Cavanagh, P. (1991). What's up in top-down processing? In A. Gorea (ed.) *Representations of Vision: Trends and tacit assumptions in vision research,* 295-304.

What's up in top-down processing?

Patrick Cavanagh

Department of Psychology, Harvard University,
33 Kirkland Street, Cambridge Massachusetts 02138, U.S.A.

Much of the work in vision in the last decade has examined what low-level vision tells high level vision. Cues such as optic flow, depth of focus, and contour intersections have been shown to be useful, reliable correlates of the 3-D structure of a scene. However, the retinal image is often ambiguous. Figure 1, for example, can be seen as either a duck or rabbit. It does not appear as a hybrid, though: only one or the other of these interpretations is seen at any given moment. In addition, the final percept — duck or rabbit here — often contains more structure than is available in the retinal image. The relative positions of the two ears of the rabbit, for example, one near, one far, are not specified in the image yet they are available in the percept and determined by our 3-D knowledge about rabbits. In these instances, the interpretation must have been influenced by top-down processes. Clearly, top-down processing speeds the analysis of the retinal image when familiar scenes and objects are encountered and can complete details missing in the optic array.

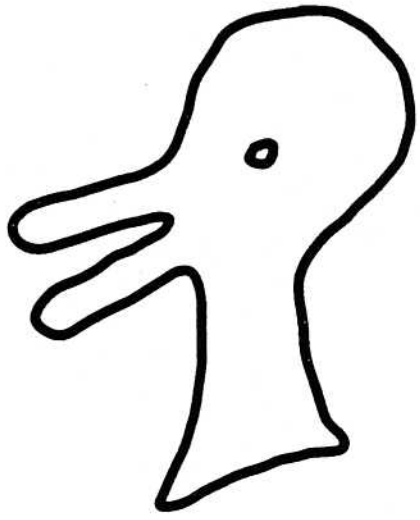

Fig. 1. An ambiguous figure that can be seen as either a rabbit (apparently staring into the sky) or a duck. The 3-D structure attributed to the various parts of the image changes in the two interpretations but the 2-D information — the location of the contours — is unaffected.

Top-down processing requires that something be up top, of course, and there have been only vague ideas about the representations that might be involved and the means by which they would influence perception. Basically, the tacit assumption is that something is up top and that this something solves otherwise puzzling visual problems. In order to start an examination of these processes, I shall describe

a particular stimulus that can only be interpreted with the aid of top-down processing but for which there is, initially, nothing up top.

In analyzing this stimulus, I shall concentrate on the early stages that lead from the initial 2-D representation on the retina to object recognition. Current bottom-up approaches to vision (see Fig. 2) assume that the 2-D representation leads to an internal 3-D model before the stimulus is identified (Marr, 1982; Biederman, 1987). Work that I have just begun takes a different approach, suggesting that object parts and boundaries should not be explicitly identified at such an early stage and that matching of raw 2-D views may be the most effective way to make the initial memory contact (Fig. 3). The basic question is the level at which image elements should be labeled as particular image tokens, whether edges, curves, 2-D shapes or 3-D volumetric features. This labeling commits the visual analysis to treat image elements in specific ways in subsequent processing and it can be disadvantageous to make this commitment prematurely.

Image ⇒ Contours ⇒ Parts ⇒ 3-D Model ⇒ Object

Fig. 2. Image analysis, as proposed by Marr (1982) and Biederman (1987) for example, proceeds from from the image through a 3-D model before indexing memory to identify the object.

I shall examine the possibility that object recognition begins, not with the construction of a 3-D model, but with a crude match of 2-D views to internal prototypes. The prototype that has the best match then guides the construction of an internal 3-D model. In other words, recognition may start with a quick table look-up process, operating on principles completely different from those implied in Figure 2. The results of this process are then "up top", available to initiate top-down processes.

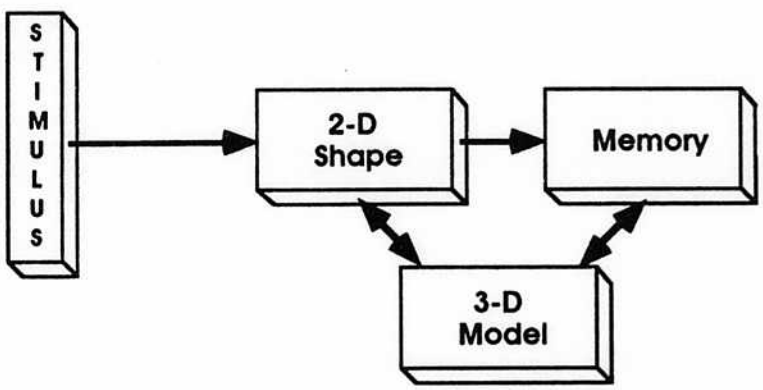

Fig. 3. In special cases, recognition may start with an initial match of 2-D image information against memory prototypes which then guide the construction of a 3-D model.

