
Pictures in the Mind's Eye: Images in our Perception of World and Art

Julian Hochberg
Department of Psychology
Columbia University

and

Mary A. Peterson
Department of Psychology
State University of New York at Stony Brook

To appear in:
M. Schuster and B. Woschek (Eds.), *Imaginal Communication*, Gottingen, West Germany: Verlag fur Psychologie
Pictures in the Mind's Eye:
Images in our Perception of World and Art

Note: English versions are available from either author

Introduction

Pictures are made and used for many different purposes, and are therefore to be judged by many criteria. Superficially, the easiest criterion to establish is that of **fidelity**: how closely the picture matches the scene it is to represent. Even this judgment is not quite so easy or objective to make as it first sounds. The more we learn about why that is so, the more we learn about the perception of the world, and as well, about what pictures are. The picture itself is, of course, traditionally a flat pigment-dappled surface, approximately evenly lit; most recently (and increasingly), pictures are luminous displays on a television or computer screen. Except in a trivial case, the scene or subject being represented is a 3D array of surfaces receiving and reflecting different amounts of light. What can picture and scene have in common?

An answer that goes back (in essence) at least to Alberti and Brunelleschi, and still heard today, is that a picture matches its subject when both provide the eye with the same array of light, and painting and subject can therefore be **surrogates** for each other, with no change in the situation that faces the viewer. If defining pictures as surrogates provided a viable approach, it should be possible in principle to produce pictures, as good as any others, simply by specifying with sufficient precision the light that would be provided to the eye by any desired scene. It was toward this goal that the camera obscura and camera lucida were developed as first steps. The photographic and video cameras almost completely the job, if the scene that is to be represented already exists. Recent rapid developments in computer graphics have extended the physical optics of 3D rotation, ray tracing, surface quality and fractal surface form, thereby coming a long way toward allowing us, at least in principle, to provide near-perfect surrogates for physical scenes and events that do not in fact exist. Indeed, if defining pictures as surrogates provided a viable approach, the matter of picture perception would be of very limited interest to psychologists: This purely physical approach to making pictures requires us to understand how the structure of the physical world of objects, scenes, and events is reflected in, and specified by, the optics of the visual ecology, but it requires very little further knowledge about the human perceiver. We will see, however, that even the best and most careful of picture makers, who wants to make an optical match between scene and surrogate in this sense, will in general find it impossible to do so. Moreover, even if such perfect optical correspondence between picture and subject could be achieved, we might still decide that the picture was a poor representation of its subject.

Optical surrogates are not the only approach to pictures. At the other extreme, people who write about art have sometimes treated pictures as a sort of arbitrary and learned pictorial language, an expression of the structure of our culture and not of the physical world (Goodman, 1968; Wartofsky, 1979). In fact, as we will see, the perception of pictures is an intimate expression of the processes by which we normally perceive the
world: It involves the structure of our sensory processes and the limits of our cognitive processes, as well as the structure of the world; other cognitive structure concerns the content of what we are prepared to perceive, that is, what we know as images. Images are "about" things (and people) that might be perceived in the world, and must interface with those things in the process of perceiving them. Some writers treat images as internalized objects having the structure and properties of object or as internal pictures of objects. As we will see, however, and as has been argued eloquently elsewhere (Goodman, 1968; Kolers & Smyth, 1970), images are very different from things, and have very different and largely unknown properties.

Images, which are important components in the perceptual process, are structures that inhere in the viewer and not in the light to the eye, regardless of how well the latter is defined and measured. Images are intimately involved in the perception of the world, but their participation is easier to demonstrate and study in the perception of pictures. This is part of what has made pictures interesting to the perceptual psychologist, to the computer scientist, and even to the neurophysiologist, and that continues to engage us at the cutting edge of these disciplines.

Let us consider first the matter of pictures as optical matches to the scenes they are intended to represent, after which we will attempt to dissect the role of images in picture perception.

