particularly critical problems exist in the availability of calibrated computer controlled optics. The camera platform characteristics discussed above place high importance on computer controlled focus and aperture. Since this is not available commercially, researchers must custom build such systems themselves.

Similar difficulties exist in the area of computer interface. Current hardware typically requires a separate motor controller board between the mechanical systems and the computer controlling them. Such boards usually plug into the computer's bus and thus are computer specific and require custom low-level driver code for each type of computer they are used with. This situation is fine for some computer systems (i.e., VME or IBM PC based systems) but is extremely difficult for others. Again, the result is that vision researchers often must write their own low-level interface and control software.

An integrated package including a pan/tilt platform, controllable optics, and a reasonable, flexible, computer interface is not currently available. This is a serious problem for the community and results in most researchers building their own hardware.

Building a Minimal System Camera platforms are fundamental to active vision research so every active vision researcher needs a platform. Since appropriate platforms are not available commercially, every active vision researcher must custom build a platform themselves. Building mechanical systems and connecting them to computers is not the professed expertise of active vision researchers (although many are extremely good at it) and considerable time and effort must go into learning what needs to be done, figuring out how to do it, and then having it all constructed. At best this results in a tremendous drain on resources and a tremendous duplication of effort. At worst it results in equipment that doesn't really perform the way it should.

To relieve active vision researchers from the burden of "reinventing the pan/tilt head" every time they buy a new camera, we feel that a common, generally useful, low-cost design should be developed for the active vision community as a whole. By funding a single effort to design, build, and distribute camera platforms to all researchers, a good deal of time, effort, and money can be saved by the research community as a whole. Money could be directed at active vision problems rather than to mechanical designers and builders.

This proposal is not meant to replace or eliminate any current programs that may exist along these lines. Rather, the proposal is meant to stimulate any existing program, or to spawn a program if none exists already. To this end, the features of a minimally useful camera platform are outlined in the next section. However, these features are meant only as a guideline and they do not constitute a design. Any platform that equals or does better than the outline would be welcome. The important thing is that some low-cost, integrated system be developed and made generally available.

A Minimal Camera System A minimal camera system for the average researcher would consist of a single camera on a computer controlled pan/tilt head. It would also be highly desirable for the camera to have computer controlled aperture and focus. It would similarly be highly desirable for the platform to consume a minimum amount of power at a relatively low DC voltage so that it could be used with existing mobile robot platforms. A simple, flexible computer interface for controlling the platform is also a must.

This minimal platform would have the following characteristics:

- Pan: Computer controlled 100°/sec pan rate, 1/10° pointing precision

- Tilt: Computer controlled 100°/sec tilt rate, 1/10° pointing precision

- Roll: Manually adjustable

- Focus: Computer controlled

- Aperture: Computer controlled

- Zoom: Optional, but if present computer controlled (50° field of view minimum)

- Camera Mount: Manually adjustable for optic axis calibration
- RS232 Interface: Direct connection to controllers an option

As a single camera system, the minimal platform would not require vergence or baseline control. However, it would be appropriate to design a two camera system that could be built on top of the minimal one camera systems. That way, a researcher could begin with a low-cost one camera platform and acquire a two camera system later on without wasting money along the way. A modular approach may also save design and manufacturing cost.

At this point we must stress again that the minimal platform is a guideline and not a design. An actual design would have to be evaluated by the research community to ensure nothing has been overlooked and the final result will enable good science. Furthermore, a better design is always welcome if the price can be kept down. The goal is to stimulate active vision research by making an effective, low-cost tool readily available.

Additional Considerations A primary consideration with respect to a camera platform designed along the above lines is that it represents a minimum. Thus, there will continue to be a need for some researchers to obtain custom platforms for some experiments. For example, real time motion tracking at the speed of humans appears to require pan and tilt speeds of greater than 250°/sec. Similarly, some fine motion tasks like tracing detailed edges or registering stereo images may require better pointing precision than 1/10° and experiments with peripheral vision will require more than a 50° field of view.

A common platform is not a standard platform. It is too early to define a standard platform on which all active vision can be based. Research based on other platform designs will continue to be an important part of the field.

