ROBERT PIŁAT
The Problem of 3D Vision in Phenomenology and Cognitive Science

In this paper I shall focus on the cognitive and ontological status of 3D visual perception. According to David Marr's account, depth perception is achieved mainly by the cooperation of the eyes (stereoscopy), although single eye adaptation (focusing) also plays an important part in the mechanism. It is not quite clear, however, how the eyes cooperate. Research on this problem is, in part, technical and involves designing algorithms for interpreting random dot stereograms. What makes the issue philosophically relevant is the role of the higher mental functions in forming the 3D visual representation of environment. What I mean here by ‘higher mental function’ is the involvement of stored knowledge in depth perception. Obviously, the term ‘higher order’ refers, in this sense, both to human and to animal cognition. The functioning of the visual apparatus in animals has as much philosophical relevance here as the corresponding processes in human beings. What is at stake is the relationship between knowledge and perception.

In pursuing the problem I shall attempt to combine phenomenological and cognitivist approaches. D. Marr's study (Vision, San Francisco 1982) will be my cognitivist source; M. Merleau-Ponty’s analysis of depth in his Phenomenologie de la perception and in Image and Word will serve as examples of phenomenological approach.

There is no shortage of objections against bridging these two disciplines. It was the founder of phenomenology who first raised objections against mixing what he called the “psychological” and “phenomenological” viewpoints. On Husserl's view, phenomenology is about the structure of possible experience, whereas psychology (today we should rather speak of cognitive science) is trying to set out theories based on actual experimental
data. Psychology works on the basis of presupposed knowledge of the world, whereas phenomenology puts that knowledge in brackets and searches for its ultimate foundation.

The cognitivists also often declare themselves against mixing phenomenological insights into the body of cognitive theory. The objection is based on the observation that the structures, regularities and even necessities which are said to be revealed by the phenomenological analysis of experience do not apply to information processing. Obviously enough, the structure of possible experience cannot be part of neuronal functions or the real-time computational models thereof. What phenomenology describes as the structure of possible experience is at most an epiphenomenon — the effect of a physical process that has no physical effects of its own.

All objections notwithstanding, there have been several interesting attempts to combine the two conflicting approaches. It was M. Merleau-Ponty who first took this path and by now many have followed. The works of D. Follesdal, H. Dreyfus, E. Marbach, J. Hintikka and J. Mohanty — to name just a few — have already built bridges.

I am not going to discuss this problem in general terms, however. Instead, I shall try to show how the bridging strategy works in a particular case: the problem of visual depth.

The question is if the only source of information for visual depth perception is the irritation pattern on both retinas. If the answer is ‘Yes’ then we only need to design algorithms transforming the 2D image (which is, on Marr's account, an oriented image obtained by applying some filtering functions to the irritation values in different retina cells — see the next paragraph for detail) into a 3D image. This is roughly the position taken by those who endorse some sort of computational theory of mind, like Marr himself. If the answer is ‘No’ then we have to assume additional cognitive resources for visual depth.

At first sight, the answer seems to be ‘Yes’. The argument seems to follow straight from Marr's theory of stereoscopic vision. Marr envisions a mathematical function, to be interpreted physiologically as a light frequency filter, which has the job of transforming the irritation pattern on the retina into an information pattern representing the quality of the light. The output pattern obtained in this way makes it possible for the visual system to discern the edges of different lightness fields — the borders between them (shape edges) correspond to zero-crossings of the hypothetical function envisioned by Marr. Thus the system may be said to accomplish a product which Marr calls a primal sketch: set of edged fields connected to the states of the retina and representing different lightness fields in the environment.
Now, from both such sketches, coming from the two retinas, the visual apparatus seems to build up what Marr calls the 2½D image, which contains information about depth. The decisive step up from primal sketch to 2½D image can be described as the application of a set of algorithms which measure disparity — the angular difference between the relative position of a point in physical space in relation to the corresponding points of the retinas.

The construction of the 2½D image is to be built up from the primal sketch. On Marr's theory the whole system is strictly hierarchical, all the steps are the result of the algorithmic processing of the previous steps.

However, there are arguments against the purely computational interpretation of Marr's findings, which are not to be ignored. Let's name just two, calling them the general objection and the specific objection respectively:

1. **General objection:** Marr ascribes some of the postulated algorithms to specific cell aggregates, which supposedly carry them out. However, it seems that at least some of the parameters of the envisioned algorithms come from general knowledge rather than from neural design. It should be noted here that even if some of Marr's ascriptions are correct, none of them really refer to the level of more advanced visual information processing. Deriving 3D shapes from 2D shapes, which is accomplished by stereoscopic vision, certainly lacks any firm neural basis on Marr's account. If the algorithms take their parameters from anything other than from simple 'cell machines', it is probably from some kind of knowledge present in the brain in distributed form. Such knowledge presupposes that the perceiving subject has already had some sensory experience. However, the very concept of having had sensory experience is not intelligible without the assumption of the awareness of the 3D space in which the subject moves and perceives. (Otherwise it would be a very counterintuitive notion of sensory experience — something like a caricature of Hume's theory.) But if this observation is correct it implies that the perceiving subject must have access to the 3D space before the subject can see it. Moreover (and this is crucial), it is not just physical space, which would make the point rather trivial (sure we have 3D moving bodies before we use our eyes!), but **it must be basically (and phenomenally) the same 3D space, which the subject sees with his eyes.**