Pictures considered as optical surrogates

Let us briefly survey the ways in which pictures must differ optically from the subject they represent and consider what we learn from those differences. First, the light that pictures and scenes offer to the eye will almost certainly be very different in wavelength, because the scene normally provides light across the entire spectrum, whereas the picture substitutes some smaller set of component wavelengths to obtain the same effect. A sophisticated computer or video display may offer the viewer as few as three dominant wavelengths: e.g., the red, green, and blue of the television phosphors (which although not usually narrow-band and pure, could be made so and could still be used to represent the entire gamut of perceivable colors). Because of this enormous optical difference, no corresponding patches of light presented to the eye from picture and scene are of the same wavelength composition. Nevertheless, picture and scene may be indistinguishable in appearance. This is so because (of course) we have only three neurophysiological channels (three kinds of retinal cones that act as three filters) through which all of our sensory information about wavelength must be gathered. It is not the wavelength in the light to the eye that is important, therefore, but the balance of its effects on the three sensory channels.

Simultaneous contrast. That is just the barest beginning. The light provided by a normal scene, including sky, highlights, shadows, etc., usually varies widely, with the most luminous patch being thousands of times brighter than the least luminous. Pictures, on the other hand, can vary only over a range of about 42:1, given the limits of pigments' reflectance and video screen phosphors. To help overcome this limitation, light and dark objects are juxtaposed, resulting in a greater apparent difference in brightness. This simultaneous contrast arises because the individual points of the display are not processed independently by the nervous system, but are subject to lateral inhibition (as Mach surmised in 1895, and as has
since been verified physiologically; for a discussion of recent research on such extended neural patterning, see Olzack and Thomas, 1986). Unlike the physical array of light, in which each point is independent of the other, the units of neurophysiological response are spatially extended, and interact over some distance. But the manner in which the regions of light and dark are arranged— the chiascuro of the painting or photograph—draws on an attribute of the perceptual system that is at least as important: that is, these interactions that determine our perceptions of brightness occur across a limited distance. More distant parts of the picture have little effect on the apparent lightness and darkness we perceive at each edge. As shown by the labels in Figure 1 (a schematization of a drawing by Seurat; Hochberg, 1979), an array of progressively brighter and darker objects can be represented by the same set of light and dark transitions, and the background can still appear to be continuous, given the appropriately gradual gradients of shade in the latter. As we will see, such limited interaction appears in all aspects of perception, and indeed is what makes chiascuro possible.

Limited spatial interactions. We learn similar lessons from the ways in which space and form are limited by the nature of pictures. As we have known since at least the 15th century, the picture and the scene provide the eye with the same pattern of light only when the picture is viewed from the same standpoint from which it was constructed. Much has been written about this issue since it was reemphasized by Piranesi in 1770, and research is accumulating as to the degree of distortion that is actually perceived (Cutting, 1987), and as to the changes that are perceived within the represented scene when the picture is regarded from different viewpoints (Goldstein, 1987). Here consider only the point, as old as Leonardo, that even at the correct viewpoint, the shape that is projected by any object that is eccentric in the picture will not be the same as the shape that the object provides when viewed frontally. With large pictures, therefore, as the viewer moves away from the viewpoint (from v to v'), rectangular buildings and circular shields, frontally represented but no longer directly ahead of the viewer, are projected to the eye as distorted trapezoids and ellipsoids. Leonardo's solution, very widely adopted in photography and in computer graphics as well as in painting and drawing, is to present each recognizable shape approximately as though it is being viewed from straight ahead. Thus, the picture of the round shield in Figure 2A is consistent only with a single standpoint (v) and inconsistent with any other (v'). Nevertheless, the inconsistencies are not disturbing when the picture is viewed from v.