Mobile Robots, Arms, Manipulators and Sensors

The use of mobile robots and robotic arms and manipulators is as much a part of active vision research as pointing cameras. The purposive, task driven nature of active vision algorithms makes them an integral part of navigation, manipulation and pointing systems. Thus, various researchers have begun to investigate the use of active vision in such systems and have been required to obtain some form of mobile robot or robotic arm to support those investigations.

At the present time, mobile robots and robotic arms of the types most frequently desired by active vision researchers are readily available and easy to use (though a little bit expensive). Furthermore, researchers typically have varying interests when it comes to robotic hardware and each laboratory has a different setup designed to investigate different tasks. Therefore, there is little consensus or desire for common robotic hardware in the way that there is a need for a camera platform. All that can reasonably be said is that robotic hardware plays an important, legitimate role in active vision research and different researchers usually require different resources. This is as it should be.

This happy state of affairs is not true when it comes to robotic grippers. There does not appear to be any hand/gripper available in the United States beyond the primitive parallel jaw gripper supplied with most robotic arms (except for the MIT/UTAH hand at over $100,000 each). This is perceived as a serious problem when it comes to research actively coordinating hand and eye. Currently such research is in its formative stages, but as it advances, researchers will want to put their efforts into the coordination and use of vision rather than into increasely clever tricks for getting a primitive gripper to do non-primitive things.

A similar situation exists with respect to non-visual sensors. In particular, two types of sensor will become increasingly important in active vision research: Touch and local inertia. Touch sensing is critically important in gripping and identifying deformable or textured objects. Integrating active vision and touch is thus seen as an important active vision topic. Unfortunately, current touch sensor technology is quite primitive and very expensive.

The term "local inertial sensing" is meant to capture sensing local orientation with respect to gravity (i.e., which way is down), local rotation (i.e., turning left, turning right), and local acceleration (i.e., falling or moving over a bumpy surface). It is possible to buy commercial sensoring elements for each of these aspects of the world, but it is difficult and expensive to integrate them into a single useful system. Coordinating vision processing with inertial sensing is also seen as an important direction for active vision research. Some estimate of local camera motion makes processing of the visual data much simpler and more reliable in many situations.

Active vision research involving robotic hardware and non-visual sensors currently involves work on many different problems by many different researchers. In the future it may be necessary to fund a common advanced gripper design if research in hand-eye coordination is to make progress.

Recommendations

Hardware is an intrinsic component of active vision research. All proposals will include the use of a pointable camera system. Many proposals will include the use of an arm, mobile platform, and/or manipulator.

We recommend the funding of design, development, and distribution of an inexpensive integrated camera platform with controllable optics.

Funding for novel hardware concepts, when researchers make an interesting case for properties not addressed by the minimal design, should be available. An example might be the need for extremely fast ($> 500°/sec$) pan/tilt mechanisms for studying some biological and tracking phenomena.

Few current active vision systems have used anything other than commercial DC servo motors. Emphasis needs to be placed on investigating novel actuator technologies in the demanding environment of small, fast, accurate eye platforms. Eric Schwartz's open loop miniature spherical torque motor presents an interesting step in this direction. Other possible areas of research are piezo, magneto-restrictive, shape memory, etc.

Over the long term, it might be a good idea to fund the design of an inexpensive, capable manipulator, as was discussed in Section 2.5.

3.2 Electronic Hardware

Advances in hardware and sensor technology have benefited and in some instances are enabling new progress in the active vision field. In this section we discuss a range of issues concerning the role of electronic machinery in active vision research.

The real-time constraint

By definition, active vision implies an intimate connection between a controllable imaging system, a possibly dynamic world, and computational processes. The processes must respond to demands originating from the outside world, and the effectors and processing must respond in timely fashion to meet the goals of the visual or robotic application. In this sense the active vision enterprise is one of real-time computation.

Hardware to support real-time vision involves at least the following components:

- Fast image acquisition. Frame-rate frame grabbers are commonplace.

- Fast low-level image processing. Currently this capability is commonly and inexpensively provided by pipeline devices.

- Fast interfaces to gaze control effectors.

