THOUGHTS ABOUT A VISUALLL GUIDED GRASP REFLEX

Richard L, Didday
Department of Mathematics and Computer Science
Colorado State University
Fort Collins, Colorado

Abstract

This paper describes a preliminary attempt to apply organizational principles taken from animal perceptual processes to the problem of providing a robot visual qrasp reflex. The goal is to provide a cheap, rapid mechanism for directinq a hand/arm to grasp one of several arbitrarily shaped objects in the visual scene. A proqram which implements a simple form of such a reflex is described.

Introduction

The structure of animal nervous systems was first analyzed in terms of reflexes. A reflex is sometimes thought of as a response which involves only a few synaptic delays and which is independent of higher centers of the brain. Thus the doctor's hammer striking (and stretchinq) the tendon just below the kneecap results in a reflexive kick no matter what we may be thinklnq about at the time. However, more current notions hold that reflexes are to be thought of as part of a hierarchical control scheme, functioninq to carry out incompletely specified commands from above. Thus, we find that the familiar knee jerk reflex does not occur when the knee's owner is in free fall9 While this particular reflex is adequate and appropriate in some cases (especially locomotion), it is under the control of other reqions of the brain. And so it goes: if we look at the nervous system hierarchically, we can discover layer upon layer of relatively autonomous subsystems controlling and controlled by others.

The visual orientinq reflex has received a substantial amount of attention in a wide ranqe of animals. Attempts to model it have led me throuqh a number of concepts which I thouqht would be interestinq to apply to the robot problem. The main notion is that this system is a rapid, parallel device which typically needs to know remarkably little about an object to respond appropriately to it. Thus, it seems to me, a robot grasp reflex could be constructed which could appropriately grasp an object in the visual scene quickly and cheaply upon a *very* crude command (i.e. without beinq told the object's exact location, orientation or shape).

The next section of this paper provides a brief review of midbrain mechanisms which mediate the visual orienting response. The last section describes a rather brief and superficial effort at applying notions from the nervous system to the problem of usinq a serial computer to control a robot hand in a simple environment. Finally there are some suqqestions for generalizinq to a more realistic case.

Author's address as of September, 1973:

R. L. Didday
Information and Computer Science
Applied Sciences Buildinq
University of California
Santa Cruz, California 95060

Animal visual orientinq reflexes

The optic tectum (or superior colllculus as this midbrain structure is called in mammals) is common to vertebrates and is implicated in visual localizing responses. In the frog, it mediates visual approach responses (snapping and orienting), while in humans it produces reflex eye movements. Let us look at its structure in the frog.

The optic tectum stands between sensory and motor systems. There is an orderly, somatotopic mappinq of points on the retina onto the tectum and from tectum to motor response. This is illustrated by microelectrode recordlnqs and by electrical stimulation of the tectum—such experiments show that a particular region of the tectum can produce a motor response which is "aimed" at the same region of visual space which projects to that spot on the tectum. 8

An early paper by Pitts and McCulloch orovides a provocative framework for describinq the operation of the frog. tectum. Picture the retina and its processes as sendinq a spatially arrayed somatotopically mapped representation of the visual scene through the ganalion cell axons to the tectum. Regions of this array which are highly excited correspond to regions of the visual scene containing visual details which "fit" the feature detecting processes of the retina. We might then imagine that each locus of excitation on the tectum triggers a motor command aimed at the corresponding point in space so that the summated effect of all the muscle commands yields a response to the "average" location—the "feature center of gravity" (see Didday4 for a more detailed discussion). This scheme would predict proper behavior when faced by one "prey object", but would nredict responses to the average spatial location of a number of visual stimuli. 'Surprisinqly enouqh, this does sometimes occur in the frog6, but usually response is made to only one of those objects present- The basic Pitts-McCulloch scheme does not predict this, usually more appropriate, type of resnonse.

The notion that the image of an object can directly lead to the proper combination of muscle contractions required to deal with it is very appealing. It offers a rapid way to deal with the problem of orientina the body with respect to objects in the environment. One possible way to allow for responses to only one of several stimuli present while maintaining the feature of allowing the stimulus to shape the response directly is to alter the Pitts-NcCulloch scheme by dividinq the incoming information into overlapping regions and providina a mechanism to select one region. The "super-position" effect which uses an object's image to tailor the response then works within the chosen region. The more rare responses "between" two stimuli would then respresent failure of the selection mechanism.

