The Problem of Naming Shapes: Vision-Language Interface

by

R. Bajcsy*
and
A.K. Joshi*

Computer and Information Science Department
University of Pennsylvania
Philadelphia, PA 19104

1. Introduction

In this paper, we will pose more questions than present solutions. We want to raise some questions in the context of the representation of shapes of 3-D objects. One way to get a handle on this problem is to investigate whether labels of shapes and their acquisition reveals any structure of attributes or components of shapes that might be used for representation purposes. Another aspect of the puzzle of representation is the question whether the information is to be stored in analog or propositional form, and at what level this transformation from analog to propositional form takes place.

In general, shape of a 3-D compact object has two aspects: the surface aspect, and the volume aspect. The surface aspect includes properties like concavity, convexity, planarity of surfaces, edges, and corners. The volume aspect distinguishes objects with holes from those without (topological properties), and describes objects with respect to their symmetry planes and axes, relative proportions, etc.

We will discuss some questions pertinent to representation of a shape of a 3-D compact object without holes, for example: Is the surface aspect more important than the volume aspect? Are there any shape primitives? In what form are shape attributes stored?, etc. We shall extensively draw from psychological and psycholinguistic literature, as well as from the recent AI activities in this area.

2. Surface and Volume

In this section, we will investigate the relationship between the surface aspect and the volume aspect from the developmental point of view and from the needs of a recognition process. By doing so, we hope to learn about the representation of shapes. Later, we will examine the naming process for shapes and its relation to representation.

This work has been supported under NSF Grant #MCS76-19465 and NSF Grant #MCS76-19466

There is evidence that a silhouette of an object, that is its boundary with respect to the background, is the determining factor for the recognition of the object (Rock 1975, Zusne 1970). If we accept the above hypotheses then the fact that the silhouette is a projected outline of the 3-D object implies that the recognition of the 3-D object at first is reduced to the recognition of a 2-D outline. This is not entirely true, however, as Gibson (Gibson 1950) has argued. According to Gibson's theory, the primitives of form perception are gradients of various variables as opposed to the absolute values of these variables. From this follows the emphasis on perceiving the surface first and the perception of the outline only falls out as a consequence of discontinuities of the surface with respect to the background.

We are pursuaded by Gibson's argument and regard the recognition process as starting with surface properties; Miller and Johnson-Laird (Miller & Johnson-Laird 1976) have suggested some surface predicates as possible primitives, such as convex, concave, planar, edge, and corner. The 2-D outline is furthermore analyzed as a whole according to the Gestalist and some salient features (Pragantz) are detected faster and more frequently than others (Koffka 1935, Goldmeir 1972, Rosh 1973); such pragmatic features are for example, rectangularity, symmetry, regularity, parallelness, and rectilinearity.

Piaget also argues (Paiget, Inhelder 1956) from the developmental point of view that children first learn to recognize surfaces and their outlines, and only later, after an ability to compose multiple views of the same object has been developed, they can form a concept of its volume.

Volume representation becomes essential as soon as there is motion of the object or of the observer. Note that the salient features of 2-D shapes are invariant under transformations such as rotation, translation, expansion and shrinking. Features with a similar property must be found in the 3-D space for the volume representation. We feel that the most important feature is symmetry. Clark's work seem to support this (Clark 1975); he shows that in language space as in the perceptual space we have 3 primary planes of reference: ground level; vertical: left-right; vertical: front-back. While the ground level is not a symmetry plane, the two vertical ones are symmetry

planes. The fact that the ground level is not a symmetry plane is supported by the experiments of Rock (Rock 1973), who has shown that some familiar shapes are hard to recognize with 180° rotation with respect to the ground level. After a careful examination of the relevant literature to date, we find that there is a claim that we can recognize shapes via some features which are more salient than others. But does it follow from this that shape is an independent attribute like color, or is it a derived concept from other features?