- Some fast interface to higher-level computing. Entire images need not be transferred into general purpose computing memory, but some image information must be available for high-level processing. E.g., special hardware can provide the addresses of "interesting" pixels, allowing a host computer to do frame-rate blob-coloring and blob-analysis.

More esoteric and specialized real-time hardware is becoming increasingly available and increasingly present in leading laboratories. Examples include:

- Smart sensors which provide on-chip processing at the sensor, resulting in symbolic labels attached to the features output by the sensor.

- Fast interfaces which can map entire images into SIMD (single-instruction/-multiple-data) or MIMD (multiple-instruction/multiple-data) computer memory at frame rates.

- Spatially variant sensors which help to reduce the number of pixels needed for image analysis by orders of magnitude.

- Hardware boosters that allow emulation of spatially variant sensors and other sophisticated early vision functionality. Notions like space-variant active vision imply a manipulation of the spatial structure of the sensor. One way to achieve this prior to the development of practical VLSI sensor implementations is by means of dedicated processors operating adjacent to the sensor. These hardware boosters typically have a function that is sufficiently specialized that currently available technology can be used to achieve high levels of performance. In addition to log-polar remappers that emulate a spatially variant sensor using a commercially available uniform resolution sensor, video-rate boosters have been developed to compute things like (a) multi-resolution pyramid images with electronic foveation, (b) Laplacian of Gaussian convolutions with large kernel sizes, (c) image warping, (d) color transformations, and (e) various histogramming and image statistics. Technologies that are relevant to "boosters" include (a) digital signal processing (DSP) chips which are increasingly available, powerful, and friendly; (b) microcontrollers; (c) gate arrays; and (d) digital and analog VLSI.

- MIMD architectures may well be necessary to support the necessary concurrency of cooperating high-level vision tasks, control tasks, reasoning and planning tasks. With spatially variant sensors the number of pixels in an image is relatively small. Thus a relatively low amount of computation is needed for most operations. On the other hand in active vision systems there are many types of tasks that are to be done at any given time. For example, different image feature detectors may being running concurrently with peripheral alert processes. These considerations suggest that an appropriate computing architecture for active space variant vision is a MIMD one. A loosely coupled, coarse grained, multiprocessor collection of medium power processors is more suited to the active vision task than a massively parallel

SIMD architecture. The SIMD machines are inflexible and lack the reconfigurability that is required to implement active, adaptive, algorithms. A significant aspect in active vision systems is its real-time asynchronous nature. The processors of a SIMD machines work in lock-step synchrony and are not well suited to efficient handling of asynchronous events. There are a number of possible approaches to the multiprocessor architecture. For example, there are dataflow architectures, message-passing systems, object oriented, blackboard based systems, and so forth. Hardware implementations of these multiprocessor systems could include networks of workstations, transputer arrays, collections of DSP boards, multiple single chip microcontrollers or even programmable gate array based systems.

Software

In addition to the hardware requirements it imposes, the real-time constraint places demands on software at several levels. First, there are practical problems in controlling real-time systems with the UNIX-based workstations that are conventionally used as hosts. One solution is to add dedicated processors running either single-user code not subjected to preemption (e.g., code downloaded onto bare 680x0 computers, or even code running under MS-DOS), or multi-threaded code running under commercial or academic real-time kernels such as VxWorks or OS9. Another practical problem arises from the fact that most image processing hardware (e.g., DataCube) is extraordinarily difficult to program. The learning curves are long, and knowledge does not transfer well from project to project. There is interesting work to be done in providing more appropriate abstractions for describing the computations to be done and in mapping those abstractions onto the available hardware.

The programming problem becomes more severe for systems that require special-purpose hardware at higher levels of the system. Image processing hardware is usually well described by a SIMD or quasi-SIMD model. However, the higher levels of active vision systems are typically designed as collections of quasi-independent processes that are best described by a MIMD model. Programming either type of system is difficult with current software technology, despite years of research. Programming heterogeneous systems with both SIMD and MIMD components has received less attention and is even more difficult. Programming languages and software environments that make programming heterogeneous systems easier would greatly facilitate active vision research as well as being interesting in its own right. Although the general problem is very hard, we believe that solving a reduced version tuned to the needs of active vision is tractable.