This tolerance of, or blindness to, the inconsistencies of spatial layout from one region to another, has been exploited in some pictorial art (notably Piranesi and Escher) and in a fair amount of experimental inquiry in psychology (Cowan & Fringle, 1978; Hochberg, 1978; Peterson and Shi, 1988) and analytic thought in computer science (notably, Haffman, 1971, and Waltz, 1975). In Figure 3, which extracts the essence of figures by Escher, Penrose and Penrose, and Piranesi (adapted from Hochberg, 1979), the figure is a spatial analog of the perpetual brightness "staircase" formed metaphorically by the gradients in Figure 1, and has much the same significance: local interpretations of orientation in depth can be largely independent of the overall array of depth information in the picture.

Such perceptual inconsistencies do not merely occur with static
pictures, as is often asserted. The same thing happens with real objects, moving in the real world if the viewer attends to a depth ambiguous region, even when unambiguous features are close by (ca. 1.8 deg) (Hochberg & Peterson, 1987; v. Hornbostel, 1922; Peterson & Hochberg, 1983). For example, consider intersections a (unambiguous) and b (ambiguous) in Figure 4: Even with definite motion between viewer and object, the cube’s perspective reverses when the viewer attends to the more ambiguous intersection (Peterson & Hochberg, 1983).

The implications of limited spatial interaction: Images and picture perception.

The ambiguity that can be manifest even with real, moving objects, when they are viewed with restricted attention, emphasizes the inherent ambiguity of seeing 3D space by means of 2D arrays of light. This ambiguity is of course what makes it possible to use a 2D canvas or TV to represent 3D objects and scenes. It is what makes it possible for artists to learn how to make the appropriate 2D patterns, as well: Each of the intersections in Figure 4 can be given innumerable 3D or flat "readings" (see Figure 5). We here propose that artists, of all people, are aware of the flat alternatives for the following reason: One of the two mechanisms involved in discerning what 2D pattern, when drawn on pad or canvas, will provide the surrogate for a 3D object or scene, consists of the act of restricting one’s attention to some individual feature, and selecting its 2D reading. Let us see what that might mean before we consider a second major mechanism that must be critical in the process of picture perception.

Analytic features. Each feature in a scene, when attended separately, permits various readings, including that of a flat arrangement of lines or edges. We know, as Gestalt theorists demonstrated vigorously from 1912 to the 1960s, that the units of perceived form or shape are not merely the mathematical points in terms of which the forms can be plotted (or, as we might put it today, not merely the pixels by which they are displayed). Neither are they the line elements into which we can dissect most patterns: Intersections a and b in Figure 4 take on a very different and mutually exclusive "readings" or three dimensional meanings, depending on the context in which they are embedded (A - J, in Figure 5), that are not inherent in the dots or lines themselves.

When considered without their context, some of these features are easier to read as being flat than others. Although we know of no experimental work to this point, it seems plausible to propose that the ease of fitting a flat alternative can be measured reliably. We expect that, to a first approximation, the ease of fitting a flat interpretation would be inversely proportional to the frequency with which a particular feature occurs in 3D layouts in the world. (A priori analyses that should be related to that measure have been done within computer science, notably by Waltz, 1975, but those need empirical surveys of what the visual ecology really provides; cf. Brunswik, 1956.)

We consider next how this notion of analytic features also implies the involvement of images as their context.

Features and context images. What we attend to usually what we look at with the center of gaze, the fovea (i.e., a region of the retina in the eye, providing an area of clear and detailed vision within the field of view). In terms of the visual angle it provides, the fovea is about equal in size to the circle in Figure 5A at normal reading distance. Outside of the
fovea, in peripheral vision, we can see little detail, such peripheral vision being about equivalent to the level of detail in a painting by Pissaro or late Monet. (Indeed, painting in such course grain is one way to provide a scene that simulates the impression made by the first glance, in which virtually all of what is seen is seen only in peripheral vision because the fovea has not yet been directed to more than one small region.) Within each glance, therefore, the viewer can attend in detail only to one or two features. Aside from what can be provided by peripheral vision, much of the context for the individual features must therefore be carried over in a kind of visual memory that summarizes and stores up the information from successive glances.