In an effort to answer this question, we set out to examine labels of shapes in the hope that if there are any shape primitives (other than angles, edges, parallelness, and the like) then they may show up in labels describing more complex shapes. One immediate observation we can make is that there are very few names which only describe a shape, such as triangle or rectangle. More commenly, label of a shape is derived from the label objects which have such a typical shape, for example, letter-like shapes (V, L, X), cross-like shape, pear-like shape, heart-like shape, etc. A special category of labels are well defined geometric objects, such as circle, ellipse, sphere, torus, etc. The question is whether we store for every shape a template or whether there are any common primitives from which we can describe different shapes.

In addition to the 2-D features mentioned earlier, primarily 2-D features, we do use 3-D shape descriptions (primitives) such as: round, having 3 symmetry planes and all the symmetry axes approximately of the same length, elongated, where the size in one dimension is much longer than the two remaining, thin, where the size of one dimension is much smaller than the other, etc. Note that many of these descriptions are vague, though often there more accurate shape labels available; for example, cone stands for an elongated object with two symmetry planes, a circular crossection, and sides tapering evenly up to a point, called appex (Webster's dicitionary).

We believe that there are some descriptions of shapes which are more primitive than others; for example, round, elongated, thin, flat, circular, planar, etc., as opposed to heart-like, star-like, and so on. As pointed out earlier, these latter descriptors are derived from the names of previously recognized objects. When we use these descriptions during a recognition process we do not necessarily match exactly all features of the template shape to the recognized shape, but rather we depict some characteristic properties we associate with the given label, and only these are matched during the recognition process. In this sense, we approximate the real data to our model and primitives. The labels which encompass a more complex structure of these properties (like cone, heart, star, etc.) when they are used in describing other shapes, are used as economical shorthand expressions for the complexity that these shapes represent. (This appears to be related to the codability notion of Chafe (Chafe 1975)).

3. Analog and Propositional Representation

In this section, we will discuss certain issues concerning the form of the stored information, necessary not only for recognition purposes (matching the perceived data with a stored model) but also for recall, and introspection of images.

There are two questions:

1. At which level the analog information is converted to propositional (verbal or non-verbal) and after this conversion, is the analog information retained?
2. How much of the propositional information is procedural and how much structural?

For simplicity, we will regard analog information in our context as picture points, or retina points. Any further labeling, of a point or of a cluster of points, such as an edge, line, region, etc. leads to derived entities by one criterion or another and therefore may be regarded as propositional.*

At this point, it is appropriate to point out that any such unit as an edge, line or region can be described in at least two different ways; one is structural or organizational, and the other is parametric or dimensional. Structural information refers to the organization of perceptual elements into groups. Figure-ground and part-whole relationships are paradigm examples of structural information. Parametric information refers to the continuous values of the stimulus along various perceivable dimensions. Color, size, position, orientation, and symmetry, are some examples of parametric information.

We are not advocating that these two types of information are independent (cf. Palmer 1975). It is, for example, a well known experience that by changing drastically one dimension (one parameter) of an object (say a box), one can change the structure of the object (in this case, it becomes a wall-like object). However, we do wish to keep the distinction between structural and parametric information. The importance of this distinction is that while structural information is inherently discrete and propositional, parametric information, is both holistic (integral) and atomic (separable). The fact that parametric information is separable is quite obvious if we just recognize that different parameters represent clearly distinguishable different aspects of the visual information. For example, color, size, position, etc. On the other hand all these parameters are represented holistically in an image, and can be separated only by feature (parameter) extraction procedures (Palmer 1975).

Parametric information is separable, however, the question is whether each parameter-feature

---

* The distinction is not really as sharp as stated here. One way to make the distinction is to look at the closeness with which a transformation of a representation parallels the transformation of the object represented. The closer it is the more analog the representation is.

has continuous or discrete values. Continuous values would imply some retainment of analog information (Kossylyn 1977), while discrete values would not. Opponents of the discrete value representation argue that a) the number of primitives needed would be astronomical, and b) the number of potential relationships between primitives would be also very large (Fishler 1977). This is further supported by experiments on recall of mental images (Kosslyn, Shwartz 1977) where these images appear in continuous-analog fashion. Another similar argument in favor of analog representation is the experiment of comparing objects with respect to some of their parameters, like size, or experiments on mental rotation (Shepard, Metzler 1971).

