THE STANFORD HAND-EYE PROJECT

by
J.A. Feldman, G.M. Feldman, G. Falk, G. Grape, J. Pearlman, I. Sobel, J.M. Tenebaum

Computer Science Department
Stanford University
Stanford, California

There is a large continuing project at Stanford Artificial Intelligence Laboratory aimed towards the development of a system capable of interesting perceptual-motor behavior. This paper presents a brief outline of the currently active efforts and suggests references for more detailed information. A more thorough discussion of the effort to organize a visual perception system is presented.

Acknowledgement

The work reported here depends on efforts by many people besides the authors and is supported in part by the Advanced Projects Agency of the Department of Defense under Contract SD-I83.

1. An Overview

The work on integrated "robots", though widely discussed, is still poorly understood even within the Artificial Intelligence community. Like any large research and development effort, a robot project is difficult to describe in a way which is comprehensive but not superficial.

In this paper we will attempt to provide an overview of our current goals and approaches to achieving them. These are a number of detailed papers on various aspects of the Stanford effort which are referred to here, including several works in progress. These latter are included because they should be available from their authors considerably in advance of their formal publication.

The overall goal of the hand-eye project is to design and implement a system which exhibits interesting perceptual-motor behavior. An important subgoal is that the problems that arise in the design of system components be solved in ways which are sufficiently general to be scientifically interesting. Thus, for example, we have put considerable effort into understanding depth perception although the special environment we are using allows for ad hoc solutions. The possible applications of our work and its relevance to the study of animal behavior have been secondary areas of interest.

Our first hand-eye system used many ad hoc solutions and was mainly concerned with the problems of combining the minimum necessary hardware and software components. This primitive, but complete system for block-stacking under visual control was completed in May I967, and has been described elsewhere [21]. The functional diagram of Figure 1 provides a sufficient description for our purposes. Our most recent work has involved the redesign of the system configuration and more careful study of each of the component programs.

Our attempt to develop an integrated hand-eye system has forced us to confront several AI problems which had received little previous attention. The two causes underlying the new problems are the complexity of the desired behavior and the innate perversity of inanimate objects. Pattern recognition, problem solving, modeling, etc. which have been studied in idealized contexts take on new aspects in the hand-eye system. The most striking result to date is that traditional approaches to these problems have not proved adequate. We are not yet in a position to make definitive statements on what is needed, but a common understanding of the issues is arising among robot builders. Nilsson's paper in this volume 14] contains a good discussion of the general situation.

The main principle which has emerged from the Stanford work is the dependence of everything on everything. For example, one might use entirely different perceptual strategies with a random access (image dissector) camera than with a scanning (vidicon) device. This inseparability contributes to high entrance cost of hand-eye research; there is, as yet, no way to experiment with a part of the program without detailed knowledge of the other parts.

Much of our effort has gone towards reconciling this mutual interdependence of programs with the inherent independence of programmers. The problem is exacerbated at a university by the need of graduate students to produce clearly separable contributions to the project.

These facts, plus the availability of systems-oriented students, encouraged us to undertake a rather ambitious system-programming project including a submonitor, a high-level language, and a new data structure. The goal of this project is to produce a hand-eye laboratory in which it will be relatively easy to experiment with new ideas in perception, modeling, problem-solving and control. This laboratory will also, hopefully, provide a testing ground for many related artificial intelligence projects currently underway; Section 3 contains a discussion of some of these.

The hand-eye laboratory will have to accomo-

date programs whose total size is several times the size of core memory. Further, as we will show in Section 2, the order in which these programs are executed cannot be determined in advance. These programs must be able to communicate with each other and with a common global model which represents the system's knowledge of the world. Since many operations require moving physical devices (like the arm and camera) which entail long delays, we would like to allow parallel execution of hand-eye subprograms. All of these requirements can be met by the addition of one basic feature, the pseudo-teletype, to the PDP-10 time-sharing monitor. A pseudo-teletype is simply a buffer set up by one job which acts as the control console of another job. Subprograms are each set up as a separate job; all active, jobs will be automatically time-shared by the main monitor. The submonitor is responsible for handling messages, some interrupts and changes to the global model and will also be able to record its actions as an aid to debugging the system.