   One might try to go around this objection by saying that the 3D space is pre-given by other senses than vision. However, even if it were true (and it probably is), we would still need to know how the 3D space provided by other senses becomes part of processes that generate the visual 3D space. We seem to get into a trap here. It looks that our visual apparatus
accomplishes depth by making use of information that already requires visual depth or at least an intuition of some depth, which is almost identical with the visual depth.

2. **Specific objection:** As I have already mentioned, the most important mechanism in getting visual depth is disparity measurement. The system must be able to estimate and compute the angular difference between the position of a certain spot as viewed by the left eye and that same spot as viewed by the right eye. This presupposes that the system ‘knows’ which spots on the retinas correspond with each other. Otherwise the system would measure numerous disparities that would be of no use for visually exploring the environment (and mix them with useful ones).

The lines of vision linking a spot in space with the two retinas cross each other in many points but only some of them are the right crossings. The others are ‘blind solutions’. Only for the right crossings does the visual system actually measure and compute disparity. But how does the system know which ones are the right crossings? Or in other words: How does the system establish corresponding points on both retinas? Now, establishing correspondence can not be accomplished by some straightforward mapping of the retinas onto each other. The only available information here it that about disparity and nothing is ‘known’ to the system about a possible state without disparity — such a state does not exist at all. It looks that the system should measure disparity to ‘see’ the correspondence and should establish correspondence to be able to measure disparity.

I would like to offer the following solution of the problem, however tentative it can be: The puzzling virtual correspondence between the retinas might be seen as a property of abstract space although it applies to functioning of actual nerve cells as a parameter of processing carried out by the neural network. Getting and storing the right parameter is crucial for the processing. Otherwise we would see a false, illusionary depth. The right parameter can be understood as the virtual center of a variation performed on primal sketches. However, the variation is not a purely computational operation. It comes about through saccadic movements of the eyeballs and probably through the movement of the whole body.

Still another reason to think that bodily and eyeball movement should be crucial for stereoscopic vision (for getting the 2½ image) is that disparity measurement is most efficient if it is based upon the high-resolution channels of the retina. These channels are not available at every point of the retina. The visual system must put them to work by pointing them onto (or at least close to) the center of the visual field. As Marr puts it “A notable feature of a system (...) would be its reliance on the eye movements for building up a comprehensive and accurate disparity map from the two
viewpoints. The reason for this is that the most precise disparity values are obtained from the high-resolution channels, and the eye movements are therefore essential so that each part of a scene can ultimately be brought into the small disparity range within which the high-resolution channels operate” (Marr 1982, 128).

If the general framework of Marr’s theory is correct, then we may say that visual depth is accomplished by the further processing of a ‘flat image’ but have to admit that an independent virtual 3D space is already there to constraint the processing.

This solution can be further supported by a phenomenological argument. Merleau-Ponty shows in his *Phenomenologie de la Percepcion* (1945) and his essay on Cezzane's painting (*Word and Image*) that depth perception is based on bodily movement in actual, physical space. There is an ongoing interaction between that movement and the constitution of the visual field. They presuppose each other. None of them is more fundamental. The bodily movement changes the visual field and the changes affect the movement. It is this interaction that makes me believe that the space I can see is the same space that is occupied by my body. Merleau-Ponty persuasively shows that the relation between the body and the visual space is important for painting. What a (good) painter does is not just mapping a spatial layout onto a flat piece of canvas. He makes us see a space as oriented to a bodily observer. The constitution of depth emerges from the intimate relationship between the body and space, hence it is a primal visual quality and not something constructed.

Concluding remark: In order to establish a correspondence between the retinas it is necessary for the system to have access to independent information about actual 3D space. However, no such piece of information is likely to come from the visual system alone. There is no other vision preceding regular vision, no looking through a keyhole at the space before actually seeing it. Functional 3D spatial information must be provided by some internal model of the surrounding space. How do we get such a model? Marr suggests the role of eye movement. Following Merleau-Ponty, I also suggest that the role of bodily movement. These two mechanisms probably combine to give pre-visual, but at the same time genuinely visual, space (this is because even at the primitive stage in question it can be distinguished from other sensory spaces, like the auditory one). Only by checking the visual data against such a mental model, a visual system can establish the retina points that should be treated as corresponding to each other. And, as remarked before, only on the basis of such a model can the visual system measure disparity and by doing so construct visual depth.
This conclusion makes clearer Merleau-Ponty’s remark to the effect that depth is not the third dimension but the first one. The visual mechanism for depth gets information not only from the lit surfaces of objects to which the primal irritation pattern of the retinas corresponds. It also uses information coming from the presupposed models of the perceived scenes.