Some of the issues that such systems should address are inherent in active vision, independent of the nature of the hardware. Resource allocation is a problem that arises in several guises. Different processes may wish to use the sensors and front-end hardware in different ways, which may or may not be compatible. Likewise, computational resources are usually scarce, so scheduling processes becomes an issue. Recent work in real-time operating systems has begun to address the latter problem and should be explored.

Custom hardware

Given recent high performance gate array, DSP and VLSI CAD support systems, it is feasible for relatively inexperienced computer science oriented people to produce "boosters" in a relatively short time: To a surprising extent, current availability of CAD support and demonstration systems have turned this type of hardware exercise into largely a software one.

VLSI sensors and processors At the lowest system level research effort will be required in the field of space-variant sensors, preprocessing at the sensor array, and smart sensor architectures. Examples where on chip processing could be especially advantageous is spatial and time derivatives. On chip analogue computations may allow coarser temporal and spatial discretization than would be necessary to accomplish the same computation by sampling full images and transferring them off chip to an external booster device. On chip it may also be possible to take advantage of the continuous nature of visual information (even motion blur can be advantageous). On chip computation and spatial variant sensor designs will greatly reduce the output bandwidth from the sensor necessary to support high performance active vision tasks. Thus while many technological and theoretical issues remain to be addressed before such devices can overtake board-level boosters, they offer the possibility of much more compact, cheaper, and higher performance systems.

Board level boosters Barriers to doing board level design are dropping and we are starting to see significant new results in the active vision area made possible by special purpose accelerator boards designed by individual vision researchers. The driving forces behind this trend include the blossoming collection of digital chips available commercially, the ease of doing digital hardware design with the low-cost CAD tools currently available, and the availability of board fabrication technologies such as machine wire-wrap and wire-bonding that require only a wire-list to construct a board. It is of course still a challenge to design an accelerator that advances the state of the art, but our experience has shown that it can be done even by small research groups with no extensive prior hardware experience.

4 Applications

4.1 Introduction

Active vision research is likely to enable new applications that are either not cost-effective or not technologically feasible without intelligent control of the data acquisition process. Recent technological advances in sensors, actuators, digital signal processors, and microcontrollers have made possible the development of small, inexpensive, real-time, active vision systems costing less than $10,000. Breaking this price barrier suggests that computer vision can at last be applied to a variety of consumer and other end-user applications. Recent algorithm advances have enabled the solution of computer vision problems that are otherwise ill-posed or intractable. This enlarges the set of tasks for which computer vision is feasible.

Our general criteria for evaluating different applications is the degree of their appropriateness for (the value added by) an active vision approach. Specific criteria include feasibility, cost and performance.

Feasibility: The application of active vision techniques can facilitate certain tasks that would otherwise be impossible using passive vision. An example is a process that incrementally builds a model of its environment, with little or no a priori information about the expected shape of objects.

Cost: Active vision can lower the cost of applying computer vision. For example, a foveated sensor coupled with a positioning system can replace a high-resolution sensor array. Since less data needs to be acquired and processed, there can be significant savings in hardware costs.

Performance: The performance of a vision system can be significantly improved using active vision techniques. Improvement can be measured in terms of reliability, repeatability, speed, and efficiency in performing specific tasks as well as the generality of the kinds of tasks performed by the system. For example, reduction of data acquired by an active vision system results in a net increase in speed of processing. The closed-loop nature of active sensing serves to improve accuracy and robustness through the minimization of uncertainty.

Appropriate applications are serial in time, because typically with an active sensor we look at, and therefore process, only one thing at a time. This does not exclude massively parallel processing at discrete times, nor does it exclude processing multiple sensors in parallel, nor does it suggest long time steps (500 degree/sec saccades are sequential, but not slow).

In this section we describe selected applications, focusing on why active vision techniques succeed rather than on the overall significance of the application. The collection of applications discussed is by no means exhaustive; the thoughtful reader can imagine many more.

4.2 Examples

This section describes selected applications: Unmanned vehicles, industrial processes, aids for the handicapped and service robots.

Unmanned Vehicles

Unmanned Ground Vehicle An unmanned ground vehicle must be able to achieve tasks such as navigating across open terrain and traveling on roads at high speeds. It must function in highly unstructured, uncertain, potentially hostile environments, performing tasks such as obstacle avoidance (e.g., avoiding rocks, trees, logs, craters, potholes), road following, and landmark-based navigation, all at high speeds.

