Visual Feedback in a Coordinated Hand-Eye System

VISION FLASH 32

by

Robert J. Woodham

Massachusetts Institute of Technology

Artificial Intelligence Laboratory

Robotics Section

AUG 1972

Abstract

A system is proposed for the development of new techniques for the control and monitoring of a mechanical arm-hand. The use of visual feedback is seen to provide new interactive capabilities in a machine hand-eye system. The proposed system explores the use of visual feedback in such operations as the pouring and stirring of liquids, the location of objects for grasping, and the simple rote learning of new arm motions.

This paper reproduces a thesis proposal of the same title submitted to the Dept. of Electrical Engineering for the degree of Master of Science.

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.

This memo is located in TJ6-able form on file VIS;VF32 >.

0. Introduction

As an introduction to this proposal, I would like to briefly outline the ideas and observations that represent the initial motivation for developing a machine hand-eye system.

0.1 Why Hand-Eye Coordination?

The classic issues of artificial intelligence, as we have seen them, involve the representation of knowledge, problem solving and learning. Vision and the BLOCKS world has been one domain for studying these issues. In a very crude way, one might characterize our work in vision at MIT into three phases:

1) Scene Recognition

the development of techniques to recognize and represent objects in a real world scene using descriptions somewhat akin to those a human might use.

eg. Binford-Horn linefinder {5}
Guzman's thesis {4}
Huffman {6}, Clowes {1}, Dowson {2} {3}, Waltz {8}
Winston's thesis {11}

2) Learning/Problem Solving

the use of these descriptions together with additional real world knowledge to develop learning/problem solving systems that demonstrate an "understanding" of the problem domain.

eg. Winston's thesis {11}
Winograd's thesis {10}

3) Manipulation

the use of the arm-hand to implement solutions generated by the problem solver/planner.
eg. the COPY demo {12}

We observe that, at MIT, there have been no recent AI theses dealing with manipulation nor any AI theses dealing with hand-eye coordination. The focus of our research has involved not the enactment of a solution to a problem but rather the generation of a plan for doing so. Winograd's thesis {10} is no less profound because manipulation was simulated on the display rather than implemented using the arm-hand.

0.1.1 An Engineering Hack?

If one's model of the ideal robot consisted of a very intelligent vision system whose output to the world was a sequence of instructions responded to by a numerical control type device, then, indeed, manipulation would be purely an engineering problem and of little interest to artificial intelligence.

However, it is my thesis that a coordinated hand-eye system represents an ideal domain for extending our study of the representation of knowledge, problem solving and learning. The following sections of the paper will attempt to develop this thesis in some detail. For the moment, I would like to continue

by making a few comments on hand-eye coordination as it relates to general issues of intelligence.

0.2 Some Meta-Comments on Hand-Eye Coordination

Mature humans are quite accomplished in a wide variety of hand-eye procedures. In our culture, we become proficient at an early age with such procedures as tying a shoe, writing and using simple tools (knife and fork, hammer and nail, etc.). Certain humans become very skilled at specialized procedures such as typing, piano playing, surgery and shooting a basketball.

What are some of the obvious points that can be made about such procedures? First of all, it is certainly true that they involve continuous interaction with the environment. Feedback, of various sorts, is used to monitor and control hand activity.

Secondly, it is certainly true that such procedures are learned. They may eventually appear automatic but they first must be learned and debugged. (If anyone doubts this, get a friend to eat with chopsticks for the first time.)

Thirdly, I think that most people believe that such procedures represent various levels of learning ability. All but the severely retarded can learn to tie a shoe. Secretarial schools say that any high school level child can learn to type proficiently. Society as a whole seems to believe that only the

top fraction of a percent can learn to be skilled surgeons. (We do not train technicians to do the cutting and tying under the direction of a physician.)

Any proposed theory of human hand-eye coordination must include and account for both interaction with the environment and a non-trivial mechanism for the learning and debugging of hand-eye procedures.

1. Hand-Eye Coordination And Artificial Intelligence

The AI Laboratory is currently embarking on a major research effort in the area of advanced automation. This work in ROBOTICS is forcing us to address several new important issues with respect to a mechanical hand-eye system.

1.1 The Robot Environment

At present, the robot's model of the real world assumes that the environment changes only as the result of some discrete action performed by the robot itself. Unexpected occurences go without notice or are fatal.

