Depth as a Disambiguating Cue in Visual Localization of Objects

Andrew S. Cromarty
Center for Systems Neuroscience
and
Dept. Computer & Information Science
University of Massachusetts at Amherst

COINS Technical Report 81-37
November 1981

Abstract

An animal or robot attempting to orient toward or away from an object in its visual field must have some form of three-dimensional map, whether implicitly or explicitly stored, in order to determine the proper direction of orientation, due to ambiguities that are present in two-dimensional retinotopic maps of the environment. Since even stationary, monocular animals can solve this problem, it is clear that stereopsis alone is an insufficient explanation of visual object localization.

A geometric proof is presented to demonstrate that the monocular retinotopic map is sufficient for such localization when augmented by a scalar-valued depth estimate such as might be provided by the lens accomodation system. This proof is obtained through calculation of a mapping from a retinotopic (or vidicon) surface to a simple output vector that could correspond to the neuronal motor outflow activity. Biological structures that may be present in the brains of lower vertebrates to approximate this complex mapping function are discussed.

C H A P T E R V

A STUDY OF THE ORIENTING RESPONSE

The problem of visual localization of objects in a three-dimensional environment is well-known and unsolved for roboticists and biologists alike. Much of the successful research of the former group involves the use of depth cues such as might be provided by, for example, optic flow and similar motion-derived information [19, 25]. Corresponding progress for the biologists, however, and especially for those studying lower vertebrates such as the frog, has been hampered by lack of sufficient data on the organization of the motor outflow pathways and the precise manner in which they receive and process visual information. In this chapter we shall demonstrate that it is possible to create an explanatory model of this system that is consistent with what we currently know about the interaction of visual input, depth information, and motor outflow in lower vertebrates such as amphibia.

It must be emphasized at the outset that this is not a model of stereopsis. Abstract models of stereopsis already

1 work on applying these models to the amphibian visual system is already in progress [20].

It is apparent from the neuroethological record, however, that at least in amphibia, stereopsis is neither sufficient nor necessary to explain the first-order phenomena of object localization.

To convince ourselves that this is true, we will consider in turn the stimulation studies of Ewert on the mapping from the tectal surface to directed motor activity and the work of Ingle on prey selection in monocular frogs.

The Tectal-Motor Map

The visual input arrives at the retina as a two-dimensional projection of the three-dimensional world, in every sense a "shadow" of the true real-world image. This 2D image is passed by the retina in a faithful one-to-one retinotopic fashion [18] to the optic tectum, which is also organized as a 2D surface. It is now firmly established in the literature that to each retinal locus there corresponds a location in the tectum which is activated when that retinal locus is excited by an appropriate visual stimulus in the environment.

Fig. 1. Ambiguity in the two-dimensional retinal projection

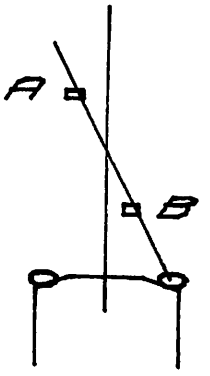


Illustration of the ambiguity resulting from the two-dimensional retinotectal representation of three-dimensional visual information. An object at point "a" is to the left of midline, requiring that the animal turn left to orient toward it; an object at point "b" requires orientation to the right. Both points map to the same tectal location, stimulation of which will always cause the animal to orient toward <u>one</u> unique location in the three-dimensional world. Note that monocular animals can disambiguate this representation.

In the tectum, then, there exists a faithful two-dimensional representation of the objects in three-space. What is the relationship between this 2D tectal representation and the motor system? Through a series of stimulation and ablation studies, Ewert has shown [12] that point stimulation in the tectum consistently produces an orienting response toward the corresponding retinotopically-mapped location in the visual field. This would appear at first blush to have answered the question of how the animal locates and orients to objects in the real world, but in fact precisely the opposite is true, as is shown in Figure 1. Note that the images of prey objects "a" and "b" will map to the same location on the retina (and hence the tectum), but the direction of orientation is opposite for the two objects. That is, for object "a" the animal must orient slightly left, whereas for object "b" the animal must orient slightly to the right. If we electrically stimulate the corresponding tectal location, the animal will always orient to one specific real-world location; that is, the tectal locus corresponds to a unique ray projecting from the animal's turning axis into the three-dimensional world. Whether this ray is directed to the animal's left or to the right, we can infer that the tectal activity is not sufficient to localize the prey stimulus, since "a" and "b"

are not on the same ray. This paradox obtains over a wide range of the visual field.