Figure 3 is presented as a sketch of the components of the visual system that are relevant for this paper. The diagram is principally concerned with the flow of shape information and not, for example, with the analysis of the position of objects or their displacement in the scene. These analyses may form part of a second processing system — the "where" system proposed by Mishkin, Ungerleider, and Macko (1982). The figure does not show the direct contribution of binocular disparity, motion parallax, convergence cues or gradients (shading, optic flow, etc.) to the 3-D model, but these are not considered in detail in this paper.

The stimuli I shall use to examine early recognition processes are figures where shape is defined by shadows (Fig. 4, righthand panel). There is more to this image than just contours, however: Many people presented with only the contours (Fig. 4, lefthand panel), for example, cannot identify it. Nor can they specify which contours might be shadow contours and which might be object contours. However, it is essential to identify the cast shadow borders in an image (this point is discussed in more detail below) because they are generally unrelated to the object contours and they seriously disrupt the interpretation of the image if they are confused with object contours — as they are in the lefthand panel of Figure 4. Clearly, any process (or person) that tries to identify parts or volumetric features in such an

image would perform poorly without some knowledge of the object. I shall argue that any approach that labels the borders in this type of image before identifying the object will be faced with several spurious borders and these extra borders will seriously disrupt the segmentation of object parts. In most natural images, there are many redundant cues that help to identify which contours are shadow contours and which are not. The images that I are studying are therefore not intended to be representative of natural images but are an especially difficult type of image that humans are nevertheless able to interpret remarkably well.

Fig. 4. Contour information alone may be insufficient for 3-D interpretation. The left panel contains the same contours as the right but is difficult to recognize on its own.

Shadows are useful for recovering 3-D information as in the righthand panel of Figure 4, where the surface structure is revealed only by shadows (Cavanagh & Leclerc, 1989). Even though shadows are useful they suffer from several ambiguities in an image like Figure 4. First, it is not evident in the image whether an area is dark because of dark pigment or because of a dark shadow (of course, the figure *is* just dark pigment even though our interpretation attributes the darkness to shadows). Different kinds of shadow contours — attached or cast, see Figure 5 and 6 — play very different roles in reconstructing 3-D structure and nothing in the high contrast images that I am using distinguishes between these two types. Finally, the interpretation of the shadows depends critically on knowing the direction of the illuminant and this is not specified in the image in any explicit manner. In one sense, we have to know where the shadows are before we can identify the direction of the illuminant but we have to know the direction of the illuminant before we can discover the shadows.

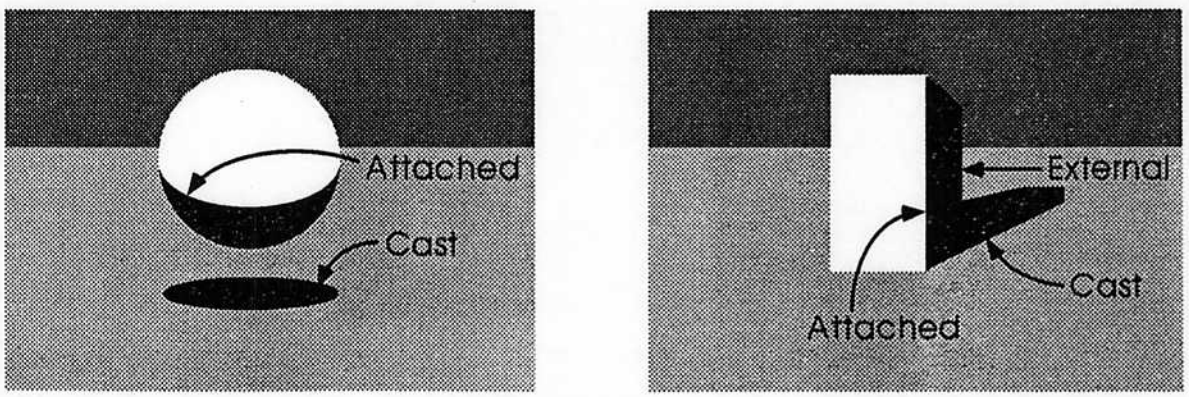

Fig. 5. Shadows have two types of borders: attached borders where the direction of the illumination is perpendicular to the surface normal (the light just grazes the surface); and cast borders where the shadow cast by one surface falls on a second surface. An object's external borders are only visible where the background and the object have a different brightnesses.