However, the real world is dynamic. It does not change only in simple discrete steps. A successful industrial robot must interact both with the ongoing process it is engaged in and with secondary background changes (eg. the addition/deletion and random placement of objects in its visual field.)

1.2 A True Hand-Eye System

Aside from simple calibration procedures, the current 'modus operandum' is to place the arm behind the robot's back (figuratively, at least), look, close the eye, manipulate, put the arm behind the robot's back again and repeat as required.

However, such a scheme provides a very limitted capability for handling a dynamic environment.

A hand-eye system capable of interacting with a dynamically changing environment must be able to continuously react to real world events. It must be able to note changes in the environment and to compare these changes with its own description of the procedure it is engaged in.

Some changes would supply evidence as to the current status of the arm-hand procedure. Others might simply be noted as irrelevant. However, fundamental to the development of such a hand-eye system is the requirement to make the arm-hand an integral part of the world of the eye.

1.3 Quality Control

Perhaps the most crucial issue now facing advanced automation is that of quality control. In any complex industrial assembly procedure, there must be a means of verifying the results of previous subassemblies (including inspection of original parts) to protect against potentially disastrous consequences.

A numerical control type approach is very limitted in this area while, on the other hand, quality control is seen as an immediate corollary of the kind of hand-eye system I shall propose.

2. A Problem Domain For Studying Hand-Eye Coordination

The following is a brief scenario of a hypothetical hand-eye system:

On a table, there are several coffee cups, a coffee pot, a bowl containing sugar cubes, a small pitcher of cream and a spoon or other object suitable for stirring. There is no particular arrangement to the objects on the table. They are randomly placed within the field of view of a vidissector eye and within the reach of a mechanical arm-hand.

A human engages in a short dialogue requesting a cup of coffee in any one of its standard configurations (ie. black, cream, cream & sugar, sugar only, double cream, etc.). The arm-hand proceeds to select a cup, pour the coffee from the pot, add the required embellishments and stir the result. The human picks up his cup of coffee and says, "Thank you!"

2.1 The Features Of Such A System

1) We would be demonstrating a generalized flexibility. Since there would be no specified arrangement of objects on the table nor a fixed recipe for coffee, the robot would have to both visually locate the objects and construct a

plan as required. Further, we would like the system to be general enough to allow for the addition/deletion of cups while operation is in progress.

2) We would be exhibiting a true hand-eye system in an environment realistically approaching that of the real world. In particular, the operation of pouring must accomodate a real world that changes dynamically, not just in discrete steps. Visual feedback, with the arm-hand in the visual field, would be an essential prerequisite to accomplish accurate pouring.

3) We would be exhibiting a somewhat generalized manipulative capability through the use of simple tools — a pot for pouring and a spoon for stirring.

4) We would be facing the issue of quality control. Visual feedback must certainly be used to monitor pouring. In addition, feedback must be used to protect against pouring into a cup that's fallen over or pouring into a cup that's already full. Similarly, feedback must also be used to keep from knocking over a cup when stirring its contents.

2.2 Is This A Good "Toy" System?

The idea of a robot coffee maker probably strikes one at first as being a good demonstration. It certainly would be that. However, in considering possible alternative problem domains for a hand-eye system, I believe that the robot coffee maker is also the most appropriate.

The coffee maker environment is rich enough to support the thorough investigation and development of the various kinds of feedback tools and capabilities that would be required in any hand-eye system. The processes involved in making a cup of coffee are quite characteristic of the kinds of processes required in a generalized hand-eye system.

The primitives required to monitor the rising level of coffee in a cup are seen as essentially equivalent to those that would be required to carefully align the edges of objects in a complex assembly procdure. The primitives required to stir the contents of a cup with a spoon are essentially equivalent to those that would be required to tighten a nut with a wrench or turn a screw with a screwdriver. Similarly, the primitives required to locate a cup for pouring are essentially equivalent to those that would be required to locate a hole for inserting a bolt or screw.

Of equal significance, however, is the fact that the coffee maker environment is also simple enough to support such an

investigation with a minimum amount of time required to deal with outside issues. I believe the current vision system can easily be modified to handle the specific objects required for the coffee maker. In any event, I can immediately begin developing techniques for visual feedback by restricting myself, for the time being, to polyhedral cups and pots.