The language and data-structure designs are closely tied to the submonitor and to each other. The language is an extension of our ALGOL Compiler [27] along the lines of the associative language, LEAP [5]. The central concept of LEAP and the underlying data structure is the association: attribute • object = value. The use of associations for world-modeling is described in detail in [16]. An important new concept in this version of LEAP is the use of local and global associative structures. Every atomic object (item) is either local or global; the associative structure local to a subprogram may contain associations including global items, but not vice-versa. Any attempt to alter the global associative structure is trapped to the submonitor which determines when the alteration should be allowed. The language contains primitives for local and global associations, message handling and interrupt processing. Preliminary versions of the submonitor, language, and data-structure are currently in operation and seem to be providing the desired increase in programming ease.

Work on the hand-eye problem proper has continued in paralled with the system development. Much of this work has been directed toward the development of a flexible set of vision programs, the subject of Section 2 of this paper. To provide a sense of direction and to bound our aspirations, we proposed a class of tasks which we hope to have the hand-eye perform. The main task is the building of fairly complex constructions (castles) out of simple blocks. The blocks were restricted to being plane-bounded and convex. The castle might be explicitly described by a set of associations relating its sub-parts or we might simply be given one or more views of it. Even this task is too difficult for the system to solve in general, but it has provided a useful context for the development of various routines.

Building a castle out of children's blocks is a problem in which there is no technical literature. Shapiro [30] has concerned himself with the development of optimal strategies for doing this with our mechanical hand which can only place a block with *l/k* inch accuracy. The first problems attacked were the development of heuristics for stability analysis and for generating the proper sequence of actions assuming ideal placement. Subsequent work will remove this restriction and attempt to develop strategies which compensate for observed imperfections in the performance of the mechanical manipulator. One of the most interesting aspects of this task is the various levels of feedback which can be used in the building process. In some cases, one need only know that a block is still in place and tactile feedback is sufficient. If the situation is more critical one might visually determine the placement error and alter the remainder of the strategy accordingly. Finally, there is the possibility of adjusting the block, under visual control until the error is sufficiently small [29]. An important part of the castle building problem has been solved by Pieper [17] in his development of an obstacle avoidance program for the arm.

The use of visual feedback in block stacking presents a rather different problem than those normally discussed in picture processing. The vision routine has the job of determining the accuracy with which some block was placed. The total scene may be very complicated and it would be absurd to perform a complete scene analysis. Furthermore, the properties of the blocks to be examined may be known in great detail and the vision routine would be able to take advantage of this fact. One of our major efforts has been directed toward solving these problems of context-sensitive visual perception. The overall system designed to do this is quite complex and is the subject of the next section.

2. The Organization of a Visual Perception System

Perception, and most particularly visual perception, is a complex process requiring a system which is sensitive to all the various levels of detail of the environment. Furthermore, since the available data is potentially overwhelming (consider the number of different viewpoints) the system must have both the mechanisms and appropriate strategies to select what data are worthy of its attention and what level of detail is best suited to the current perceptual goal.

We will concentrate on these two aspects of visual perception - levels of detail and strategies for attention. Data from a scene may be structured to varying degrees. At the lowest level lie the intensity and color of the light at a particular point in the visual field; at a higher level are those objects in the visual scene which we dignify by the use of nouns; at a still higher level one notices interrelationships and relative motion between objects. At the highest level one is aware of the total situation - as "Danger! Collision imminent." Each of these levels of perception is necessary and we must integrate all of them. Ordinarily, we are conscious only of our perceptions of objects and

situations, but the fact that we can learn to draw indicates that lower level details are perceived and can be made accessible to consciousness.

It is curious that we must learn to draw - as if the lower levels of visual patterns are coalesced into objects at a preconscious level. This notion gives rise to a simplified theory of perception held by many workers in perception and pattern recognition. The theory is embodied in a strategy of perception which places attention first at the lowest level of detail and then extracts successively higher levels until the organization of the entire scene is understood. Thus, by processing intensity and color distributions one obtains texture, edges, and corners. From this information regions are extracted and these in turn are associated into bodies. Then the bodies are identified as objects and their various interrelationships are derived. Thus:

points -> lines -> regions -> bodies -> objects ->

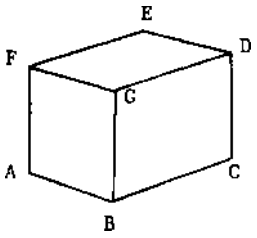
scene

Essentially, all the early work on visual perception, including our own, proceeded along these lines. To some extent, the beautiful work of Guzman [9] °n finding the distinct bodies in a perfect line drawing had an undesirable effect on the field. Guzman's program was so successful that it sent people on a quest for the perfect line drawing program. Although we have had considerable success [7,11] at generating line-drawings, it has become apparent that the strict bottom-to-top processing sequence is not optimal. We will present some general discussion on the organization of vision systems and then describe our current efforts.