There is a great deal to learn about how much information is provided by peripheral vision, and about the nature of the context that is carried by integrative visual memory. It seems plausible, however, that those constraints are so weak that the attentive viewer can elect to consider only individual features in the scene, or small clusters of such features, and thereby be free to see the 2D reading even in a 3D scene. Indeed, if we examine the work of those artists who explicitly want to flatten the picture, we see that they avoid local features that are strongly 3D, replacing them with local features that are more likely to be seen as flat. For example, in the picture of Cezanne's father sketched in Figure 6 (adapted hers from Hochberg, 1984), intersections that would be strongly 3D in their local effects are replaced by features in which edges that lie at different distances in the 3D scene are coincident in the 2D projection. This technique makes deliberate use of the Gestalt "law of good continuation" (well before Gestalt theorists called attention to it and named it) to flatten the picture.

Abundant examples can be found in which artists use 2D local features in the effort to flatten the canvas, often producing paintings that are clearly pictures of recognizable 3D objects in 3D scenes, but that do not look one bit three dimensional (e.g., Figure 7 which is a sketch of a painting by Vuillard). In our opinion, this is not a trivial attribute of the work of a particular set of artists, but is an important clue as to the way in which images are involved in perception, and as to what images are like. Let us try to explain why 3D scenes may be clearly represented by pictures that were deliberately made so as to eliminate the depth cues and other features that carry depth information.

Recognizing three dimensions from two. When the gaze is directed among the different places in a picture or a scene in the world, many features never receive central attention. Moreover, even when features are attended, it may be for no more than the duration of a single glance (about 1/3 of a second). Although they are simultaneously present in the picture or scene, and of course can be returned to at will, individual features are essentially viewed within very little momentary context, and much of the picture is never looked at. To this, add the strong possibility that most information about depth is accessible only to central vision (and this is especially true if the viewer is stationary with respect to the scene). We can see, therefore, how the depth information that is actually picked up might be quite sparse, or even deliberately ignored in both the momentary glance, and in the succession of glances as well.

Given that scenes must be sampled glance by glance, and that the glances must themselves proceed under some guidance (that is, by some set of
visual questions that the viewer elects to ask about the scene, guided by
the initial view; we must ask about the early cognitive representation of
the contents of the scene, and the structure by which information about
those contents, as brought by successive glances, is stored. Cognitive
representations about our sensory explorations of the world are usually
called schema. Our representations of visual content -- of the specific
objects and their layout, and of our expectations about what detail we will
find if we look at some particular place -- are what we mean by visual
images. In this analysis of perception, images play a central role,
determining what information we will pick up in the first place, what we
will store and (at least to some extent) when we have learned all we are
going to learn, and when we can stop looking. Among psychologists concerned
with perception in general and picture perception in particular, the process
is generally talked about as one of "hypothesis testing," but that does not
help at all unless we can say something reasonably specific about what the
hypotheses are and what tests they entail.

An extremely promising step toward cataloging and understanding how
hypotheses are first entertained in the process of looking at some object or
scene has just been offered by Biederman (1987). Building on theorizing by
computer scientists (notably Binford, 1971, 1981) who are concerned with how
machine vision might be achieved, Biederman (1987) presents a very
persuasive argument to the effect that we have a small vocabulary of
primitive shapes; that these combine to form the objects that we can
recognize in the very first moment of inspection (and that, we should add,
means very largely in peripheral vision); and that our recognition of these
units does not depend on first perceiving their distance and depth. This is
not the old argument that we first perceive a 2D depthless world, and then
construct a 3D world from it: In order to be useful, the units must be
shapes (e.g., see Figure 8) that remain essentially intact in the 2D field
of view, regardless of their slant or orientation to the line of sight.
Object and scene perception at this level are neutral with respect to the
third dimension, not flat or 2D.