These data provoke the obvious suggestion: that the animal uses binocular information to further constrain and thus disambiguate this localization problem. But the results of several further experiments make it quite clear that binocular information is unnecessary for the animal to localize prey relatively well, even for objects in the binocular region of the visual field.

First, Ingle [21] has shown that frogs which are blind in one eye are capable of locating and snapping at prey in the binocular portion of the visual field, with only slightly decreased targeting success for prey as far into the binocular region as fifteen degrees past the midline. In particular, they snap slightly short of objects in the nasal visual field; the standard deviation of the snapping distance is the same. Severe degradation of orientation, in the sense of (for example) orientation to the wrong side of midline, is not observed. This observation suggests that binocular vision (i.e. stereopsis) may not be the dominant mechanism in 3D localization of such visual stimuli.

Second, in further prey-selection studies, Ewert [12] found that if monocular toads were visually stimulated, it was not the locus of snapping but rather the latency until snapping that differed as compared against the normal toads, and furthermore, "when the snapping area in the 'blind' tectum of the monocularly blind toad was stimulated electrically below threshold when a visual prey object was present, the toad behaved normally and snapped frequently" [15]. This suggests that the principal role of binocular input to the tectum is not three-dimensional prey localization, but rather reinforcement of the signal at the corresponding location on the contralateral tectal surface. (This could be useful, for example, as part of a prey-selection system; that is, the animal will be most likely to snap to the two-dimensional locus where the activity between the two tecta coincides.)

A Constraint-based Analysis of Visuomotor Activity

While it is clear that stereopsis is cannot be the mechanism by which such animals locate objects in their 3D environment, it is also clear that no truly two-dimensional surface can provide the information necessary to disambiguate this inherently three-dimensional problem. We are left, then, with two possibilities:

1. The retina provides three-dimensional information. This approach is along the line of Marr [26], who suggests that the retinal image provides enough information to allow recreation of a depth-tagged "2 1/2-D sketch" of the environment, through a mechanism that performs spatial frequency filtering.
2. There is another source of depth information not present in the retinotopic representation of the image.

The first possibility is at least plausible in higher organisms where (a) there is enough processing capacity (brain) to justify the suggestion that spatial frequency filtering could occur, and (b) there is already some evidence for spatial filtering of the retinal image [4, 9]. In lower vertebrates, however, neither of these conditions can be confidently asserted. In addition, there <u>is</u> another source of depth information, namely that provided by the lens accomodation system. It has been proposed [21] that this single scalar-valued quantity is used in animals like frogs to provide information on the third dimension of the visual world; we will now turn to formal study of that

hypothesis.

Because the precise mechanism of accomodation is not well understood, we shall simply assume that the animal has such a system, and that (a) the animal can focus on objects in the visual field, and (b) some quantity corresponding to the extent of flexion/contraction of the lens muscle is available to central brain regions. Note that we do not claim that accomodation information is not available in the tectum; rather, we rely on the neuroethological work of Ewert and Ingle to demonstrate that, whatever the role of such information in the tectum, it does not seem to be "solving the 3D orientation problem" there, at least according to the currently available data. Indeed, the stimulation studies alone strongly suggest that the tectum is not the brain region wherein the 2D map and depth data are integrated. We therefore suggest that the tectal representation of the world is only two-dimensional, as Ewert's work implies, and that this 2D map is augmented by lens accomodation information farther downstream, perhaps in the reticular formation.

To demonstrate the plausibility of this proposal, we shall prove that a two-dimensional (that is, (x,y)-coordinate) specification of the location of the object's "shadow" on the retinotopic surface requires only a

single additional scalar-valued quantity to allow the animal to effectively locate objects, and that the lens accomodation signal we have postulated serves perfectly adequately as the scalar quantity needed.