The essential goal in discovering the shadows in a figure is identifying the cast shadow borders. These borders have a special status in images because they do not correspond to any discontinuity in the object but to a discontinuity in the illumination. The cast shadow border is not a material border and basically needs to be ignored in order to patch together the pieces of surface that actually belong together.

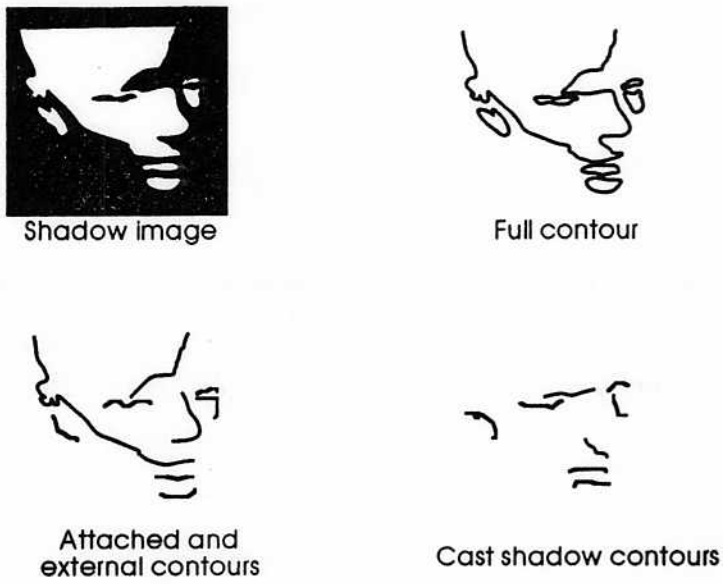

Fig. 6. The contours of the shadow image (top left) are difficult to interpret on their own (top right) but the attached and external contours (bottom left) are easily recognized. The cast shadow contours (bottom right) present a meaningless jumble of lines.

I claim that the recognition of shape from shadow in a stimulus like that in Figure 4 starts with an initial 2-D matching process. The reason is that there is no alternative for these images: the stimulus cannot be interpreted directly from image data without knowing what the object is. As mentioned above, the interpretation of the image requires that the different types of borders — object border, cast and attached shadow borders — be identified or parsed in the image. The only way to parse the image contours into attached or cast borders without any top-down guidance is presented in Figure 7 and I shall show that this parsing fails on images that we can nevertheless interpret.

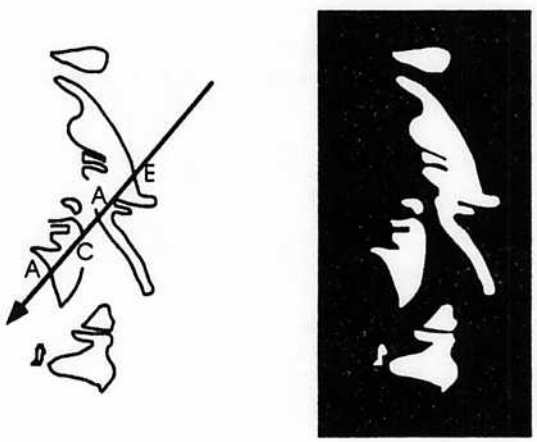

Fig. 7. A line parallel to the direction of the illuminant can uniquely label the light to dark transitions which always fall in the same sequence: external (E), attached (A), cast (C), attached, ... [e.g. {E,A},{C,A},{C,A},...] no matter where the line is traced.

If we assume that the direction of illumination can be determined (in the extreme, all directions can be examined until a consistent interpretation is found), then it is possible to distinguish cast from attached shadow borders in the image and from the attached shadow borders to recover the object's 3-D structure. Figure 7 demonstrates that along any line parallel to the direction of illumination, borders alternate in a fixed order. If the background is dark, the order is always external, attached, cast, attached, cast, attached, and so on, with cast and attached repeating in pairs. Using this rule, the border type can be identified throughout the image.