The application of traditional passive vision to this problem presents difficulties. Passive vision sensors with both wide field of view (perhaps 180 degrees) and high spatial resolution are expensive to manufacture, and supply so much data that processing becomes computationally prohibitive. Passive vision techniques that involve reconstruction of the world are extremely time consuming, and thus not suitable for high speeds; the problem requires use of specific information at specific times and not all information all the time.

Active vision techniques circumvent these problems. Active gaze control can point inexpensive space-variant sensors (which have wide fields of view and varying spatial resolution) at regions of interest, efficiently satisfying both angular and spatial requirements. Active vision techniques for tracking features such as road edges, for example by predicting the location of the road tangent point from one instant to the next, avoid 3-D reconstruction, and can be efficient enough to satisfy real-time requirements. In a multi-agent unmanned ground vehicle scenario, active vision can be valuable in holding gaze on other agents.

Unmanned Underwater Vehicle Mine clearing in shallow water (3 to 30 meters) requires navigating to the area of interest, detecting mines, transiting to the mines for low-light imaging, and using vision to extract features that are then used to attach detonators (either buoyant or weighted).

In conventional approaches, the vehicle carries the sensor to a number of different positions to construct a composite image of all sonar and optical returns. This image is then processed to locate mines. It is generally agreed that no existing conventional approach will succeed in even a crude fashion. The shallow water environment is characterized by strong, intermittent wave surge that generates particulates in suspension, obscuring optical sensors for long-range mine detection. The wave surge buffets the vehicle, disrupting structured search and making it necessary to ride out the surge, re-position the vehicle, and re-acquire the objects being sensed. Finally, the proximity of the seabed and the water surface attenuate and produce multiple paths for sonar returns, limiting both object detection and acoustic navigation accuracy.

Active perception behaviors on-board the underwater robot can improve mine detection efficiency and improve navigation accuracy. The capability to pan and tilt permits more efficient area searches for the acoustic anomalies caused by mine-like objects. Once within low-light optical range, visual servoing to the mines is significantly more accurate than the absolute coordinate positioning obtained from acoustic transponders. Finally, navigation among the mines will also be more accurate using relative servoing from bathymetry features than using acoustic navigation.

Unmanned Air Vehicle Unmanned aircraft that can land in unstructured, potentially hostile sites could play a vital role in making military and reconnaissance operations safer (by subjecting human pilots to less risk) and cheaper (by crashing fewer drones).

The environmental information required to land an aircraft must be acquired at a high resolution and over a wide area (relative to the dimensions of the vehicle). The bandwidth required for teleoperated landing makes telemetry to a passive vision system very expensive, and under conditions where stealth is desired, not feasible. Further, vibration of the aircraft corrupts the imagery, causing severe difficulties for processing. Finally, with a single fixed camera it is difficult to locate features of the runway that are relevant to landing, such as the vanishing point.

Active vision techniques to fixate and to perform saccades with a variable-resolution sensor can reduce the telemetry bandwidth (for teleoperated landing) and the amount of processing (for autonomous landing) required by several orders of magnitude. Further, these techniques can eliminate the noise due to the mechanical vibration of the aircraft, avoid image blur that accompanies fast motion, and serve as a visual compass. By continually pointing the camera at the vanishing point of the runway, active vision techniques simplify image feature detection and provide stable feedback for control of the aircraft.

Outer Space Operations Outer space operations such as maintaining satellites, repairing external equipment, and assembling large structures must be conducted in the unprotected void. Monitoring and controlling these operations based on visual images is difficult because of the extreme illumination conditions. Objects that are illuminated directly by the sun appear very bright, as do objects illuminated indirectly by the light reflected from smooth, white surfaces. Objects that lie in shadow appear very dark, as does deep space itself.

Passive vision techniques perform poorly under these conditions. Typical photodetectors in cameras do not have the dynamic range required to produce images of such scenes with reasonable contrast. Automatic gain control ameliorates some of the problem, but effectively loses information, and further, introduces calibration problems.

