WORKING PAPER 85

SOME ISSUES FOR A DYNAMIC VISION SYSTEM

Mark A. Lavin

Massachusetts Institute of Technology

Artificial Intelligence Laboratory

Vision Group

December, 1974

ABSTRACT

This paper is a thesis-proposal-proposal: a discussion of some issues which seem relevant to the problem of dealing with visual scenes undergoing change. The problem area is broadly stated, some relevant points are noted, and a possible scenario for a thesis is discussed.

Work reported herein was conducted at the *Artificial Intelligence Laboratory*, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense and monitored by the Office of Naval Research under Contract Number N00014-70-A-0362-0003.

Working Papers are informal papers intended for internal use.

This paper discusses several issues relating to vision systems which can deal with scenes that "move" (change over time). I hope to pursue research with such systems for my doctoral dissertation; this paper is a kind of thesis-proposal-proposal. It is not a progress report; I intend it primarily to solicit comments and suggestions from you, the reader.

Psychology's View of Motion Perception

Response to movement in the visual field occurs in even the most primitive organisms which can be said to have "vision systems" [1]. So, we might turn to Psychology for some explanation of the process. However, when we examine studies in Motion Perception, the results are rather confusing. As with many other areas relating to perception, there is a lack of clear-cut correspondence between physical events ("real movement" in the world, or in the retinal projection) and "apparent" or "perceived" phenomena. Motion Perception also has a bag of "illusions" [2] to further complicate the issue.

Most critically, the motion which an organism "perceives" is partially predicated on factors in the visual field which have little to do with motion *per se* (e.g., stimulus intensity, shape of surrounding field, etc.). Perhaps there really isn't a separate well-defined function within the visual system which we could term Motion Perception. James J. Gibson states:

> *The general conclusion is that the problem should not be stated as one of the perception of "motion" (of the kind studied in physics) but (a) as various problems of the perception of change--of environmental events as given to the eye by transformations of the optical pattern, and (b) as one aspect of the general problem of the sensory control of behavior by feedback stimulation.* [3]

This says to me that the problem of constructing a vision system to respond to moving scenes (a *dynamic vision system*) cannot be addressed simply by adding "motion-detectors" to the retinal array. Rather, we must re-examine what we mean by the VISION PROBLEM.

Toward a New Model for the Vision Problem

Following Gibson's lead, I propose the following as a definition of the VISION PROBLEM:

> *The task of a vision system is to establish and maintain a coherent [4] description of the environment, using input to the eyes [5].*

The critical point of this definition is the emphasis on "maintain." As the environment changes, events (mainly motion) occur in the visual field. The vision system should respond to these events by modifying its model of environment. The main difficulty is that we are trying to use two-dimensional information (the retinal projection) to infer three-dimensional facts (the shape, location, and movement of objects in 3-space). The constraints on physical objects, and our familiarity with the "real world" partially help us out of this difficulty.

The need for dealing with changing scenes seems obvious to me. However, the only recent work in the MIT AI Lab on looking at scenes undergoing change [6] is that of Bob Woodham. However, that work was limited to very specific kinds of motion as a means of providing feedback for effector control. I propose to look at the problem more globally.

Why has so little emphasis been given to the problem of dynamic vision? Perhaps it seemed too hard: The fact that a good line-finder takes so long to process a single frame may have discouraged anyone with a sub-megayear attention span. Perhaps the problem seemed too difficult to debug: For static scene-analysis, it is reasonable to store test pictures for repeated tests; for testing a dynamic vision system, storing what amount to movies might prove too costly [7]. Perhaps the problem appeared irrelevant: perceiving motion is a separate issue which can be dealt with later; but this seems to be just what Gibson was arguing against. Finally, maybe the problem just seemed too easy: given that we have adequate (albeit slow) methods of analyzing static scenes, couldn't we just apply these methods iteratively to yield information about changing scenes (in the same way that a series of still frames yield a movie)?

I want to argue forcefully against such a "quasi-static" approach to dynamic vision. For one thing, there seems no reason to update the model when a scene does not change (at least as it appears in the retinal array). Further, when parts of a scene change, it seems unreasonable to reprocess the entire scene. Finally, even for those parts of the scene which do change, it seems likely that the entire processing need not be re-applied...some parts could be salvaged. At very least, the knowledge from previous frames might be useful to guide processing of this new frame (e.g., by anticipating new object locations).

I next discuss three areas which research in dynamic vision systems will have to deal with.