The coffee making system involves an environment that is sufficiently dynamic so as to require a degree of interaction that would constitute a significant advance over previous work in machine hand-eye coordination. The primitives developed for the coffee maker would be applicable to a host of other hand-eye tasks. At the same time, the coffee maker represents a problem domain that is very accessible and manageable given the current status of the MIT vision system.

There are a number of subproblems that need to be solved in order to support such a coffee making system. In what follows, I give my initial thoughts on these subproblems.

3. The Cup

As a part of my 6.544J termpaper {13}, I conducted a simple experiment to study human hand-eye coordination. The experiment consisted of throwing simple objects into a wastepaper basket at various distances. The rim of the wastepaper basket was covered with strips of adhesive tape coated with luminous paint. The experiment was conducted in a photographic darkroom.

Under darkroom conditions, the visual world was totally dark except for the fine-grained, uniformly lit elliptic ring seen as the projection of the rim of the wastepaper basket. Even under these conditions, the experimental subjects were able to determine the location and orientation of the wastepaper basket sufficiently well for accurate throwing.

The elliptic projection of a circular surface conveys a great deal of information. Much about a cup could be specified to a hand-eye system simply in terms of the elliptic image of its rim. Assuming only that the eye is elevated with respect to the table, the vidissector would see the rim of a standing coffee cup as an elliptic ring. The eccentricity of this ellipse can be used to determine the elevation angle θ of the eye with respect to the table.

Consider figure 1. The major axis a of the elliptic image is formed directly by the diameter of the cup. The minor axis b of the elliptic image is formed by the perpendicular distance between rays of light reaching the eye from points P and R.

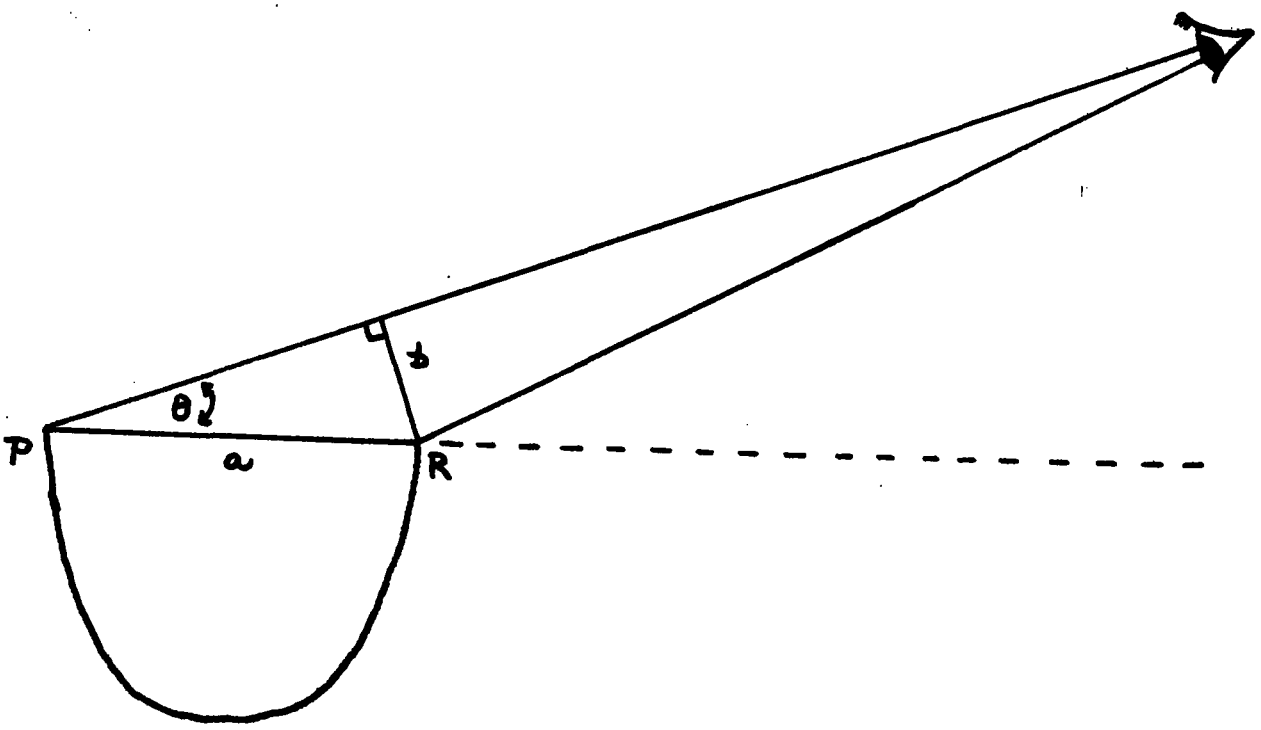

figure 1.