Active vision techniques can succeed in these harsh conditions by adapting camera parameters to the scene. For example, they can improve contrast by controlling the aperture, and they can avoid troublesome areas by shifting gaze. In many cases, the optical parameters are highly interrelated, so it is not possible to change only one parameter without violating some constraint. Thus, it is more likely that the active vision technique would reset a significant number of parameters.

Industrial Processes

Active vision offers a number of advantages for industrial processes. Two industrial processes are now briefly discussed in terms of the criteria stated in the introduction.

Tracking Tracking is critical to applications in automated welding, painting, and cutting. Most seam tracking applications (welding and cutting) are already implemented using a subset of active vision techniques. For example, a sensor mounted on the end-effector is used to servo the end-effector so that the sensor always tracks the seam at a specified offset. Further improvements in cost and performance are likely to be achieved using foveated sensors.

Rivet inspection The robot must check whether all the rivets in an aircraft wing are present and installed correctly. This task stresses accurate measurement over large distances combined with high resolution examination of the individual rivets for aspects such as circularity and flushness.

Aids for the Handicapped

Imagine being quadriplegic, sitting in an automated wheelchair equipped with a camera system and manipulator. You command the system to go to the kitchen by pointing with some input device to an image of the kitchen door. The wheelchair could then visually navigate to the kitchen, avoiding obstacles along the way (visually-directed cruise control). You would arrive at a desired location in the kitchen by pointing to a series of navigational targets. Through a series of visual interface commands, you instruct the manipulator to open the cabinet door, pick out your favorite mug, pour a cup of coffee, place a straw in the cup, and then hold the cup in a position from which you can drink.

Traditional computer vision systems are impractical for these applications. First, specific geometric object models can be hard to obtain for individual users (everybody's cabinets look different). Second, geometric maps of all places that a user may want to go to are unavailable, and furniture rearrangement is common. Third, these computer vision systems are not real-time.

Active vision offers solutions to these navigational and manipulation problems. Each situation requires that the camera actively move, focus, and maintain a view of the object or pathway of interest. The camera must also input enough 3-D information for the manipulator to grasp an object. The entire task must be visually servoed since the object models or object locations are unknown with respect to global coordinate systems. To pick up an object, for example, fixation can be used to segment the desired object from the background. A 3-D object model can be built by controlling the focus and/or the motion of the camera. Visual servoing techniques could then be used to pick up the object. Many of the manipulation techniques are also applicable for navigation. The camera can track features of the scene, use its motion to identify obstacles in the path, use its motion to determine if other moving objects are on a collision course with it, or use fixation to determine a target location.

Service Robots

Household Service Robots Here, we will concentrate on the household tasks of tidying and vacuum cleaning, which require the following capabilities:

- Navigation through an indoor, potentially cluttered, unstructured environment with many unknown obstacles. Objects such as chairs, tables, and toys may be arbitrarily placed in the environment.

– The ability to grasp and move objects, both large and small, and to put them back in place.

An active vision approach could make these tasks much easier to accomplish. First, the ability to actively control the camera parameters allows the robot to perform visual search for what it is seeking. It can look on the floor, on top of tables, or under tables. Also, active camera control can allow it to get alternative views of objects which may be in shadows or in poor lighting. Second, the ability to perform stereo through vergence control can greatly aid in determining the 3-D position of objects that need to be grasped. Achieving obstacle avoidance by servoing the motion of the robot relative to a fixation point on the obstacle can make the problem of obstacle avoidance with little clearance tractable, since the absolute position of the robot in the room need not be known (as with the case of passive vision).

Sentry robot The robot tours a list of stations, stopping at each to carry out a simple surveillance or manipulation task. The robot is equipped with a map and may choose its own routes between locations. Sample tasks at stations are checking license plates for unauthorized vehicles and verifying that doors are securely locked. This task stresses long-range navigation using a map and its functional integration with a variety of station tasks.

Other Applications

Other applications will surely benefit from active vision approaches. They include the following: Terminal homing to target for smart weapons; detecting and tracking intruders; determining a computer user's intentions by tracking their eyes and hands; rummaging through garbage to find items to recycle; defensive driving monitor for automobiles; fruit and vegetable harvesting; sorting irregular postal pieces; overcoming the time delays in remote teleoperation; assessment of toxic waste sites; airport runway surveillance, where very large stretches of runway must be monitored, but no single resolution can provide enough detail at long distances to determine whether a vehicle or person is authorized.