We can now understand how it might be that we can recognize 3D scenes
in pictures that look totally flat: The recognition of the woman in Figure
7 may proceed quite independently of depth and orientation if we indeed
recognize shapes through depth-neutral components. This is the second
mechanism by which painters can learn to "see the world flat" (i.e., to
perceive what 2D pattern would serve as a surrogate for some 3D layout of 3D
objects): by analyzing the scene into some vocabulary of depth-neutral
"component images." Finally, this analysis helps us to understand what
happens at the level of the local features in such paintings as Figures 6
and 7: Intersections 6a and 7a are quite flat by themselves, but they do
not actually contradict 3D intersections at those places. Thus, they
provide no perceptual challenge to the flatness of the picture nor to the
tridimensionality of what it represents.

Interactions between features and context

How whole and part interact has been a continuing and fundamental
problem in perceptual psychology, and it is neither solved, nor sidestepped,
by separating them into different domains (e.g., component images and
features). In Figure 9B, we have a set of black features on a white disk;
in Figure 9A (adapted from Bradley, Dumas, & Petry, 1976), is an ordered
arrangement of such discs. If we say that the arrangement consists of the
views of a cube made of black wires, that is readily perceived in Figure 9A, and indeed, the cube will soon appear to reverse in depth, thereby offering us assurance that the perceived structure of the resulting object's image is tridimensional, although most of the cube is obscured by the apertures through which the intersections are seen. But the intersections taken by themselves do not appear to reverse in this way. The Bradley et al picture tells us that not all images are neutral with respect to 3D, and it seems to show us that it is the 3D structure of a more inclusive object image that (at least in this case) determines the spatial functions of the local features.

We know that in the mind of the viewer, some images are more readily visualized and manipulated than others (Peterson and Weidenbacher, 1987). The differences seem to lie in the content or the meaning of the images. Similarly, it is difficult to answer rapidly whether the shape in Figure 10a is to be found in Figure 10b, but much easier to find 10c in 10d (which is clearly something that Matisse was experimenting with here, as in other paintings and cutouts). In our version of the Rubin figure-ground stimulus shown in Figure 11, the central shape is relatively dominant in 11A, and alternates more readily with the surround in 11B, when the upright images of two faces can be readily fit to the surround. The surround seems less dominant when the figure is inverted (Figure 11A). Quantitative evidence to this point comes from experiments in which viewers are asked to try to see the center or the surround as figure: Viewers are both more successful at seeing the surround as figure and less successful at seeing the center as figure when the picture is upright rather than inverted (Peterson & Weidenbacher, 1988). The geometry and configurations are the same in the upright and inverted conditions; both the local component images and the overall object images may be different.

We need to know much more about the nature and content of the images that provide the perceptual context for our individual glances and about how those images interact with the features provided in each glance. We also need to know whether the components that underlie that recognition, and that must comprise much of the content of our thoughts about objects, are anything like the finite set that Biederman proposes, and are indeed neutral as to depth as our present argument assumes. Given that artists have, by our account, wrestled with these problems (less systematically, but more freely than psychologists) for centuries, we will surely continue to find their work a fertile source of perceptual theory.
References


Cautions

Figure 1: Local contrast in chiaroscuro: The apparent lightness of the figures at their rightmost edges is greatly enhanced by the heightened contrast between the figures and their backgrounds (i.e., by the juxtaposition of bright and dark regions). Without a sharp contour or edge, brightness differences are perceived as differences in illumination (i.e., as modeling), not as lightness differences: Regions a and g are brighter than regions b and f, respectively, but the figures appear to be of homogeneous lightness (or reflectance) except for the black bow. It is important to know that the contrast at g is virtually unaffected by the brightness at g. The section between g and f is elaborated from a sketch of Seurat's *The Black Bow* (adapted from Hochberg, 1979).