Pylyshyn (Pylyshyn 1977) cautiously argues against the analog representation for the same object viewed under different conditions as a result of the semantic interpretation function (SIF). The SIF will extract only those invariances characteristic for the object in a given situation, and thus reduce the number of possible discrete values and their range for a given parameter. The invariances are determined by laws of physics and optics, and by the context, i.e., the object sizes will remain fixed as they move, the smaller objects will partially occlude the larger object, etc.

We would like to propose a discrete value representation for parametric information with an associated interpolation function (sampling is an inverse of interpolation) and a clustering procedure. During the recognition process, a clustering procedure is evoked in order to categorize a parameter while during an image recall an interpolation procedure is applied to generate the continuous data. Our model seems not to contradict Kosslyn's findings, that is we assume as he does, that the deep representation of an image consists of stored facts about the image in a propositional format. Facts include information about:

a) How and where a part is attached to the whole objects.
b) How to find a given part.
c) A name of the category that the object belongs to.
d) The range of the size of the object, which implies the resolution necessary to see the object or part.
e) The name of the file which contains visual features that the object is composed of (corners, edges, curvature descriptions of edge segments, their relationships etc.).

The only place where we differ from Kossyln's model is in the details of the perceptual memory. While his perceptual memory contains coordinates for every point, our perceptual memory has identified and stored clusters of these points, like corners, edges, lines, etc. From these features and the interpolation procedure, we create the continuous image. This is very much in the spirit of a constructive vision theory as proposed by Kosslyn and others. A similar argument can be used for preserving continuity in transformation of images, such as rotation (Shepard, Metzler 1971) and expansion (Kosslyn 1975, 1976). The contraction process is the inverse of expansion and therefore will envoke the sampling routine instead of the interpolation routine. The problem of too many discrete values and their relationships, as stated by Fishler, is taken care of by the fact that for each parameter there is an associated range with only a few categories such as small, medium, and large. As pointed out by Pylyshyn, it is the range of parameters which is context dependent and thus differs from situation to situation. This view also offers some explanation that often incomplete figures are perceived as whole.

We also want to postulate that analog information, as we specified it, is not retained, and if there are ambiguities due to the inadequacy of the input data, a new set of data is inputed. This is supported by several psychological experiments, for example, by asking people to recognize a building where they work from accurate drawings and sloppy pictures (Norman 1975). The overwhelming evidence is that people prefer a sloppy picture to the more accurate one, for recognizing their own building. Even the experiment of Averbach and Sperling (Averbach and Sperling 1968) concerning the visual short memory after 1/20 sec exposure to letters does not contradict our hypothesis that we maintain in this case, edges rather than picture points, although it allows the other interpretation as well.

We now turn to the second question. Sihce propositional information can be represented by an equivalent procedure (giving a true or a false value), the question of propositional information vs structural information can be replaced by the question: What are the necessary procedures that have to be performed during a recognition process and what type of data they require? Clearly, the parametric information is derived procedurally There are well defined procedures for finding color, size, orientation, etc. The part-whole relationship as well as the instance relationship clearly have to be structurally represented (Miller and Johnson-Laird 1977).

While the structural information is derived from symbolic propositional data and the transformations performed are, for example, reductions, and expansions, the parametric information is derived from the perceptual data and the transformations performed are more like measurements, detections, and geometric transformations.

In the context of 3-D shape representation we believe in a combination of procedural - parametric and propositional nodes organized in a structure. Take an example of representing a shape of a human. We have the part-whole relationship: head, neck, torso, arms, legs, etc. Head has parts: eye, nose, mouth, etc. These concepts are propositional - symbolic. From the shape point of view, however, head is round, neck is short and wide elongated blob, the arms and legs are elongated and the torso is elongated but wide. Although these labels correspong to 2-D as well as 3-D shape, there is a mechanism: projection transformation which transforms elongated 3-D into elongated 2-D shape. In any case, round,

elongated, wide, short, are procedures - tests whether an object is round, elongated, etc. We know that round (circle) in 2-D corresponds to sphere in 3-D, elongated (rectangle, or ellipse) to a polyhedra or cylinder, or ellipsoid.