4.3 Summary

We have identified a number of applications, including unmanned vehicles, industrial processes, aids for the handicapped and service robots, that active vision techniques can make more cost-effective or technically feasible.

In most of the applications considered, computer vision has already proved useful (although perhaps not highly successful). But that is vision of the conventional kind: Either off-line, or requiring supercomputers or parallel processors for real-time performance. We recognize that most biological visual systems exploit massive parallelism, and that we will need to also, especially to solve some of the harder problems mentioned. However, we are convinced that active vision, with its emphasis on real-time interaction with the real world, will succeed in applications that satisfy one basic requirement: The task to be performed must be sequential at some level. If not, then data from a high-resolution passive sensor can be processed in parallel, providing better performance.

5 Conclusion

We have outlined some of the exciting areas of research in active vision, including foveal sensing, gaze control, attention, eye-hand coordination and integrating active vision with architectures for robot control. The section discussing each area explained some of the open problems and research goals in that area. The "Benchmarks" section gives some suggestions for how progress may be evaluated towards those goals, and stresses the importance of developing measures to make this possible.

Technological progress in computer-controllable mobile camera platforms and lenses, foveal sensors, and real-time image processing architectures is vital to research and applications of active vision. Cameras are becoming smaller, lighter and easier to maneuver, and systems for orienting them smaller and less expensive. Similarly, off-the-shelf image processing equipment is becoming faster and less expensive, and custom equipment is becoming easier to build. There is enough overlap in the needs of researchers in the field that cooperation will help to reduce the cost of research equipment such as mobile camera platforms.

A number of factors point towards the potential importance of active vision to applications of computer vision. These include the dramatic reductions in complexity permitted by selective attention to incoming data and spatially-variant sampling; advances in robustness by the use of cues well-suited to robot tasks; and close integration with robot architectures. The community has both demonstrated the uses of advances in cameras and real-time image processing technology and is contributing to advances in these and related areas. We believe that elements of active vision will be important in the application of computer vision to many real-world problems.

It should be stressed that research in other areas of computer vision besides those mentioned in this document will be instrumental in research in active vision. In fact, we hope that many of the ideas expressed in this document will be adopted by the general vision community. Active vision is the natural result of considering vision in the context of an active organism and its changing environment.

6 Bibliography

Although the concept of active perception is relatively new to computer vision, it has a long history in the psychology of perception, going as far back as Helmholtz (1910). Important references from this field that have influenced research in active vision include:

J. J. Gibson. *The Ecological Approach to Visual Perception.* Houghton Mifflin, 1979.

A. L. Yarbus. *Eye Movements and Vision.* Plenum Press, 1967.

The following list is a selection of papers presenting the views of some of the leaders in active vision (listed in approximate order of first publication).

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

P. J. Burt. Smart sensing with a pyramid vision machine. *Proceedings of the IEEE*, 76:1006–1015, 1988.

Eric Krotkov. *Active Computer Vision by Cooperative Focus and Stereo,* Springer Verlag, 1989.

J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1:333–356, 1988.

A. L. Abbott and N. Ahuja Surface Reconstruction by Dynamic Integration of Focus, Camera Vergence, and Stereo In *Proceedings of the Second International Conference on Computer Vision*, pages 532–543, 1988.

Dana H. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

John Aloimonos. Purposive and qualitative active vision. In *Proceedings of the International Conference on Pattern Recognition*, 1990.

Many other references to papers published in the area of active vision may be found in the Rosenfeld computer vision bibliography, available to those who have access to the Internet by anonymous FTP from ADS.COM (in the VISION-LIST-ARCHIVE directory). Additions to the bibliography are published annually in the journal Image Understanding, previously entitled Computer Vision, Graphics and Image Processing. The work done by the authors of this document is well indexed by this bibliography; they will provide more extensive discussions of many of the ideas expressed here, and they will lead the reader to a rich variety of work relevant to this rapidly evolving field.