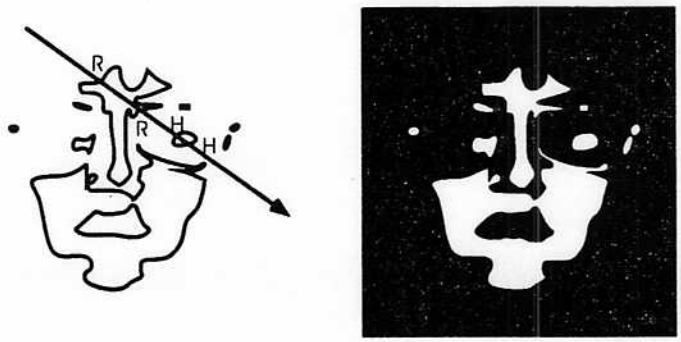

Fig. 8. In this image, some regions may be dark either because of low reflectance (R: hair, dark glasses, lipstick) or because of shadows. Other regions are light because of high diffuse relectance (skin) or because of specular relections (H: highlights on the glasses). The labeling scheme of Figure 7 fails here because the light to dark transitions can no longer be uniquely identified.

However, this parsing only works in an image where the borders are all shadow borders. If the image also contains reflectance and highlight borders (shown as R and H in Figure 8), no consistent labeling is possible. Shadow borders cannot be distinguished from reflectance or highlight borders in the image even if the direction of the illuminant is known. Therefore, the image contours cannot be parsed. ***The image is nevertheless recognizable.*** I conclude that some information other than that available in the image must guide the interpretation. But this cannot be true either since most people perceive the righthand panel of Figure 8 as a face without being told what it is beforehand. There is no other information available. Even if I want to invoke top-down processing, there is nothing up top!

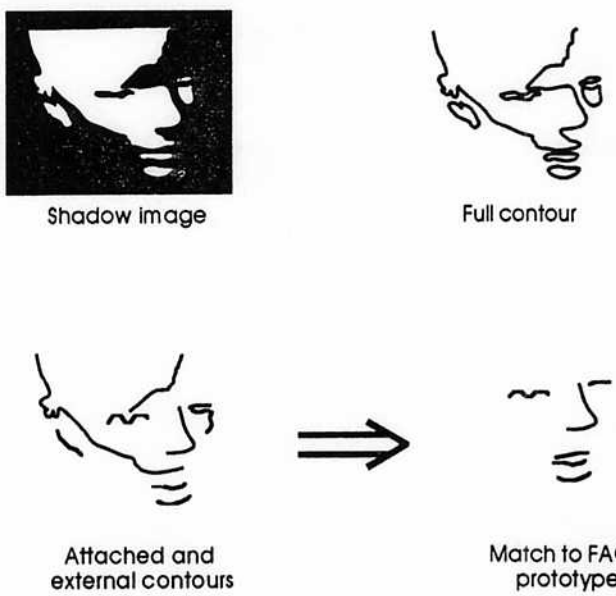

Fig. 9. The contours of the shadow image (top right) contain an easily recognized subset that can match to characteristic contours of a face prototype.

How can we get knowledge about the object before we recognize it? I suggest that a very different type of process performs a rapid, crude match against the image data to identify the type of stimulus. This hypothesis or prototype then guides the interpretation of the image. Note that even though the contours of shadow figures such as those in Figure 4, 7 and 8 are difficult to recognize, there is sufficient contour information in them for a match to simple prototypes in memory. As we saw in Figure 6, there is a subset of contours in these figures that is easily recognizable: the attached shadow and external contours. It cannot be known beforehand which contours are attached shadows and which are cast but a matching process that is capable of recognizing subsets of contours in the presence of irrelevant contours could extract the best match.

Figure 9, for example, contains some characteristic contours of the object mixed in with, and indistinguishable from, many irrelevant cast shadow contours. An initial 2-D match is therefore advantageous only if image information can be matched to memory prototypes without first selecting which subset of image contours must be object contours. Matched filtering techniques developed for 2-D recognition in the 1960s demonstrate possible methods for performing the necessary steps: identify targets based on a partial set of contours in the presence of unrelated contours (Bieringer, 1973; Van der Lugt, 1964); specify which contours participated in the match; fill in the missing contours (Collier & Pennington, 1966); and identify the residual image contours unrelated to the matched object (Caulfield, 1974, residual contours must then be explained by other scene components such as shadows, and other, occluding objects). This match process must be able to match against all possible prototypes simultaneously. It is probably most reasonable to think of storing only a small set of characteristic protoptypes — a face, a car, a cylinder etc. — and not a prototype for each possible instance of these broad classes.