Consider a monocular amphibian contemplating a visual stimulus (Figure 2). The animal's eye does not face directly forwards (0^{o}), but rather in a direction somewhat offset from the forward direction (0^{o}). The visual stimulus itself appears at some angle 0^{o}_{vis} with respect to straight ahead. And to make matters more complex, none of these is the angle through which the animal must turn, since the turns are performed around an axis passing not through the eye, but rather through some point farther back in the body, perhaps in the area of the dorsal vertebrae.[1] We shall refer to this point as the "pivot point", and the angle of turning (with respect to straight ahead) as 0_{turn}. We can then generate a proof by geometric construction, as follows:

Consider the triangle specified by the three points P (pivot point), B (bug), and E (eye). Our goal is to calculate the angle 0_{turn}, the angle of turn for the animal to orient to a visual stimulus at point B (taking the

---

1. It is interesting to note that the word "vertebra" comes to us from the Latin <u>vertere</u>, "to turn".

Fig. 2. Visual geometry during orientation

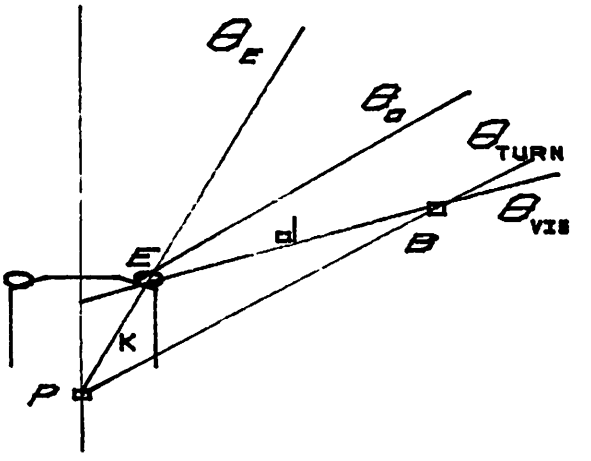


Diagram of the geometry of the frog's visual processes during orientation. Zero degrees is straight ahead of the animal; θ_o is the angle between zero degrees and the direction in which the animal's eye faces; θ_{vis} is the visual angle for a visual stimulus "B"; θ_{turn} is the angle through which the animal must turn to be facing "B"; "P" is the pivot point within the animal's body, i.e., the axis of turning; "E" is the animal's eye. Distance "d" is the length of side EB and "k" is the length of side EP of triangle EPB.

current orientation to be O^{o}). We will speak of angles and three-dimensional (x, y, z) specifications of the location of points in space interchangably, since this merely requires a coordinate transform, that is, the use of spherical or rectangular coordinates respectively. It should be clear that the "shadow" of all points along a ray EB will always map to a single point in the retina (and hence the tectum).

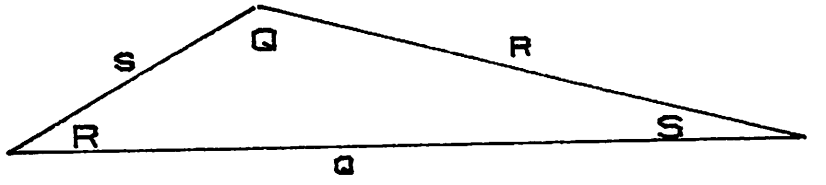
By assuming that the animal's skeleton is more or less rigid, that is, that the distance between (for example) P and E is invariant for a given animal, we can see that we always know the length of line segment EP and the angle O_e (see Figure 2).[1]

From the Law of Sines we know that, given a triangle QRS (see Figure 3), we have that

---

1. We note in passing that we will always take the animal to be facing straight ahead at the time of visual stimulation, both for the purpose of simplifying the explanation and because treatment of the more general case would require a better understanding of the precise location of the pivot point P. It should be immediately obvious to the reader that this simplification can be made without loss of generality.