When we view only one view of a scene or a photograph, we analyse the 2-D outline. However, when we have more than one view at our disposal or when we are asked to make 3-D interpretation then we reach from the 2-D information to corresponding 3-D representation. This is the time when volume primitives like sphere, cylinder, and their like come into play. These primitives do not seem to be explicit (we do not say a shape of a man is a sphere attached to several cylinders) in the representation. Rather what is in the shape representation are the feature primitives, (like the symmetry planes, the ratio of symmetry axis) attached to other pointers, which point also, if appropriate, to labels like sphere, cylinder, flat object, polyhedron, etc. These labels are in turn used for shortening a complex description.

An implementation of a 3-D shape decomposition and labelling system is under development (Bajcsy, Soroka 1977). Earlier we have experimented with a partially ordered structure as means to represent 2-D shape (Tidhar 1974, Bajcsy, Tidhar 1977) in recognition of outdoor landscapes (Sloan 1977) and in the context of natural language understanding (Joshi and Rosenschein (1975), Rosenschein (75)).

Note that not always are we able to describe a shape as a composition of some volume primitives like sphere, cylinder, or a flat object. As an example in the case is a shape of a heart. A heart has 2 symmetry planes and it is roughly round, but its typical features are the two corners centered, one, concave and the other convex connected by a convex smooth surface. Here clearly, any attempt to describe this shape, by two ellipsoids or some other 'primitive' is artificial. Thus, the representation will have only feature primitives but no volume primitives.

Of course, there are cases that fall between. As an example, consider a kidney shape where one can say it is an ellipsoid with a concavity on one side.

What are the implications from all of this?

1. We do not measure or extract spheres, cylinders and their like as primitives, but rather we measure convexity, concavity, planar, corners, symmetry planes, which are primitive features.
2. These features form different structures to which are attached different but in general, not independent labels.
3. While these structures represent explicit conceptual relationships, the nodes are either labels or procedures with discrete values denoting, in general, N-ary relations.

4. Conclusions

In this paper, we have considered the following problems:

1. How much of analog information is retained during recognition process and at which level the transformation from analog to propositional takes place?
2. How much of the information stored is procedural (implicit) and structural (explicit) form?
3. What are the primitives for two dimensional and three dimensional shapes?
4. How is the labelling of shapes effected by the way the shapes are represented? By studying the shape labels can we hope to learn something about the internal representation of shapes?

Clearly, these four questions are intimately related to the general problem: representation of three dimensional objects.

We are led to the following conclusions. Our conclusions are derived primarily on the basis of our experience in constructing 2-D and 3-D recognition systems and the study of the relevant psycological and psycholinguistic literature.

1. Analog information is not retained even in a short term memory.
2. Our experience and the analysis of the relevant literature leads us to be in favor of the constructuve vision theory. The visual information is represented as structures, with nodes which are either unary or n-ary predicates. The structures denote conceptual relationships such as part-whole, class inclusion, cause-effect, etc.
3. The shape primitives are on the level of primitive features rather than primitive shapes. By primitive features we mean, corners, convex, concave and planar surfaces and their like.
4. The labels of shapes, except in a few special cases, do not describe any shape properties and are derived from objects associated with that shape.
5. In order to preserve continuity, we need interpolation procedures. We assume that several such procedures exist, for example, clustering mechanisms, sampling procedures, perspective transformations, rotation, etc. These are available as a general mechanisms for image processing.

We certainly have not offered complete solutions to all the issues discussed above, but we hope that we have raised several valid questions and suggested some approaches.

References

1. Averbach, E., and Sperling, G . Short-Term Storage of Information in Vision in: Contemporary Theory and Research in Visual Perception, (ed.) R.N. Haber, NY, Holt, Rinehart and Winston, Inc. 1968

2. Bajcsy, R., and Soroka, B.: Steps towards the Representation of Complex Three-Dimensional Objects, Proceedings on Int. Artificial Intelligence Conference, Boston, August 1977.