Thus a rough 2-D match could select the best candidate object — a face, boot, hat, or whatever. The 3-D information stored with this prototype could then guide the construction of an internal 3-D model — verifying that it is consistent with the image contours and resolving ambiguities of shadow borders, occluded objects and incomplete contours. A match between a subset of the image contours (and these will only be among the external and attached shadow contours such as, in Figure 10, the nose profile or the lips) has the important consequence of identifying the *residual* contours — the contours not explained by the prototype. A second process must then "explain" these residuals. On the right in Figure 10, the residual contours might be attributed to shadows or material boundaries or to other objects that are occluding the object that matched the prototype. These hypotheses for the residuals must then be verified in the image to see whether the original hypothesis can be maintained; for example, if a residual contour is to be interpreted as a shadow border, the image should be darker on the shadow side of the border. Figure 11 depicts this process for one piece of residual contour from the right side of Figure 10.

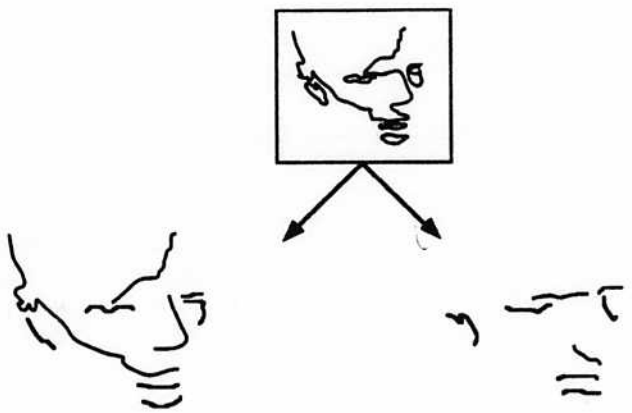

Fig. 10. The image contours include the informative attached shadow and external contours on the left and the more or less irrelevant cast shadow borders on the right. Only the contours on the left can be expected to participate in a match to a simple prototype although certainly not all those shown here would be involved — perhaps only the nose, eyes, lips and ear contours. Those that don't participate in the match are the residuals that must be explained through different assignments in the scene. These will always include the cast shadow contours shown on the right.

If there is insufficient support in the image for the 3-D aspects of the initial prototype, it would be discarded. This is the probable fate of a face prototype for the top image of Fig. 10 — there are no dark regions to support 3-D shadow explanations of the many contours that are not characteristic face (lips, nose, forehead, etc.) contours. The prototype with the next best 2-D match would then be selected to guide the 3-D modeling. However, there is no other obvious 3-D prototype for this particular image. The only remaining explanation is that the contours are all material borders so that the enclosed areas are seen as separate 2-D islands.