Fig. 3. Law of sines



$$\frac{q}{\sin(Q)} = \frac{r}{\sin(R)} = \frac{s}{\sin(S)} \qquad (1)$$

In particular, this tells us that

$$\frac{\sin(\text{angle EPB})}{d} = \frac{\sin(\text{angle EBP})}{k} \qquad (2)$$

where d = length of side EB and k = length of side EP. Note that if we can solve for angle EPB, then we have the angle θ_{turn}, since it is obvious from inspection that

$$\theta_{turn} = \theta_t + \theta_e \qquad (3)$$

where θ_t = angle EPB. In fact, we can solve for θ_{turn} in just this fashion:

$$\frac{\sin(\theta_t)}{d} = \frac{\sin(\text{angle EBP})}{k}$$

where θ_t = angle EPB. From inspection it is obvious that angle EBP = $\theta_{vis} - \theta_t$, so we have that

$$\frac{\sin(\theta_t)}{d} = \frac{\sin(\theta_{vis} - \theta_t)}{k}$$

which we can rewrite as

$$\frac{k}{d} = \frac{\sin(\theta_t - \theta_{vis})}{\sin(\theta_t)}$$

In order to isolate the θ_t term we use the following trigonometric identity:

$$\sin(a-b) = \sin(a)\cos(b) - \cos(a)\sin(b)$$

where $a = \theta_t$ and $b = \theta_{vis}$. This gives us

$$\sin(\theta_t - \theta_{vis}) = \sin(\theta_t)\cos(\theta_{vis}) - \cos(\theta_t)\sin(\theta_{vis})$$

so that by substitution we find that

$$\frac{k}{d} = \frac{\sin(\theta_t)\cos(\theta_{vis}) - \cos(\theta_t)\sin(\theta_{vis})}{\sin(\theta_t)}$$

This now allows us to divide numerator and denominator by $\sin(\theta_t)$, so that we have

$$\frac{k}{d} = \sin(\theta_t)\cot(\theta_{vis}) - \cos(\theta_t)$$

and now we can isolate θ_t and solve for θ_{turn}:

$$\sin(\theta_{vis})\cot(\theta_t) = \frac{k}{d} - \cos(\theta_{vis})$$

$$\cot(\theta_t) = \frac{k}{d\,\sin(\theta_{vis})} - \frac{\cos(\theta_{vis})}{\sin(\theta_{vis})}$$

$$= \frac{k}{d\,\sin(\theta_{vis})} - \cot(\theta_{vis})$$

$$\theta_t = \operatorname{arccot}\left[\frac{k}{d\,\sin(\theta_{vis})} + \cot(\theta_{vis})\right]$$

$$\theta_{turn} = \operatorname{arccot}\left[\frac{k}{d\,\sin(\theta_{vis})} + \cot(\theta_{vis})\right] + \theta_e \qquad (4)$$

But what are "k" and "d"? Variable "k" is side EP, which is the distance from eye to pivot point, a biological constant "known to" or "knowable by" the animal; and "d" is simply the distance from the eye to the object, which is precisely the information that the lens accomodation signal

supplies. Thus we have an expression for θ_{turn}, the turn angle for the proper turn (that is, the turn towards the correct three-dimensional locus), expressed purely in terms of anatomical constants, the two-dimensional locus of tectal activity, and the lens accomodation signal "d".

Whither the arccotangent neuron?

If we merely wish to construct a mechanical system to perform this function, an equation such as (4) may be an adequate result; however, for the biologist, the question of how this mechanism can be neurally implemented still remains. Must one conjecture that neurons calculate such complex trigonometric functions?

Certainly all we learn of the nervous system makes it unwise to predict that it cannot perform such computational tasks. However, it is difficult to imagine how such an equation would be implemented, and to maintain an intellectual grasp on our modelling we shall instead explore a second alternative.

Physicists are familiar with the technique of assuming linearity for certain trigonometric functions within small bounds; this is a standard approach to solving the equations for pendulum motion, for example. What if such a simplification were employed by the motor outflow system to approximate this rather more complex equation (4)?

One result we might expect to observe is a smooth, monotonic degradation of match between the predicted and true target locations during snapping as we consider points successively farther from some "central" location in the visual field. This is in fact exactly what Ingle observed [21]: the more nasal the stimulus location, the more likely the animal is to snap short of the stimulus. This experiment is not conclusive evidence for the use of approximation, of course, especially because the most temporal regions are not easily tested.[1] At least, however, such an experimental result might seem less peculiar when viewed as the result of a failure of a linearity

---

1. Amphibia typically snap only to prey objects that are more or less directly in front of them; more temporal objects elicit an orienting response rather than a snapping response. There is disagreement as to whether or not this observation applies as strongly to frogs as to toads, however [21]. (Note that toads are more active predators, stalking their prey rather than awaiting it, and that their eyes are directed somewhat more forward.)

approximation used by the motor system.