The Domain: Movement in the Real World

If we want the dynamic vision system to respond to change in the environment, we must ask: what kinds of changes are possible? To a large extent, this boils down to the question: What kinds of movement can occur in the real world [8]? Basically movement in the scene (or the corresponding retinal array) derives from three sources:

(1) Movement of Objects in the scene [9]

(2) Movement of the viewing point

(3) Movement of the source of illumination

I think it's useful to further divide categories (1) and (2) according to whether the movement is independent of, or under control of the perceiving organism [10].

How do the various types of movement relate to the problem of dynamic vision?

1) Obviously, motion is the major contributing factor to change in the environment, and hence to the requirement for a <u>dynamic</u> vision system.

2) Information <u>about</u> motion (e.g., "the big red block is moving northwest at Mach 2.3") is an important feature of the environment which the dynamic vision system should record in its model. This seems particularly critical both for "anticipation," (e.g., "any minute that block's going to cross over Oregon"), and for manipulator applications, where movement must be related to forces (e.g., "we're going to need more than a catcher's mitt to stop it").

3) Motion information may shed light on the static nature of objects in the scene. I like to think that there is some analogy here to the bug==>feature transformation which Waltz performed with shadows. That is, movement may not be confounding factor in shape recognition (something that the recognition should be invariant with respect to) but a positive aid. I am thinking primarily in terms of the close relation between depth perception based on stereopsis, and on motion parallax (static vs. dynamic way of getting two views of a scene).

4) Finally, it is motion in the scene (or in the retinal array) which should "drive" the dynamic vision system. This will be discussed more in the following section.

The Processing in a Dynamic Vision System

Attempts to extend current vision systems to handle dynamic vision will run into the objections raised above for a quasi-static system. Getting around this will probably be the major problem in my proposed work. Let me suggest that the system should be a "change-driven process(es)" in the following senses:

> ***(1) If the purpose of processing is to maintain a description, no actual processing should occur when the environment (or the retinal array) does not change*** [11].

Such an idea immediately suggests a retina full of little change detectors, which fire when local intensity changes. Any such firing results in an interrupt which would trigger processing. This can be extended beyond these gross levels as follows: to the extent that any subprocess of the system consists of mapping some input description (e.g., a line drawing) into an output description (e.g., a Winston net), that subprocess should be executed only when its input description changes [12]. The benefit of such a scheme is dependent on how much the system looks like a highly layered (pass-oriented) affair, which may be undesirable.

> ***(2) When the input does change, it would be preferable to map the change into a corresponding change in the output, where possible, rather than recomputing everything.***

As an example: it may be possible to map changes in vertex angles in a line-drawing directly into changes in the orientation of the objects or movement of the viewing point. For complicated domains, it is not immediately clear that such incremental processing is possible.

There is one clear flaw with change-driven processing: how does it start up? I must waffle a little bit here, but I believe that if the environment is not too bizarre and relatively familiar, it may be the case that situations of "starting up" are really quite rare. In any case, all the capabilities of a static vision system would presumably be available to the dynamic vision system as well (the issue of designing a control structure which will "fail softly" back to brute force methods is an interesting one).

The Output of the Dynamic Vision System: the Environmental Model

What must such an environmental model look like? Clearly, current models (e.g., Winston nets, for the BLOCKSWORLD) will have to be augmented with primitive descriptors relating to motion [13]. Beyond this, I believe that the bag of world-knowledge sufficient for the task will have to go somewhat further: better ideas about space and time will be necessary. In addition, the descriptive mechanism will probably contain more good ideas about Kinematics (that goes almost without saying) and Dynamics (e.g., "objects supported by a single object tend to undergo indentical translations to it, barring outside intervention;" or, "if you push something hard enough it will fall over"). An interesting question is whether a lot of new facts are needed to deal with motion, or whether most of that information is already contained in the static descriptions of objects and scenes.

The success of Waltz's program suggests that the "right" descriptive mechanisms may yield the process for producing those descriptions as a relatively low-cost side-effect.

Toward a Thesis

I have argued above that dealing with changing scenes is an interesting problem area in machine vision research. I hope I have illustrated that it has enough ramifications and difficulties to make it non-trivial. But a problem area is not the same thing as a thesis topic (remember the "Summer" vision project?). Sometimes, the distinction may not be great: if one's problem is understanding line-drawings of scenes with shadows, an obvious thesis topic is to write a program which understands (builds descriptions of) such scenes.