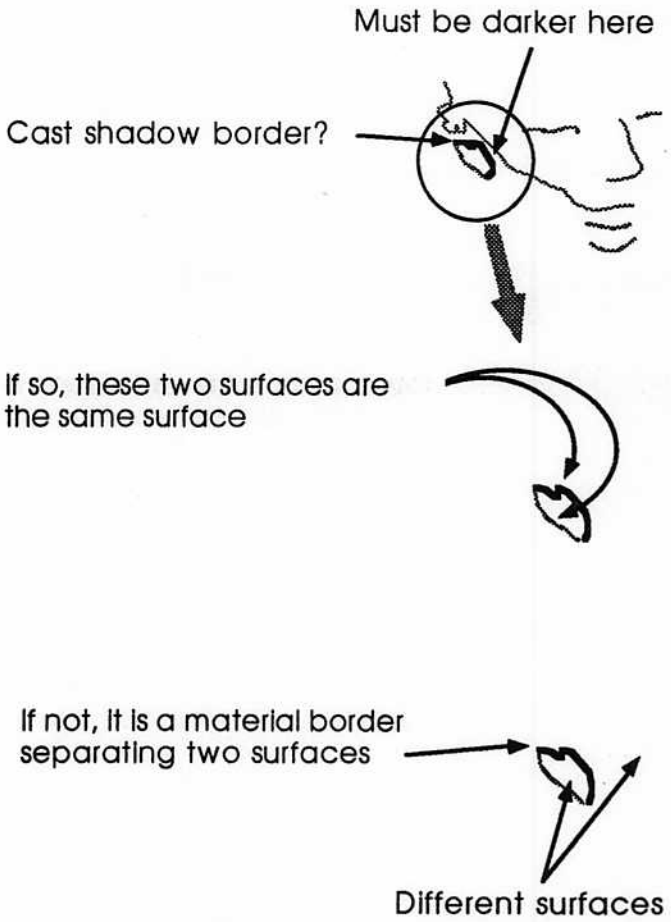

Fig. 11. Each residual contour must be explained by an additional scene property. If the particular contour shown as a thick line in the ear area is to be labeled as a cast shadow border, the adjacent region to the right must be darker than that to the left.

The memory prototypes required for this inital 2-D match are quite unlike those required by recent approaches (Marr, 1982; Biederman, 1987) where memory prototypes are object-centered and can serve to identify any arbitrary 2-D view of an object in the scene. In contrast, if the initial match is based on stored 2-D views, each object prototype would have to have numerous 2-D views as part of its representation in memory. The number of necessary views would have to be especially large if size- and orientation-invariant coding is not used by the visual system (although see Cavanagh, 1984, 1985). In one sense, this is not an insurmountable problem even in the worst case since 2-D matching is ideally suited to the massively parallel processes hypothesized in neural net memory systems (Anderson, Silverstein, Ritz & Jones, 1978; Kohonen, 1977).

Note that the 2-D views bundled together as memory prototypes are viewer-centered (see Fig. 14). There is evidence that the visual system does, in fact, operate on viewer-centered representations and not 3-D object models when accessing memory. Rock and his colleagues (Rock, DeVita & Barbeito, 1981; Rock & DeVita, 1987) have demonstrated that views of wire-frame objects seen from different directions are reliably identified only when they have the same retinal projection, indicating that 2-D viewer-centered representations may mediate recognition.

In the test figures that I have used, top-down processing operates from a rough prototype for the object in the image and, in cases where there is no other source, the prototype is provided by a 2-D match to the image. The model of recognition therefore has three stages: first, a 2-D match of the image against memory prototypes selects the "best" prototype; second, this prototype guides the construction of a 3-D model, checking the image for consistent support as the interpretation develops in detail; finally, the completed 3-D model corresponds to recognition.

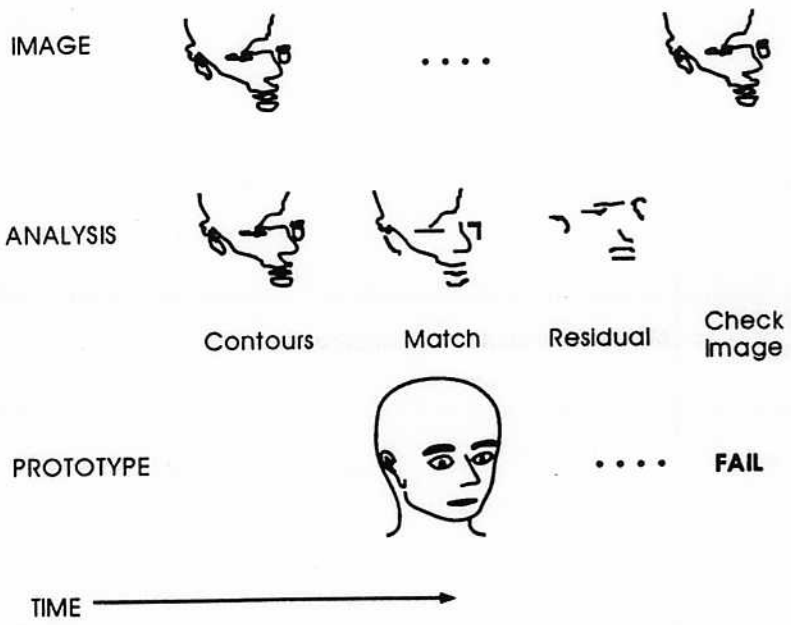

Fig. 12. The analysis of the image depicted on three levels: the displayed image which remains unchanged after its initial presentation; the analysis steps of extracting image contours, matching to memory prototypes, identifying the residual contours not contained in the prototype, and checking image for support for local interpretations of each residual contour; finally, the prototype is available to direct top-down processing following the match to memory. The face prototype fails in this test image containing only contours since no support can be found for 3-D interpretations of the residuals (e.g. shadow borders or occluding objects).