Thus, the elevation angle θ satisfies the relation

$$\sin\theta = b/a$$

so that

$$\theta = \arcsin(b/a)$$

In an ideal optical system, the image size S_i, the object size S_o, the focal length of the lens f, and the distance of the object from the lens d are related as follows:

$$\frac{S_i}{f} = \frac{S_o}{d}$$

If S_o is the known diameter of the cup, if f is the known focal length of the vidissector lens and if S_i is the length of the major axis a as measured from the vidissector data, then the distance of the cup from the eye becomes

$$d = (S_o f)/S_i$$

This distance d plus the elevation angle θ could be used to determine the location of the cup with respect to any arm-hand coordinate system (provided the position of the arm-hand "shoulder" is known relative to the eye).

However, in practice, it is to be expected that small errors in the measurement of S_i (due to imperfect determination of the elliptic image, imperfect focus, etc.) would induce large errors in the measurement of d. The distance, d, as computed in this manner, would provide only a crude approximation to the location of the cup. When required, visual feedback based upon the approach of the arm-hand will be used to refine the determination of the cup's position (see section 4.). But, unless you want to do something with the cup, there's no real need to know its position all that accurately.

Currently, the MIT vision system avoids making use of objects of known size to determine object distance, preferring instead to use optimal focus and/or the horizontal plane hack. I have no particular axe to grind in this regard. The above is only to indicate that a crude estimate of d can easily be obtained.

However, the determination of the eccentricity of the elliptic image of the rim of the cup has other uses. If the elevation angle θ of the eye with respect to the table is known (as presumably is the case with humans), the eccentricity of the elliptic image can be used to determine if the cup is upright or if the cup has fallen over. (A degeneracy can occur if the cup

has fallen over away from the eye.)

4. The Hand

4.1 How To Control The Hand?

In any hand-eye system, an interesting dimension to consider is the extent to which the hand is driven using absolute coordinate calculations versus the extent to which the hand is driven using relative coordinate calculations (ie. feedback).

Currently, the vision system is used to obtain the absolute 3-D coordinates of points of interest. The 2-D image of the scene provides two constraints while a hack (optimal focus, horizontal plane, range finding, etc.) provides the third. In such a system, the hand can be directed without the use of visual feedback. However, such a scheme depends upon both a highly reliable vision system (no errors in support hypotheses, etc.) and a highly accurate arm-hand (ie. it gets exactly where we told it to go).

In any hand-eye system, it is quite easy to determine the absolute 3-D coordinates of the arm-hand's position in terms of the orientation parameters of the various joints. However, the difficult problem is, given a point in 3-space, what joint coordinates will place the arm-hand at that point. The solution to this problem generally involves inverting the large matrix representing the coordinate transformations through the various joints.

In general, the transformation matrix will have singularities. Even if the matrix can be inverted, there is no

guarantee that the simplest mathematical solution is the most aesthetically pleasing.

Recently, two systems (Wichman {9}, Shirai {7}) have been developed which use visual feedback to obtain relative coordinate information to correct the inaccuracies of hand movement. These systems still use vision to obtain absolute 3-D coordinates and only use visual feedback (relative coordinates) to fine tune the result.

I would like to propose the idea of controlling the coffee maker hand as much as possible by the use of visual feedback. In some sense, we would still be faced with the same problem as before. The 2-D image of the scene (including the hand in the field of view) provides only two constraints as to the position of the hand relative to an object. We still require a third constraint.

4.2 Using Feedback To Obtain The Third Constraint

Very little is actually known about how a human coordinates his hand and eye in 3-space. Although we do not know enough about neurophysiology to deny it, it seems unlikely that the human nervous system inverts large coordinate transformation matrices. We do know that a human possesses a number of redundant sources of feedback. (All six human proporioceptive systems could easily be involved in a human coffee maker.) At the

same time, we know that vision alone can support human hand-eye coordination. If the incoming sensory pathways from the muscles and joints of the human arm-hand are interrupted at the dorsal spine root, the movement of the arm-hand can still be voluntarily controlled by looking at it, so long as the outgoing motor pathways remain intact.