I think that a good thesis on understanding changing scenes will at least touch on the three issues described earlier:

> *(1) Enumerating the kinds of change or motion which can occur in real scenes (or some restricted domain) and the kinds of concommitant changes these produce in the retinal array (the actual dynamic vision problem involves mapping the latter into the former).*
>
> *(2) Creating a descriptive system for the chosen domain which is easily updated when changes occur, and which can embed information about change.*
>
> *(3) Devising a control structure (maybe no more than a big DO loop) which can guide the processing. The underlying assumption is that the quasi-static approach is clearly unacceptable.*

Toward a Scenario

With all the above as preamble, let me now suggest a specific (but highly tentative) thesis topic:

A line-drawing movie (or its symbolic equivalent, some sort of display list) would be created to model the view seen by an imaginary vehicle on an excursion through an imaginary city made up of (you guessed it!) BLOCKSWORLD-type buildings. The movie would be "shown" to the dynamic vision system, whose task would be to produce a description of the path it took and the "sights" it saw along the way. These data might be presented by having the program draw a rough plan of the imaginary city and its route through it.

The problem could presumably be elaborated by having elements of the scene (other than the viewpoint) moving. If the simulated movie could be produced on-line, the program might initiate commands to the vehicle to search for some other object, or to pursue some moving target.

A Final Word Regarding the BLOCKSWORLD

As noted above, my proposal is yet another excursion into the BLOCKSWORLD/line-drawing tradition of vision programs, a fact which is sure to bring many fans of AI research to their feet and heading for the exits. Let me briefly suggest two justifications for this choice:

(1) As mentioned previously, BLOCKSWORLD scenes can be stored very compactly, for diagnostic purposes.

(2) I feel that many issues relating to movement and change transcend the particular kind of domain in which such changes occur (particularly with respect to control-structure issues). As such, I do not feel that working in the BLOCKSWORLD really constitutes "cheating."

Of course, it's very easy to make facile comments (consider, for example, this one here). I hope that the end result of this research will be such as to allow subsequent programs to deal with moving scenes represented in other formats (e.g., a region/blob-oriented representation such as is used by Yakamovsky in his road scene hack).

NOTES

[1] Lettvin has described the bug detectors (retinal cells which respond to small, concave, moving objects) in the frog. Even insects respond to motion: try catching a fly.

[2] Some examples are the Phi phenomenon, apparent movement, and motion aftereffect. But are illusions a bug or feature? Often they reflect the subtle underlying processing strategies. For example, the phenomenon of colored shadows seems to be a result of the eyes' very nice hacks for maintaining color constancy.

[3] The quote is from an article by James J. Gibson in a book, ***Readings in the Study of Visually Perceived Movement,*** by Irwin Sigel (Harper and Row, 1965).

[4] A nice word..."coherent." It connotes consistency, elegance, and utility. It also puts me in mind of McDermott's thesis, which I think has many implications for a system trying to acquire knowledge serially.

[5] Perhaps this is too parochial: for example, there is strong evidence that the vestibular system (middle ear balance) interacts rather strongly with the vision system. And mightn't the following be considered forms of vision: binaural auditory localization, bats's sonar, the infrared sensors of some poisonous snakes?

[6] As opposed to looking at static scenes before and after change, as in the Copy Demo.

[7] One interesting possibility is to store the set of scenes symbolically, and have some program (e.g., a dynamic hidden line processor, as discussed in Lavin's Working Paper 66) regenerate the series of test scenes on line. I suspect that many of the issues encountered in building an effective generator along these lines would carry over to the dynamic vision system. In this vein, I believe Pfister's recent Ph. D. thesis on DALI (***The Computer Control of Changing Pictures***) might be very interesting.

[8] Of course, turning on a light bulb produces change with no "motion." A dynamic vision system would have to account for this as well.

[9] I would further distinguish the cases where parts of objects move. For example, I perceive an expanding balloon in terms of a scaling transformation on a single object, not the translation of its many surface elements away from the center.

[10] Psychologist Richard Held has some nice points points this distinction, particularly regarding development of motor-vision coordination.

[11] I imagine an input device with a coarse, wide-angle set of "event detectors" (basically little scene-subtractors) which would call attention ot places in the field where change is going on. Then, finer "foveal" elements could look at these locations more carefully.

[12] In some cases, changes in the input description to one layer will not propogate to the output description for that layer; demanding that every layer be change-driven capitalizes on such early cutoff situations.

[13] We will need descriptors about the motion of single objects (e.g., directions, rates, and terms like yawing, pitching, rolling, etc.); prepositional relationships between objects (e.g., "toward", "at", "out from under", etc.). Citing McDermott's work again, I see the need for a limited, but rich set of qualitative Descriptors.