The model of vision which we fine useful involves routines at various levels, cooperating in an attempt to understand a scene. There is a large body of psychological evidence [6,32] indicating the dependence of perception upon global information and upon preconceived ideas. Many of the well known optical illusions fall in this class. One can also show that there are simple scenes which are ambiguous in the absence of global information, but are easily resolved in context.

A most striking case of this is the ground plane assumption [23], which has become a cornerstone of all robot perceptual systems. From a monocular image it is impossible, in general to calculate the distance of an object from the camera. If, however, the object is lying on a known plane (one whose transformation to image coordinates is available) then the depth of the object's base vertices is known. This particular piece of global information has been implicitly used for depth information, but has many other uses.

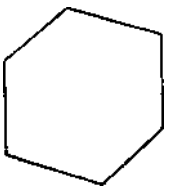
Consider the following line drawing.



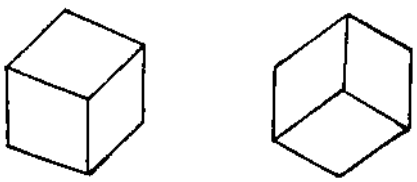
If one knew that this object were lying on the plane determined by ABC which was known, then one would know the projection of each point in the image onto the ABC plane. Each point e.g. F must be on the line determined by its projection onto the ABC plane and the lens center. If the line AF is perpendicular to the plane we then know the length of AF.

Further, we can often determine whether or not AF is perpendicular to the plane from the information available. The lens center, point A and the projection of point F determine a plane, which contains the line AF. If this plane is perpendicular to ABC then the line AF is also, for objects which are at all regular [26]. If one knew the lengths of AF, BG, and CD and their angles with the ABC plane, then the coordinates of F, G, and D are computable and assuming F, G, D and E are in a plane is sufficient to determine E. Thus the ground plane assumptions plus some global regularity conditions allow for the complete description of an object from a single monocular view. Of course, these conditions may not hold, but Falk has some encouraging results in object recognition using these kinds of techniques .

A somewhat more basic problem arises in the consideration of the following image:



which might have come from, among other things:



The interior edges might very well be less distinct and be missed by the program which first tried to form a line drawing. At some higher perceptual level, a program could detect the ambiguity and attempt to find the interior edges. With the contextual information available, the

system could then use highly specialized tests to determine the presence of an edge. Further, since the area involved is relatively small, it might also be reasonable to apply very sensitive operations which are too costly to use on an entire scene. In both cases we see how our system organization facilitates a perceptual strategy involving selective attention. A vision system which worked strictly bottom-to-top would have no notion of attention. There would be a standard line finding operation, followed by an attempt to fit intersections, etc. These are inherent limitations [28] in any such system balancing noise sensitivity with ability to perceive detail. The flexible organization discussed here allows for the use of different hardware and software components in different contexts and has much greater potential.

Those readers unfamiliar with the field will probably feel that we have set up an elaborate straw man. Cognoscenti will recognize the man as real enough, but will be looking for a way to make our grand design operational. The remainder of this section will be devoted to a discussion of how we are attempting to do this.

The goal, once again, is to produce a flexible visual perception system capable of selective attention and of integrating information from all levels of perception. An obvious prerequisite for such a system is a monitor, language, and data structure capable of its support. Our proposed design was described in Section 1.

A second necessary ingredient of any such system is a large set of flexible basic vision routines. Among the necessary functions are: reading raw data, changing the camera position and parameters, edge finding, corner fitting, region finding, analysis into distinct bodies, identification of particular objects, and complete scene analysis. Work is under way in all these areas but we will be content to describe briefly some of the work which seems to be most interesting.

One important aspect of the general vision system is accomodation, the adaptation of the input mechanisms to the visual environment. Selective attention can then be implemented in the vision hardware by choosing accomodative strategies which reflect current perceptual goals. For example, the camera could be sensitized to a specific color characteristic of a desired object (via a color filter). This effects a gross reduction in the volume of information which must be input and subsequently searched to determine its relevance.