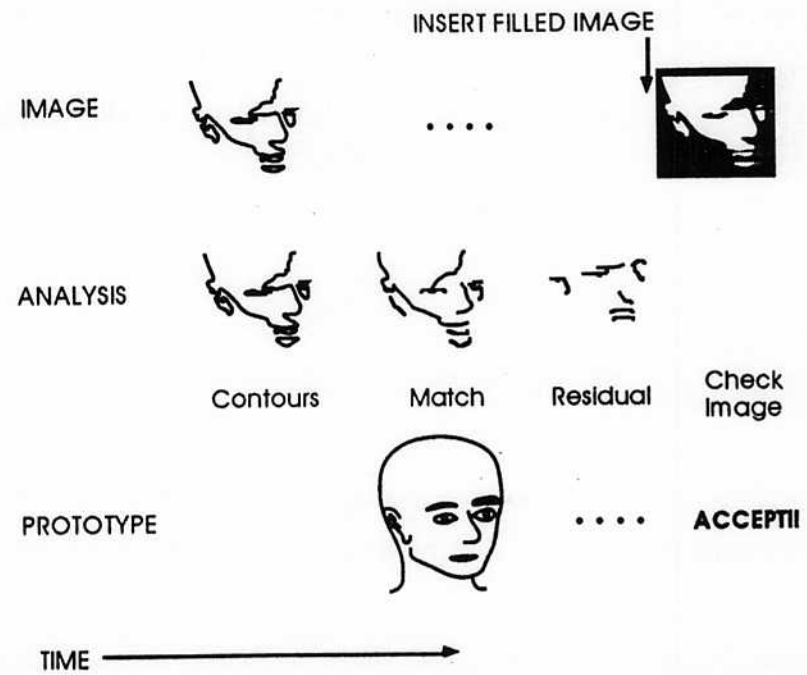

FIG. 13. In this example, the contour image is replaced in the display with a filled image just prior to verifying the residuals. Brightness levels are appropriate for shadow explanations of the residuals so the 3-D structure of the prototype is accepted even though only contours were presented initially.

If this model of early recognition is correct, it has an interesting and unexpected consequence. If the contour version of a shadow figure (e.g. the lefthand panel of Figures 4, 7, or 8) is presented, the 2-D match to the prototype should occur *even if the figure is not recognized.* The match occurs as shown in Figure 12, but it is rejected later because of lack of support for the residuals. It may be possible therefore, to switch the image from a contour version to a filled version at the appropriate moment and obtain recognition in the same total time — as if the filled version were present from the start (Figure 13). I have begun experiments to test this prediction.

In summary, what's up top? In the examples that I have presented here, it appears that some type of rough prototype may be the representation that guides the interpretation of the image. In most natural images where many redundant cues are available, the prototype may be chosen based on 3-D information. In the high-contrast images that I have used, no 3-D information is available from the image either directly or through pictorial cues such as perspective, contour intersections, or deep concavities. I claimed that for these images an initial 2-D match selected the best prototype to guide image interpretation. This initial match does not constitute recognition, however, and an experiment was suggested to demonstrate that this early match occurs even for stimuli that cannot themselves be recognized. The stored prototypes may be limited to basic object types, some as complex as a face and others as simple as a cylinder, but do not require a prototype for each instance of a class.