Clearly, there are a great number of ways to implement such a function, and three different implementations are contrasted in Didday4. One implementation utilizes cells which function much like the "newness" and "sameness" cells described by Lettvin, et. al.'.

Arbib and Didday1 describe how this basic framework might be used as a locallzina reflex which acts in conjunction with recognition mechanisms to form a percelvlna system capable of using memory to interact with a complex environment.

The basic organization provides a means of using visual features to control motor behavior which must be accurately tailored to local details of an object In the visual scene. Behavior which depends on detailed analysis and knowledge of more subtle aspects of the objects is carried out by other reqions (cortical visual centers in mammals). The optic tectum or superior colliculus thus is seen as performing a reflexive localizinq role.

A grasp reflex for robots

This section describes a first attemot to define a system which is organized like the visual mechanisms depicted in the previous section. The motor behavior which the system is to control is arm and hand movement.

I have made a number of assumptions in what follows, many of them with an eye to simplifyinq the problem. Many of them could be relaxed with present knowledge and further effort. Some thoughts on such expansions of the power of the reflex are described later. The most serious simplifying assumption is that the world may be treated as two-dimensional.

One of the principal notions put forth by Arbib and Didday[1] was that the qoal of perception is in determining appropriate motor behavior. This led to the idea that the task of perception is not to construct a delicately detailed internal reconstruction of the input scene but rather to help choose possible motor responses. Thus we find sensory feature detectors elegantly matched to possible motor response (or perhaps vice versa). The feature detectors soon to be described owe their details to the hand they are to help control.

The hand is assumed to consist of two "fingers" on a palm as shown in Fig. 1. Details of the hand which are reflected in the desiqn of the feature detectors are the length and width of the fingers and the maximum opening distance of the fingers. As seems to be the case in animal nervous systems, motor programs (are assumed to) exist which, when given the appropriate parameters, can direct the hand so that the palm lies over point (x, y} at an angle O and with the fingers w units apart. It is then the task of the grasp reflex to analyze the visual scene and produce values of these four parameters which are appropriate for graspino some object in the scene. Four parameters suffice because the system to be described functions in two dimensions only.

Further simplifications arise from the assumption that the two-dimensional visual input is obtained from "above" the real world scene and in such a way that all objects are seen from the same distance. Thus an object's size is immediately recoverable from the size of its retinal image. Finally I have assumed that all objects are darker than the surround.

Figure 2 provides an outline of the conceptual organization of the system. The implementation is for a serial computer, thus where layers of parallel operating cell types would exist in animals, this implementation simply maintains lists of the location, orientation and level of excitation of active "cells". The terminology from nervous systems is used throughout, however.

There are five classes of "cells" which process the visual scene to produce a specific motor command. All are "directional". The first two are sensory feature detectors which receive inputs from local regions of the retina and detect "tongues" (class a) and "Inward boundedness" (class b). These somewhat obscure descriptions are ellucidated in Fig. 3a and b. There is already a larqe amount of detail of the nature of the hand included at this level. The maximum opening width of the hand determines the width of the excitatory region of the class a cells. The lenqth of the class b cells corresponds to the length of the hand, and the width of the inhibitory cell row in the class b cells corresponds to the widths of the fingers. All these would change 1f a different hand were to be used.

In a nervous system implementation, there would be some number of overlapping cells of both classes for each point on the retina, and the outnuts of these feature detectors would lie in some topographically mapped region. In the current, serial implementation, a list is created with an element for each class a cell which has a non-zero level of excitation, and similarly for class b.

Class d cells are excited by the concurrence of a class a cell at angle e and a class b cell at the same anole in a reaion "to the left of" the class a cell. Thus a class d cell siqnals the joint existence of a "tonque" and a reaion appropriate for the left finoer to occuny. Class e cells do the same but for the right finger.

Class c cells are excited by concurrent class d and class e excitation which is centered on the same class a cell. Thus class c excitation sinnals an appropriate angle and position for the nalm of the hand to occupy.