The camera parameters currently under computer control are the pan and tilt angles, focus, magnification and digitization level. There are two hard problems in accomodation which arise from the need for a common world model. When the camera is panned, it gets a new view. The images of objects in this new view must be placed in correspondence with the old Images of the same objects. An even more difficult problem is to compute accurately the perspective transformation [23] applicable in the new situation. Sobel [26] has developed techniques for these problems, relying heavily on the literature of photogrammetry.

A major area of interest at Stanford has been the development of low-level edge and line finders The visual system of the original system was little more than a good edge follower pus a routine which used the ground plane assumption and the existance of only cubes to locate objects. There have been extensive analytical and practical studies of various spatial filtering and edge finding techniques [11, 28]. More recently, we have begun to look at feature verifiers which will use global information and a prediction to help identify a feature.

There are also programs which do fairly well at corner finding, region extraction, etc. These are fairly flexible and might be incorporated into a vision system organized as we have suggested. The real problem is to develop routines for these tasks which are sensitive to possible errors and ambiguities and know when to ask for help. A related issue is the language for communicating between vision programs at various levels. We have just begun to seriously confront these issues

We are currently completing an interactive version of our grandiose vision scheme. Grape is extending his programs [7] to allow for user intervention at several stages in the scene analysis process. As intermediate stages of analysis are displayed, the user will be able to interrupt and add information to the system. Using this system and some hard thought, we hope to come up with a reasonable first cut at the multi-level vision system. The process of refining this system and adding to its basic capabilities will, like the poor, always be with us.

3. Related Work in Artificial Intelligence at Stanford

The robot problem, in some sense, encompasses the entire field of artificial intelligence - there is nothing in artificial intelligence work which would not be useful in the ultimate robot. The precise degree to which various other efforts should be coordinated with a robot project is unclear. Traditionally (for the past three years), the M.I.T. group has kept quite strictly to hand-eye problems which the S.R.I, group has concentrated on combining as much of its work as possible. The Stanford group is somewhere between - there are a large number of artificial intelligence projects at varying distances from the hand-eye effort.

One closely related development is concerned with improvements in the devices used for the mechanical hand and eye. The research or, vision devices has been largely analytical [3] but consideration is being given to building a laser system which will directly produce a three-dimensional image. The work on arms and hands is conducted largely by the mechanical engineering department and has been rather more active. This

effort has produced one dissertation [17], two complete arm systems, and a variety of proposals for others [25]. In the visual perception area, there are attempts to solve such problems as face and person recognition. There is also a significant effort underway to operate a motorized cart under computer control. The cart and its sensing devices are operational and the programming for this task has begun. Although the cart project will use some of the vision routines developed in the hand-eye effort, its goals are quite different. The main problems being attacked in the cart project are vision from a moving object and the related problems of control. This project is expected to grow considerably in the near future.

The most relevant of the many theoretical efforts is the work on the use of automatic theorem proving methods as a technique for building strategies [12]. Some such mechanisms will eventually be part of the hand-eye system and there are efforts to axiomitize some hand-eye tasks. However, there are very difficult theoretical and practical problems to be solved before a theorem prover will be able to develop strategies as flexible as the one for castle building described in Section 1.

The work on systems programming discussed briefly in Section 1, contains a number of interesting problems in its own right. The use of many parallel programs operating on a single global data structure is a problem of considerable current interest. Even more intriguing is the possibility of problem-directed resource allocation. The control program for a particular hand-eye task will attempt to choose an optimal sequence of vision, manipulation and computation routines for achieving its goal. It seems reasonable that such a control program could allocate resources (core, processor, etc.) better than a blind scheduling algorithm; we are designing the system to allow for experimentation along these lines.

Certainly one would like the ultimate robot to communicate with people in natural language. There is a large effort under Colby [3] to develop models of human belief structures and programs which can construct these belief structures from natural language statements. Another important continuing effort is that of Reddy [22, 3I] on speech recognition. This work has been quite successful and has actually been combined with the original hand-eye system in a demonstration program. Much more elaborate natural language communication systems for hand-eye could be produced if there were a scientific advantage to be gained.

One project in natural language processing which seems particularly relevant is that of Becker [1]. He is developing a model of human cognitive structure which attempts to encompass both perceptual and verbal behaviour. Currently in its early stages of development, this model may become a serious contender for the basis of the general problem solver in the next generation robot.