Although we have been reluctant to use objects of known size in vision research, there is no reason why the robot system should not know as much as required about its own arm-hand. One could tape a ruler to the robot finger. More subtlely, one could tape circular markers to the robot fingers. In the same fashion as illustrated for the cup, these markers could be used to obtain both orientation and crude depth perception for the arm-hand.

In section 3., a crude method for locating cups was introduced. Another such method involves adjusting the vidissector lens to obtain optimal focus. A third method, called the horizontal plane hack, makes use of the previously determined equation of the table plane to locate feature points that are known to touch the table (or to be a known height above the table).

Once crude estimates of the position of the arm-hand and the object of interest are obtained, a simple hill-climbing procedure based on the movement of the arm and of the shadow cast by the arm can be used to locate the object for grasping, pouring, etc.

4.3 The Rote Learning Of Motion Primitives

Observations with children seem to indicate that new hand-eye procedures are learned and debugged by carefully controlling and monitoring the introduction of new degrees of freedom. A young child learns to drink a glass of milk by first concentrating all motion in the shoulder joint. At this stage, the child is unable to lift a glass to its mouth while at the same time holding it level. A young mother soon learns to maintain the level of milk in the glass low enough so that the meniscus of the liquid reaches the edge of the glass only when the glass reaches the child's mouth. Gradually, the child begins to introduce freedom in elbow and wrist movement. By carefully observing the effects of each new degree of freedom, the child is able to learn the complex procedure of motion through space without tilt or rotation.

In much the same way, I believe that the robot can be made to learn such complex motions as stirring and pouring. Rather than go through complex arithmetic calculations involving all possible degrees of freedom, the robot can use visual feedback to note the effect of controlled degrees of freedom. These observations can be used to refine its approximation to the desired motion through space.

The kind of learning required here would be very simple.

The robot would have good descriptions of what it means to stir and pour. Visual feedback would be used to apply these descriptions to the particular stirring or pouring task at hand.

5. Visual Feedback During Pouring

Any mechanized system that attempts to pour materials from one container to another can most effectively make use of visual feedback. It is highly unlikely that one would want to consider the alternative of having accurate sensors in the arm-hand detect the loss of weight from the source container and then use specific gravity calculations to determine the volume of materials poured.

Visual feedback can be used both to determine when the required volume of materials has been poured and to adjust the actual rate of pouring.

Let us assume that our cup is a light color and that the coffee is dark. The following represents a snapshot of the cup as the coffee is poured into it:

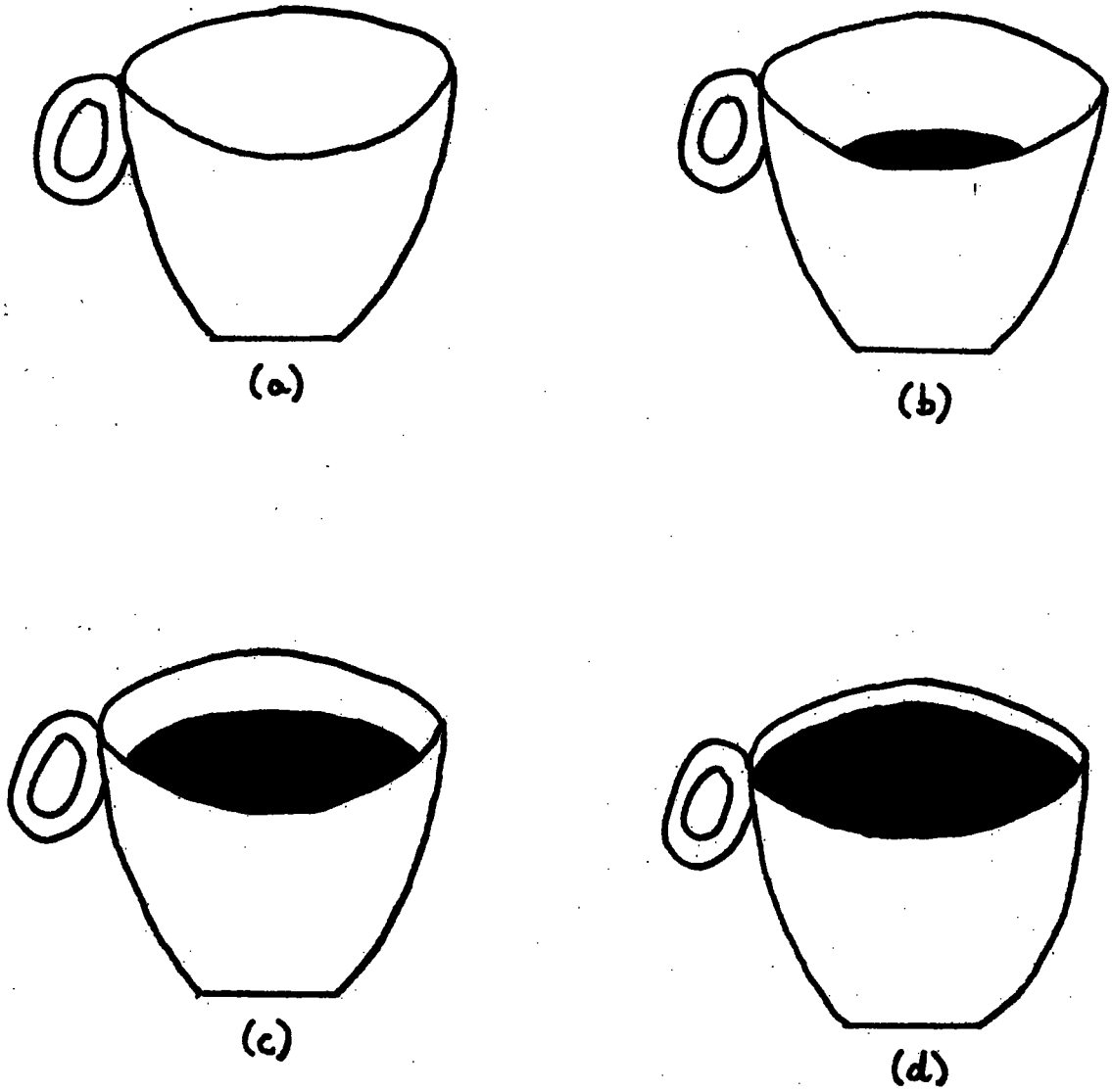

figure 2.

A detailed real time analysis of vidissector data along the minor axis of the image of the cup's rim can be used to determine both the absolute level of coffee in the cup and the rate at which the height of the coffee in the cup is rising. When the cup is nearly empty, visual feedback can be used to

permit the arm-hand to pour quite rapidly. As the cup begins to fill, visual fedback can be used to slow the rate of flow to protect against overshooting and spillage.

This feature of using visual feedback to control the rate of pouring is quite important since we would like to allow the coffee pot to be full, half full, or nearly empty. No fixed pouring action would be appropriate in all cases. Finally, visual feedback of this sort would also allow us to determine when the coffee pot was empty (ie. when no more coffee can be poured out).

6. The Use Of Tools

I do not view the robot's use of simple tools as a separate issue in itself. Rather, I view it as a test of the generality of the robot's model of its own arm-hand.

When humans make use of simple tools, they possess a rather remarkable ability to incorporate those tools as extensions of themselves. Consider the example of tapping a pencil on the table. In a very real sense, one experiences the touching of the table at the point where the pencil touches the table, not, as is the real case, in the fingers where they touch the pencil.

In the case of a robot coffee maker, I see no essential difference between the robot's stirring the contents of a cup with its finger or with a spoon. If the model of the arm-hand allows a stirring motion at all, it should be general enough to allow the arm-hand to stir with a spoon. Similarly, if the model of the arm-hand allows the hand to rotate about a point, it should be general enough to allow the hand to rotate about the point representing the tip of the coffee pot spout (ie. pouring).

Thus, although the proposed system will appear to be demonstrating a somewhat sophisticated use of tools (coffee pot, spoon, etc.), in reality, it will be demonstrating a generalized model of its own arm-hand. The use of visual feedback to control the operations of pouring and stirring (as outlined in section 4.) is immediately generalizable to arbitrary pots and spoons.

7. Summary

The work proposed above is to be carried out using the facilities of the MIT Artificial Intelligence Laboratory consisting of the PDP-6/PDP-10 computer system together with the mechanical arm-hand and the vidissector eye.