The selection mechanism operates on the totality of class c excitation, and selects that spatial region which has maximal class c excitation, just as the selection mechanism in the froq model selects that reqion which has maximal "foodness" potency. Since an object which is small enough to fit entirely within the opened hand could be qrasped equally well from any direction, there will in general be many class c cells excited. To aid the selection mechanism in such cases, a bias is applied to those class c firinq levels which result in arm positions which are less "awkward" beinq favored. That is, hand orientations that call for the least deviation from the hand/arm startinq position are preferred.

Once a region has been chosen, the class c, class d and class e excitation levels in that reqion are used to derive the motor command (the positional parameters for the hand). The hand position and orientation command is obtained by computing the sum of the x and y coordinates and angle for each class e cell in the region weighted by the Firing rate of each cell. It is this weighting process which allows the determination of position and angle to a greater precision than the spacing of the cells. The command for finger width is computed in a similar manner from class d and class e cells within the chosen region.

Figs. 4-6 show three visual scenes on which the simulated mechanism has been tested. The input scene is a 50 x 50 array of squares. The actual incut to the program is an array which for each square has a number corresponding to the proportion of that square which is dark. Fig. 4 illustrates a blob shaped object which has a lobe to the right which (qiven the feature detectors used) is good for grasping. This region produced maximal class c excitation, so the parameters are tailored to qrasp the object there. The "hand" is sketched in the chosen orientation.

Fig. 5a illustrates an object which is squared off at one end and pointed at the other and which is at an angle with respect to the coordinates of the input grid. Fig. 5b illustrates the regions corresponding to those class a cells which were excited. The response is made up of weighted sums of this information, and so is appropriate to the position of the object even though no cell is oriented colinearly with any side of it.

Fig. 6 Illustrates the ability of the mechanism to choose one of several objects present. Given the forms of the class a and class b cells, a squared off end is more appropriate for graspino than is a

blob, so the upper object is chosen. The approach from the right was chosen due to the bias toward smaller arm movement.

Conclusion

In its current, preliminary form, the grasp reflex carries out the operation "go get the best looking thing to grasp". What is "best" is determined by the two classes of feature detection units operating on the input scene. Another layer of feature detectors which process the scene before it gets to the grasp reflex could bias what is "best". This new layer (which would be under program control) could serve to enrich the power of the grasp reflex by allowing different sorts of things to be "best" at different times. If, for example, "handles" have some particular feature characteristics, appropriate settings of this new layer could cause the reflex to pick an object up by the handle (for using it) one time, and pick it up by some other appropriate part (for giving the object to someone else) the next.

Like any reflex, the mechanism has a fairly narrow range in which it can function appropriately. If, for example, the reflex chooses to grasp an object which turns out to be too heavy to move, some higher level function which makes a more sophisticated analysis of the situation will have to be alerted. In cases when the grasp reflex is adequate, it can serve to perform the task of orienting the hand very rapidly at very little computational effort.

The scheme must be further enlarged to incorporate "negative" sorts of information, e.g. excitation caused by objects not in the chosen region could be used to prevent those motor responses which would bump into other objects. Also, negative information could be used to allow picking up an object with the fingertips when it is shaped so that a full hand grip is impossible. More importantly, negative information is needed when the object is so small that it can be picked up from any orientation. In this case, even weighting in favor of minimal arm movements will not preclude the possibility that the average location nf excitation lies within the boundaries of the object-thus producing a command which would tend to place the hand on top of the chosen object. One other possibility that shows promise here is to use information from the object to alter the size and shape of the regions used by the selection mechanism.

The most important alteration would be to have the system function in a three dimensional world. Feature detectors responsive to visual features such as those Guzman[5] utilized might prove helpful here. Most of Guzman's features are local in nature—if complex global features prove to be necessary then the reflex organization becomes unattractive—the reflex is of value only if it is extremely fast.

It is conceivable that the organization described here is really only appropriate to the parallel nature of the brain. Much of the efforts in the analysis of Didday[2,4] were concerned with the problems of global versus local information flow. In the brain, global functions are very costly (require many interconnections)--in a serial computer they are somewhat more expensive than local ones due to the combinatorics Involved. It remains to be seen whether taking ideas of organization from the brain is of use to robots. Perhaps the organization into hierarchically controlled reflexes will provide less direct insights than those suggested here, but it seems to have been so universally employed in brains as to be of some basic value-

References

1 Arbib, M. A. and Pidday, R. L. (1971). The organization of action oriented memory for a perceivina system. Part I: The basic model.. J. Cybernetics, 1, 3.