As these projects and the hand-eye system develop, we expect them to have an increasing effect on one another. The remaining problems are immense, but the entire approach seems more sound and realistic than was the case a few years back.

References

1. Becker, J., "The Modeling of Simple Analogic and Inductive Processes in a Semantic Memory System", Proceedings of the 1st International Congress on Artificial Intelligence, Washington, I969.

2. Colby, D., and Enea, H., ''Heuristic Methods for Computer Understanding of Natural Language in Context-Restricted On-Line Dialogues," Mathematical Biosciences 1, 1-25 (1967).

3. Earnest, L.D. (I967). "On Choosing an Eye for a Computer". AI Memo No. 51, Stanford University, Stanford, California.

h. Ernst, H.A. (I96I). "MH-1, a Computer Operated Mechanical Hand". Doctoral Thesis, M.I.T., Cambridge, Massachusetts.

5. Feldman, J.A., and Rovner, P.D., "An Algol-based Associative Language", AI Memo No. *66*, Stanford University, Stanford, California.

6. Gibson, J., The Senses Considered as Perceptual Systems, Boston, Houghton-Mifflin, 1966.

7. Grape, G.,"Untitled, I969", forthcoming.

8. Green, C, "Theorem Proving by Resolution as a Basis for Question Answering Systems", Machine Intelligence 4, American Elsevier, 1969.

9. Guzman, A., "Decomposition of a Visual Scene into Three-dimensional Bodies", Proc. FJCC, I968, p. 29I-304.

10. Guzman, A., "Some Aspects of Pattern Recognition by Computer", MAC-TR-37, Project MAC, M.I.T., Cambridge, Massachusetts.

11. Hueckel, M., "Locating Edges in Pictures", forthcoming AI Memo, Stanford University, Stanford, California.

12. McCarthy, J., and Hayes, P., "Some Philosophical Problems from the Standpoint of Artificial Intelligence", AI Memo No. 73, Stanford University, Stanford, California, November I968 (to appear in Machine Intelligence 4, American Elsevier,

13- McCarthy, J., Earnest, L., Reddy, R., and Vicens, P., "A Computer with Hands, Eyes, and Ears," Proc of FJCC '68 (I968).

Ik. Miller, W.F., and Shaw, A., "A Picture Calculus" in Emerging Concepts in Graphics, University of Illinois Press, I968.

15. Nilsson, N., "A Mobile Automaton", Proc, 1st International Congress on Artificial Intelligence, March 1969.

16. Paul, R., Falk, G., Feldman, J., "The Computer Representation of Simply Described Scenes", Proc. Illinois Graphics Conference, April, I969.

17. Pieper, D., "The Kinematics of Manipulators under Computer Control", AI Memo No. 73, Stanford University, Stanford, California, (Dissertation in Mechanical Engineering).

18. Pingle, K., "Hand-eye Library File", Artificial Intelligence Operating Note No. 35, Stanford University, Stanford, California, August I968.

19. Pingle, K., "A List Processing Language for Picture Processing", Artificial Intelligence Operating Note No. 33, Stanford University, Stanford, California.

20. Pingle, K., "Visual Perception by a Computer", Proc. Summer School on Automatic Interpretation and Classification of Images, Pisa, Italy, August I968.

21. Pingle, K., Singer, J.A., and Wichman, W.M. (I968), "Computer Control of a Mechanical Arm Through Visual Input",Proc. IFIP Conference, Edinburgh, I968.

22. Reddy, D.R., "On the Use of Environmental, Syntactic, and Probabilistic Constraints in Vision and Speech", AI Memo No. 78, Stanford University, Stanford, California, I969.

23. Roberts, L.G., (I963). "Machine Perception of Three-Dimensional Solids". Optical and Electro-Optical Processing of Information, MIT Press, Cambridge, Massachusetts.

24. Shapiro, G., "Advanced Hand-Eye Manipulating", Internal Memo, Stanford AI Project, Stanford University, Stanford, California.

25. Scheinman, V., "Considerations in the Design of a Mechanical Arm", Internal Memo, Stanford AI Project, Stanford University, Stanford, California.

26. Sobel, I., "Visual Accommodation in Machine Perception", forthcoming.

27. Swinehart, D., "Golgol III Reference Manual", Artificial Intelligence Operating Note *48,* Stanford University, Stanford, California .