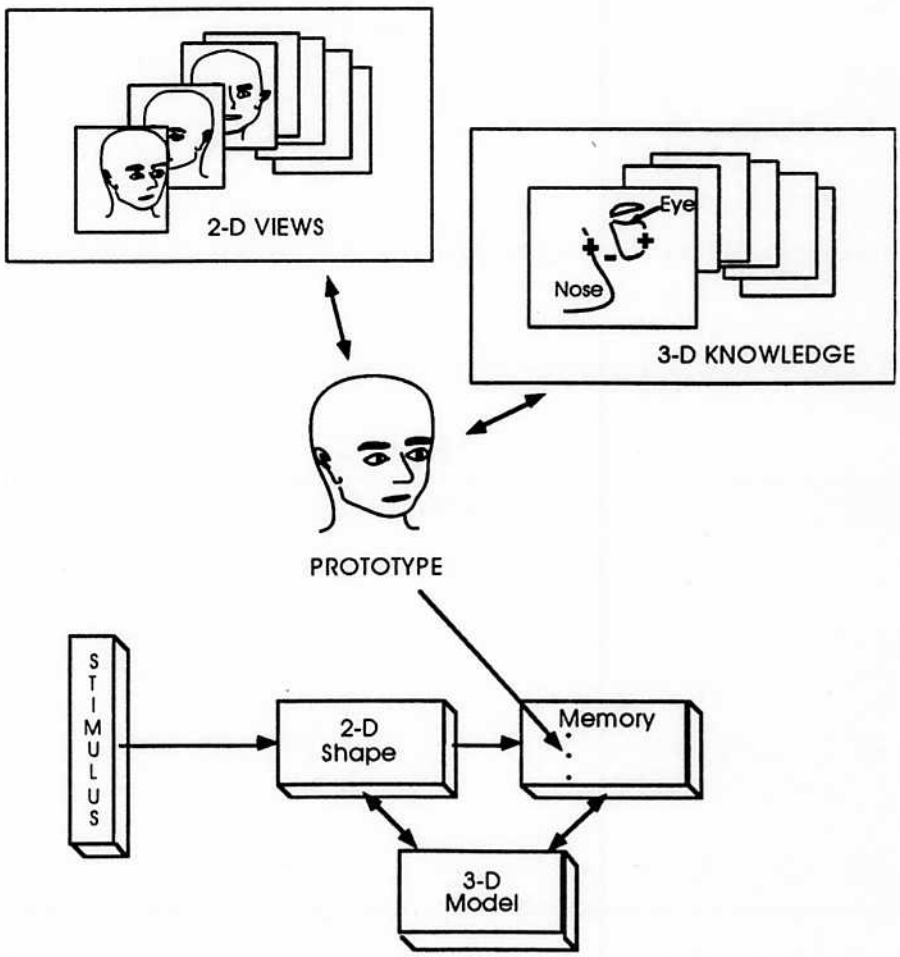

Fig. 14. Memory prototypes include a set of 2-D views from several viewpoints and 3-D knowledge about the object such as directions of curvature along the object contours that are visible in the various 2-D views as well as identification of the parts. Prototypes for those parts would contain additional information as well. The prototypes could be identified from 3-D information available in the image or in its absence from a 2-D match to individual views. Once the prototype is identified, it guides the completion of the 3-D model of the object.

REFERENCES

Anderson, J. A., Silverstein, J. W., Ritz, S. A.,& Jones, R. S. (1977): Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review* **84**,413-451.

Biederman, I. (1987): Recognition-by-components: A theory of human image understanding. *Psychological Review* **94**, 115-147.

Bieringer, R. J. (1973): Optical correlation using diffuse objects. *Applied Optics* **12**, 249-254.

Caulfield, H. J. (1974): The rabbit-and-hat problem. In P. Greguss (ed.) *Holography in Medicine*, Surrey, U.K.: IPC Science and Technology Press, 3-26.

Cavanagh, P. (1984): Image transforms in the visual system. In P. C. Dodwell & T. Caelli (eds.) *Figural synthesis*. Hillsdale, N. J.: Lawrence Erlbaum Associates, 185-218.

Cavanagh, P. (1985): Local log polar frequency analysis in the striate cortex as a basis for size and orientation invariance. In D. Rose & V. G. Dobson (eds.) *Models of the visual cortex*. London: John Wiley & Sons, 85-95.

Cavanagh, P. & Leclerc, Y. (1989): Shape from shadows. *Journal of Experimental Psychology: Human Perception and Performance* **15**, 3-27.

Collier, R. J. & Pennington, K. S. (1966): Ghost imaging by holograms formed in the near field. *Applied Physics Letters* **8**, 44-46.

Julesz, B. (1971): *Foundations of Cyclopean Perception*, Chicago, University of Chicago Press.

Kohonen, T, (1977): *Associative memory*. Springer-Verlag, Berlin.

Marr, D. (1982): *Vision*. Freeman: San Francisco.

Rock, I., Di Vita, J. & Barbeito, R. (1981): The effect on form perception of change of orientation in the third dimension. *Journal of Experimental Psychology: Human Perception and Performance* **7**, 719-731.

Rock, I. & Di Vita, J. (1987): A case of viewer-centered object perception. *Cognitive Psychology* **19**, 280-293.

Mishkin, M., Ungerleider, L. G., & Macko, K. (1983): Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences* **6**, 414-417.

Van der Lugt, A. B. (1964): Signal detection by complex spatial filtering. *IEEE Transactions on Information Theory* **IT-10**, 139-154.