2 Didday, R. L. (1970). The simulation and modelina of distributed information processing in the frog visual system. Stanford University Inform. Systems Lab. Tech. Rep., 6112.

3 Didday, R. L. (1971). Simulatina distributed computation in nervous systems. Int. J. Man-Machine Studies, 3, 99.

4 Didday, R. L. (1972), Structural models of simple sensory-motor coordination. Int. J. Man-Machine Studies, 4, 439.

5 Guzman, A. (1967). Some aspects of pattern recognition by computer. MAC TR 37, MIT.

6 Inole, D, (1970), Visumotor functions of the froo optic tectum. Brain Behav. Evol. 3, 57.

7 Lettvin, J. Y., Maturana, H. R., McCulloch, W. S. and Pitts, W. H. (1961). Two remarks on the visual system of the froq. In Sensory Communication. Ed. W. W. Rosenblith, pp. 757-776. Cambridge, Mass, : MIT Press.

2 Pitts, W. H. and McCulloch, W. S. (1947). How we know universals: The perception of auditory and visual forms. Bull. math. Biophys., 9, 127. This paper is reprinted in McCulloch, W. 5. (1965). Embodiments of Mind. Cambridoe, Mass. : MIT Press,

9 Roberts, T. D. M. (1967). Neurophysiology of postural mechanisms. Butterworths, n. 66.

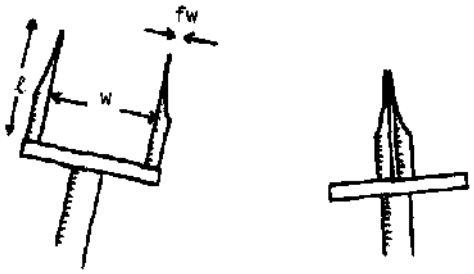


Figure 1

The Hand, open and closed, w= 5 units, ℓ=5, fw="1".

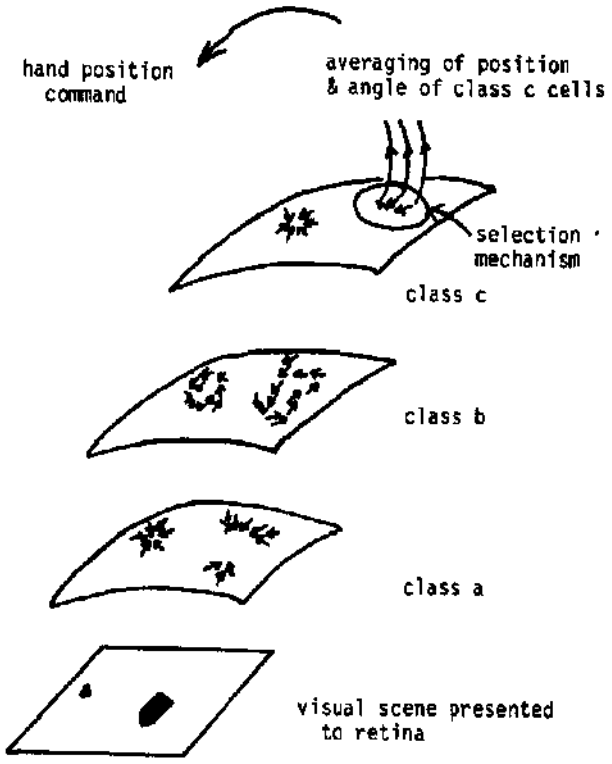


Fig. 2 A visualization of the scheme for determining the x, *y* , o position of the hand. Cell firing is indicated by a small arrow whose orientation corresponds to the orientation of the receptive field of the cell. The selection mechanism chooses the region containing maximal excitation.