28. Tenebaum, J., "An Integrated Visual Processing System", forthcoming.

29. Wichman, W.H., "Use of Optical Feedback in the Computer Control of an Arm", Engineers Thesis, Stanford University, Stanford, California, August I967.

30. Thompson, M., "Manual of Photogrammetry", 3rd Edition, I966.

31. Vicens, P., "Aspects of Speech Recognition by Computer ", Dissertation in Computer Science, Stanford University, March I969.

32. Yarbus, A.L., "Eye Movements and Vision", Plenum Press, New York, I967.

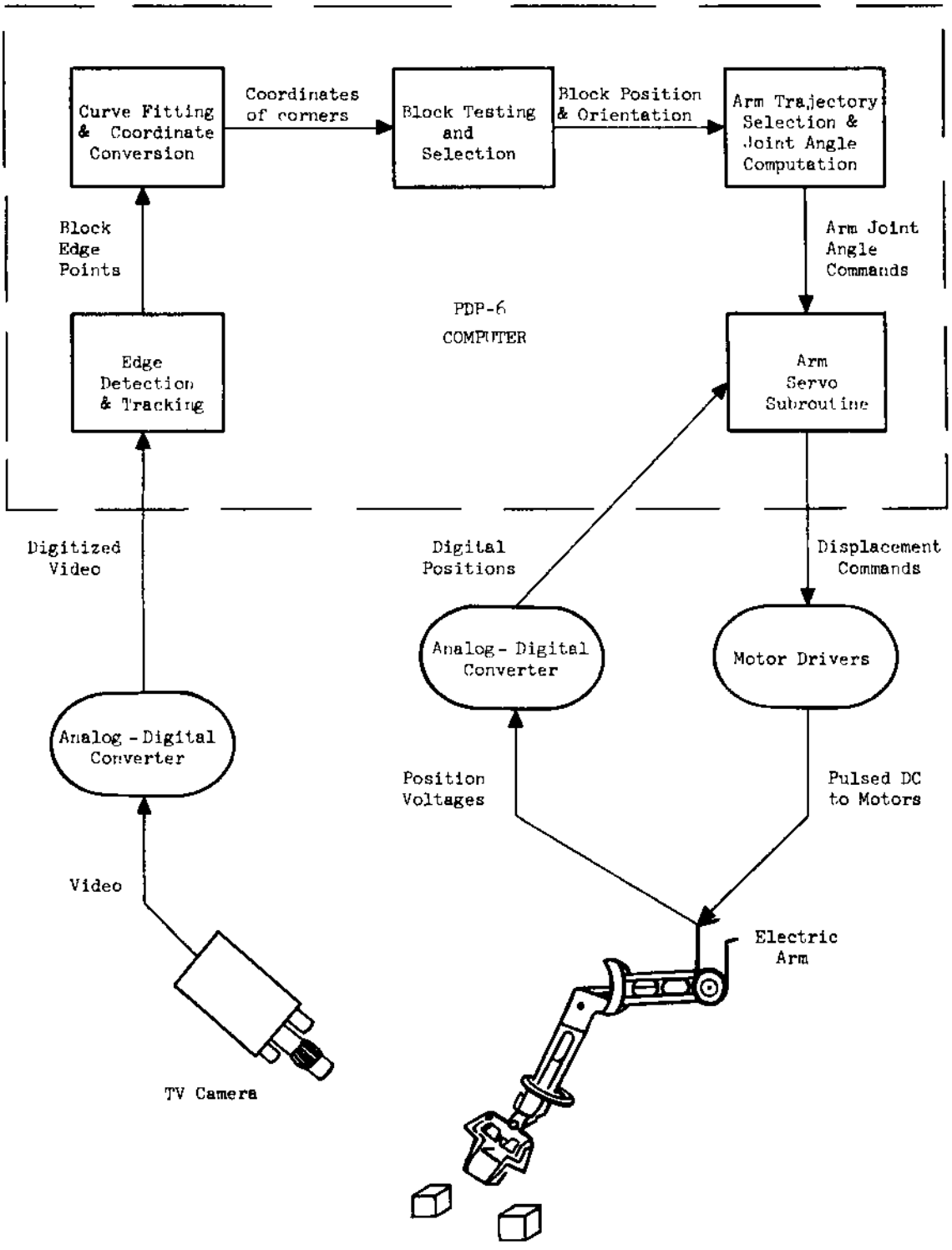


Figure 1. The Initial Block-Stacking System