Unfortunately, there is currently insufficient experimental evidence to warrant a more specific proposal on the nature of this approximation function. It does seem plausible, however, that a parallel-to-serial, maximum-finding anatomy such as that proposed by Didday [7] for prey-selection could implement such an approximation, albeit farther downstream than Didday had postulated.

Discussion

The phenomenon of normal orienting and snapping behavior in monocularly blind amphibia is not as perplexing as it might seem. All information needed to disambiguate the orientation problem is there, if we assume that visuomotor brain regions have access to lens accomodation information.

Of course, such a proof serves only to demonstrate that sufficient information exists to solve this problem in monocular animals given the lens accomodation information and the two-dimensional locus of activity; in some sense, we already knew that this had to be true, as a result of the

research [12, 15, 21] previously described. However, the question of whether or not this is in fact the correct explanation remains to be seen. While there is reason to believe that stereopsis cannot be the primary means of depth perception[1] (at least for prey localization), we have no data to justify rejection of a "spatial frequency filter" model; it is rather the case that we simply cannot find any evidence <u>for</u> such a model in the present literature on lower vertebrates.

What role, then, might binocular input play? Several suggestions have been made in the course of this paper. In more detail, we might consider at least the following:

1. <u>Additional depth information for prey localization</u>: The fact that binocularity is not necessary for prey localization does not mean that stereopsis does not contribute depth data, but merely that it is not the

---

1. Additional evidence has recently been provided by Udin and Collett [32] to suggest that the binocular intertectal projection in amphibia contains approximately two orders of magnitude fewer fibers than current stereopsis models [8, 20] require to achieve the visual resolution observed during amphibian orienting behavior. This is not conclusive in and of itself, since there may be drastically different models of stereopsis that <u>could</u> operate with so few intertectal connections; however, when taken in conjunction with the results [12, 15, 21] already described, their observation supports the conclusion that stereopsis is not the dominant mechanism in amphibian prey localization.

sole or dominant source of depth information. It is still possible that binocular input supplements the accomodation-based prey localization process.

2. Depth perception for other behavioral tasks: Recognition and avoidance of barriers and predatory objects is almost certainly not performed by the tectum [22], and it is not difficult to imagine that barrier localization relies on a depth discrimination method different from that used in prey selection; stereopsis could be that mechanism.

3. Prey recognition reinforcement: Binocular information could be used for reinforcement of the two-dimensional tectal activity that corresponds to "preyness". For example, binocular regions might map to each other such that an optimally prey-like object is one which (a) is in snapping range (as determined by accomodation) and (b) produces tectal images which "overlap" when the activity levels of the two tecta are compared. If this were the case, we might expect to find interference patterns in the tecta when images are not optimally located. Interesting interference effects have in fact been observed by Didday [7] and by Collett [5]; unfortunately,

physiological study of such interference phenomena has not proceeded apace with these behavioral observations. Nonetheless, the possibility that binocular input is used for reinforcement of the 2D map rather than for depth inference is an exciting prospect deserving of additional study.

There is an alternative explanation of monocular orientation which has not been discussed in this chapter, but which would justify the use of stereopsis models of depth perception in lower vertebrates: it is possible that the research results of Ewert, Gaze, and/or Ingle are misleading or in error. This possibility deserves mention only because it is possible to imagine that binocular or accomodation information somehow modifies the patterns of tectal activity induced by a visual stimulus, such that the retinotectal map is effectively <u>not</u> retinotopic. Given the weight of historical precedent, however, and especially because so little is known about the biology of the accomodation and stereopsis systems in amphibia, any such explanation must be viewed as highly speculative in nature.

Acknowledgements

I am indebted to Paul Grobstein for first bringing to my attention the inability of stereopsis-based models to explain the orienting response in amphibia, at the Visuomotor Conference at the University of Massachusetts at Amherst in November 1981. I would also like to thank my many colleagues in the Department of Computer and Information Science for their helpful comments on a draft of this paper, and especially Rich Sutton and Martha Steenstrup for brainstorming with me on the three or four rather less tractable versions that the geometric proof went through before it reached its final form.