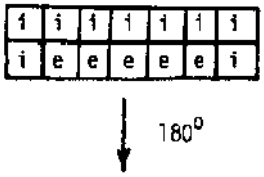


Fig. 3a

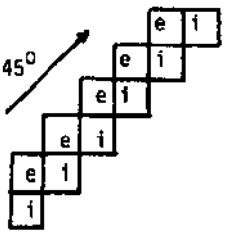
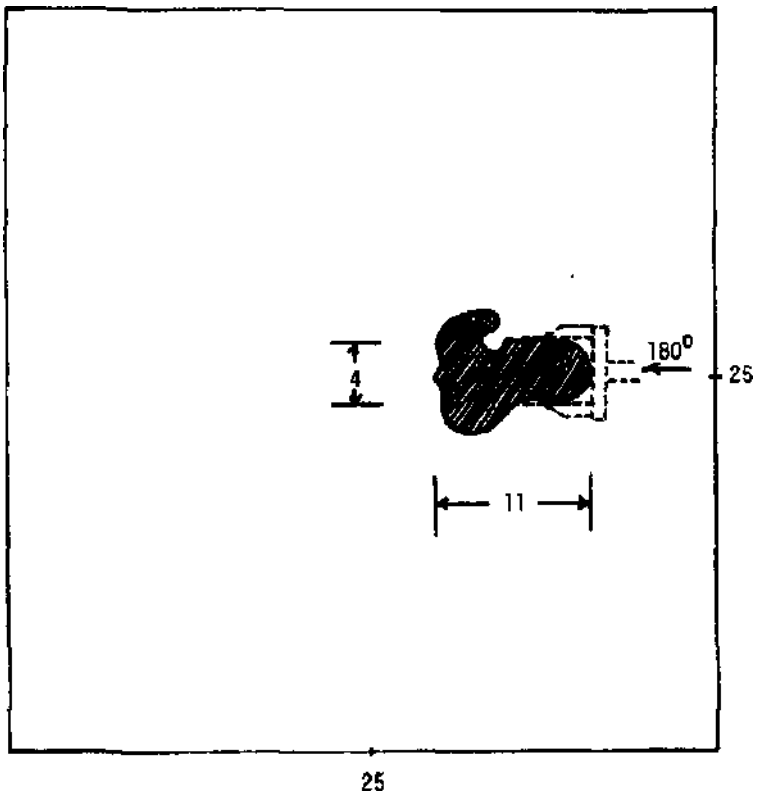


Fig. 3b

Fig. 3 The receptive fields of typical class a and class b cells. Visual inputs to squares narked "i" have an inhibitory effect, those falling on squares marked "e" have an excitatory effect. There is a class a and a class b "cell" for each point in the visual scene and in each orientation 0°, 45°, 90°, 135°, 180°, 225°, 270° and 315 . Combinations of cells at these orientations and having differing firing rates can deal with visual events lying at any angle.

Class a cells detect "tongues" -- i.e. places that are not too wide for the hand to grasp. Class b cells detect regions which a finger could fit around.



F1g. 4 A blob. The hand parameters chosen would place the hand in the position shown. In this computer run, eleven different excited class c cells were in the chosen region.

Fig. 5a An object oriented at about 256°. Weighted combinations of class c cells at angles 100 , 225°, 270° and 315° combined to specify the appropriate hand angle.