At the moment, I am considering the work to consist of five major phases which are to be attacked in the following order:

Phase I

Phase I consists of developing the LISP functions and predicates required to monitor the pouring of liquid into a cup. In particular, I propose to implement the following:

> On the table, there is a standing, empty polyhedral cup. Slightly above the field of view of the vidissector, there is a human holding a coffee pot. The human slowly begins to pour coffee into the polyhedral cup. When a specified level has been reached in the cup, the machine rings the bell on a teletype (or some such thing) and the human stops pouring.

Phase I is seen as an effort to develop and test the LISP primitives in real-time. I feel that it is important to become

aware of the real-time capabilities (and lack of same) of our system at an early stage of the work.

Phase I will make almost exclusive use of the current vision system so that the proposed LISP primitives can be quickly implemented and tested.

Phase II

We require the ability to perform such procedures as rotating, at variable speed, about an arbitrary point in space (pouring) and rotating in a small circle in a horizontal plane (stirring). As a first pass at this problem, phase II will consist of monitoring a target held, in inverted lollipop fashion, by the arm-hand.

Under visual control (specifically, using hill climbing on the convergence of shadows and feature points in the 2-D image), we will attempt to touch points in the visual world with the target. (eg. direct the target to touch the corner of a cube)

Once such capabilities exist, they can be extended, for example, into the generation of arm-hand motion appropriate for the stirring of the contents of a cup with a spoon.

Phase III

Phase III will consist of the extention of phase II to more complex motions based upon an attempt to incorporate the rote learning of sequences of motion primitives.

Using the motion primitives developed in phase II together with higher level descriptions of stirring, pouring, etc., we will attempt to "learn" appropriate sequences of arm-hand motions to accomplish the required tasks. Visual feedback will be used to monitor and criticize the first pass attempts at each new motion.

Phase IV

Phase IV will consist of integrating the capabilities developed in phase III with those developed in phase I.

At this point, we should have a system capable of pouring and stirring coffee without human assistance.

Phase V

Phase V is seen as the open-ended attempt to add additional features to the system. Some of these, as suggested above, would be:

(i) the ability to detect whether a cup is already full and hence not use it.

(ii) the ability to detect whether a cup has been knocked

over and right it as required.

(iii) the ability to add sugar cubes and cream before stirring.

(iv) primitive obstacle avoidance (ie. knowing where things are) as exemplified by allowing the random addition/deletion of cups while coffee making is in progress.

(v) the use of cylindrical (as opposed to polyhedral) cups.

References

{1} Clowes, M., "On Seeing Things," Artificial Intelligence Journal, Vol. 2, No 1, Spring 1971.

{2} Dowson, D., "What Corners Look Like," Vision Flash 13, AI Laboratory, MIT, June 1971.

{3} Dowson, M., Waltz, D. L., "Shadows and Cracks," Vision Flash 14, AI Laboratory, MIT, June 1971.

{4} Guzman, A., "Computer Recognition of Three-Dimensional Objects in a Visual Scene," MAC-TR-59, thesis, Project MAC, MIT, December 1968.

{5} Horn, B. K. P., "The Binford-Horn Line Finder," Vision Flash 16, AI Laboratory, MIT, June 1971.

{6} Huffman, D., "Impossible Objects as Nonsense Sentences," Machine Intelligence 6, (ed. Meltzer, B. & Michie, D.), Edinburgh University Press, Edinburgh, 1971.

{7} Shirai, Y., "Guiding a Robot by Visual Feedback in Assembling Tasks," Electrotechnical Lab., Tokyo, Japan.

{8} Waltz, D., Doctoral Dissertation and AI Technical Report in preparation, AI Laboratory, MIT.

{9} Wichman, W., "Use of Optical Feedback in the Computer Control of an Arm," AI Memo 56, AI Project, Stanford University, August 1967.

{10} Winograd, T., "Procedures as a Reprsentation for Data in a Computer Program for Understanding Natural Language," MAC-TR-84, thesis, Project MAC, MIT, February 1971.

{11} Winston, P. H., "Learning Structural Descriptions from Examples," AI-TR-231, AI Laboratory, MIT, September 1970.

{12} Winston, P. H., "Wandering About the Top of the Robot," Vision Flash 15, AI Laboratory, MIT, June 1971.

{13} Woodham, R., "Towards a Theory of Hand-Eye Coordination," 6.544J termpaper, MIT, May 1972.