Figure 2. Large pictures that are made to be looked at from more than one viewpoint, like A (a sketch of David's *The house of the Sabine Women*, adapted from Hochberg, 1984), are as surrogates inconsistent from one region to the next. The round shield is consistent only with one viewpoint (y in B); when viewed from some other point (y'), it would have to be drawn as an ellipse on the picture plane (P) to be an optical match for a round shield, and is therefore inconsistent with the frontal view at y. From y', the circle in the painting would be consistent with a shield that is an ellipse like that in Figure 8B, which is not what it looks like from y' or any other viewpoint. Of course, the same inconsistencies exist in all of the objects and figures, but those they are not so easy to recognize as in the case of a circle.
Figure 3. When attention is near the right-hand end, surface a is perceived as being viewed from above; with attention at the other end, surface b is perceived as being viewed from below. The intermediate panels are relatively free to be perceived at either orientation, depending on where attention has been directed, and the inconsistency between the two ends is only perceived after some study of the picture.

Figure 4. The figure represents a partially covered skeletal cube, seen from below and left. Both as a real moving object, and as a static picture, the spatial structure is unambiguous when attended in one place, a, yet reversible and ambiguous when attended at another nearby place, b (Hochberg and Peterson, 1987; Peterson and Hochberg, 1983).

Figure 5. Some of the alternative ways, flat as well as volumetric, in which the pictures at left (A, E, and K) which are the intersections of Figure 4) can be "read" or interpreted. Some of those readings would require a precise and therefore unlikely accidental alignment in order to provide the picture at left (e.g., B, D, G, L-N). The feature at A is the traditional depth cue of interposition, strongly evoking the continuation of the interrupted (horizontal) line behind the uninterrupted one (vertical), as for example in G.

Figure 6. Especially since impressionism, artists often deliberately replace features that would provide "illusionistic depth" (i.e., an optical match) with more neutral features or even with features that oppose a depth reading. In this sketch of Cezanne's painting of his father (adapted from Hochberg 1984), most of the intersections that would provide unequivocal depth (e.g., junction g; see A in Figure 5) are replaced by intersections in which the edges form smoothly continuing lines at the junction, and which would require a precise alignment by accident if they had arisen from surfaces at different distances (e.g., a, b). The pictures remains three dimensional in what it represents, and particularly so in peripheral vision, but appears flat whenever the gaze falls near one of these altered junctions.

Figure 7. Only one of the intersections in this sketch of Vuillard's "Woman in Black" (adapted from Hochberg, 1980) retains the unsinged depth cue of Figure 5a, and that junction, g, must operate against the represented three dimensional relationships: it would have the interrupted line (here, the vertical) continue behind the uninterrupted edge of tablecloth and skirt. And in g, as in the other intersections (a, b), unlikely coincidences of alignment would be needed to provide the drawn intersections if they arose from objects at different distances. Despite the absence of depth information, what the picture represents remains recognizable.

Figure 8. Some aspects of form are largely or wholly independent of accurate knowledge of depth or orientation. A sphere provides the same outline (A) from any viewpoint, and although an ellipse (B) may be provided in the picture by a circle viewed from a slant, and although an ellipsoid at a slant might be represented by a circle (e.g., B as an object would fit the circular picture of the shield when viewed from y' in Figure 2B), no change in distance or orientation would make it a square, G. Biederman (1987) has proposed that a limited set of primitive perceptual volumetric components or geons (e.g., D; E), in various combinations, underlie the processes that occur in the first moments of object-recognition.
Figure 9. When regarded as the corners of a wire cube viewed through a set of apertures (a), depth is evident in this picture (adapted from Bradley et al., 1975), as is evidenced by abrupt spontaneous reversals. When the disks are regarded as individual features, which is made easier by destroying the overall implied structure (as in b), both depth and reversals are drastically reduced or absent.

Figure 10. See text. The drawings in b and d are inverted and upright sketches, respectively, of the main features of Matisse's "La Danseuse".

Figure 11. Figure-ground reversal should be easier and more complete in A than in B. See text.
NB:

Ed. should be printed so that the circle at
X is 2K are about 16 mm.

Fig. 5

Fig. 6