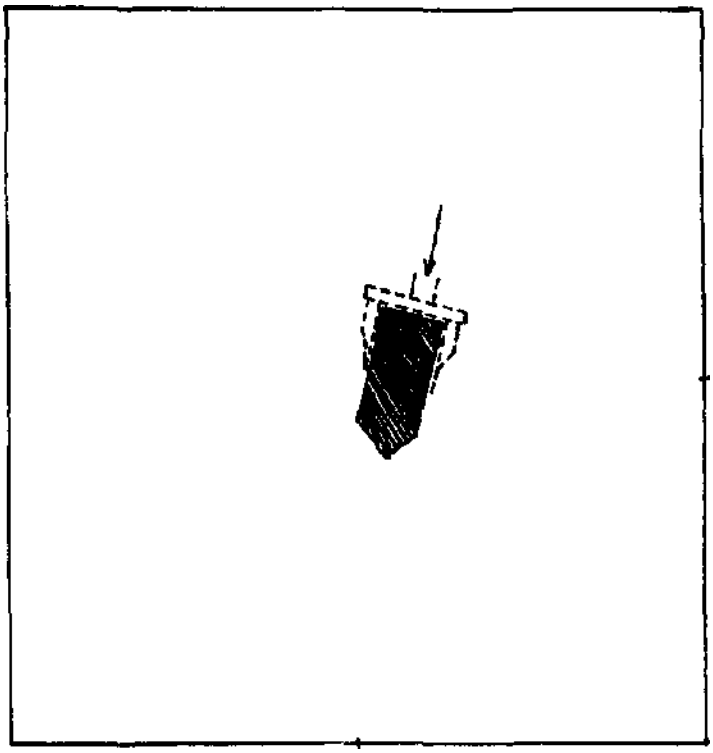
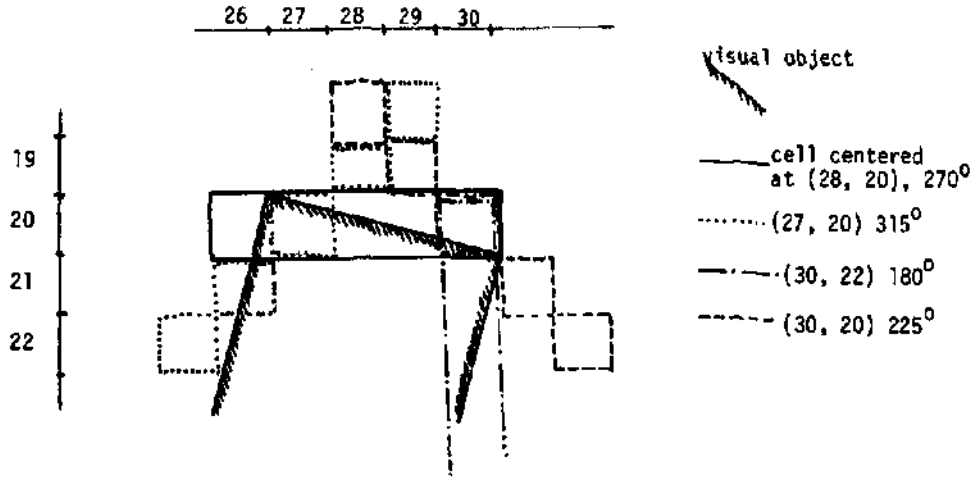


Fig. 5b Details of the process showing the class a regions which participated in making up the class c firing which in turn tailored the motor command. Amount of overlap with the object determines firing rate.



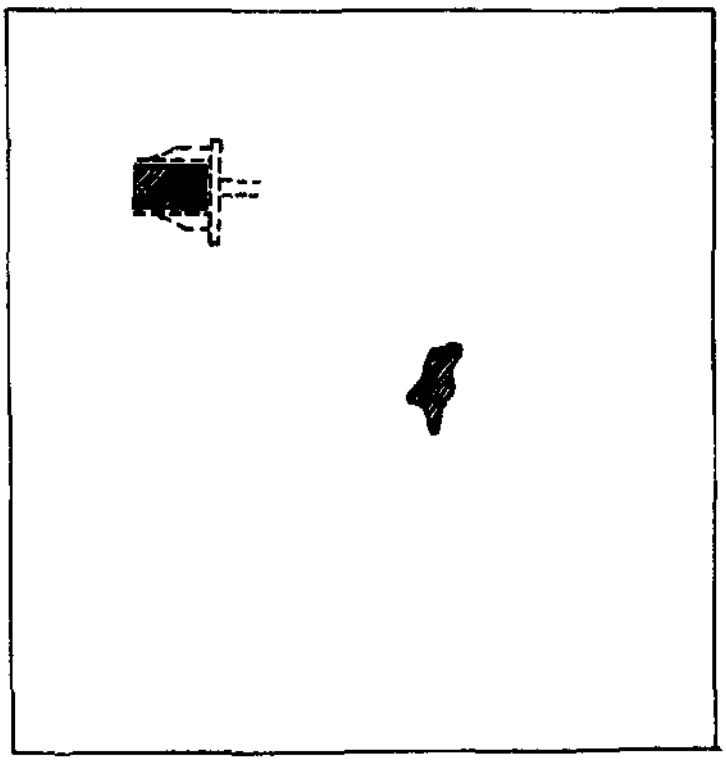


Fig. 6 A scene consisting of two objects. The hand is capable of grasping either -- the region around the right end of the upper object was maximally excited so the hand parameters were tailored to it.