3. Bajcsy, R., and Tidhar, A.: Using a Structured World Model in Flexible Recognition of Two Dimensional Patterns, Pattern Recognition Vol. 9, pp. 1-10, 1977.

4. Clark, E.V.: What's in a Word? On the Child's Acquisition of Semantics in His First Language, in: Cognitive Development and the Acquisition of Language, (ed.) T.E. Moore, Academic Press, NY 1973, pp. 65-110.

5. Clark, H.L.: Space, Time Semantics, and the Child, in: Cognitive Development and the Acquisition of Language (ed.) T.E. Moore, Academic Press, NY 1973 pp. 27-63.

6. Chafe, W.L.: Creativity in Verbalization as Evidence for Analogic Knowledge, Proc. on Theoretical Issues in Natural Language Processing, Cambridge, June 1975 pp. 144-145.

7. Fishler, M.A. On the Representation of Natural Scenes, Advanced Papers for The Workshop on Computer Vision Systems, Univ. of Massachusetts, June 1977, Amherst.

8. Gibson, J.J.: The Perception of the Visual World, Boston, MA, Houghton, 1950.

9. Goldmeir, E.: Similarity in Visually Perceived Forms, Psychological Issues 8, 1972, No. 1 pp. 1-135.

10. Koffka, K . Principles of Gestalt Psychology, New York, Harcourt, Brace 1935.

11. Kosslyn, S.M.: Information Representation in Visual Images, Cognitive Psychology 7, pp. 341-370, 1975.

12. Kosslyn, S.M.. Can Imagery Be Distinguished from Other Forms of Internal Representation? Evidence from Studies of Information Retrieval Times, Memory & Cognition Vol. 4, 1976, No. 3, pp. 291-297.

13. Kosslyn, S.M., and Shwartz, S.P.: Visual Images as Spatial Representations in Active Memory, in: Machine Visions, (eds.) E.M. Riseman & A.R. Hanson, NY Academic Press (in press) 1978.

14. Miller, A , and Johnson-Laird, P.N.: Language and Perception, Harvard Univ. Press, Cambridge, MA 1976.

15. Norman, D.A., and Bobrow, D.G.: On the Role of Active Memory Processes in Perception and Cognition, in: C.N. Cofer (ed.) The Structure of Human Memory, San Francisco, W.H. Freeman, 1975.

16. Palmer, S.E.: 'The Nature of Perceptual Representation: An examination of the Analog Propositional Contraversy, Proc. on Theoretical Issues in Natural Language Processing, Cambridge, June 1975 pp. 151-159.

17. Piaget, J., and Inhelder, B : The Child's Conception of Space, New York: Humanities Press, 1956.

18. Pylshyn, Z.W.: Representation of Knowledge: Non-Linguistic Forms, Proc. on Theoretical Issues in Natural Language Processing, Cambridge, June 1975 pp. 160-163.

19. Rock, I.: Orientation and Form, Academic Press, Inc. Ny 1973.

20. Rock, I.: An Introduction to Perception, MacMillan Publ. Co., NY 1975.

21. Rosh, E.H.: On the Internal Structure of Perceptual and Semantic Categories, in: Cognitive Development and the Acquisition of Language. (ed.) T.E. Moore, Academic Press, NY 1973, pp. 111-144.

22. Shepard, R.N., and Metzler, J.: Mental Rotation of Three-Dimensional Objects, Science, 171, 1971, pp. 701-703.

23. Tidhar, A.: Using a Structured World Model in Flexible Recognition of Two Dimensional Pattern, Moore School Tech. Report No. 75-02, Univ. of Pennsylvania, Philadelphia, 1974.

24. Zusne, L.: Visual Perception of Form, Academic Press, 1970, NY and London.

25. Sloan, K.: World Model Driven Recognition of Natural Scenes, Ph.D. Dissertation, Computer Science Department, University of Pennsylvania, Philadelphia, June 1977.

26. Joshi, A.K., and Rosenschein, S.J., "A Formalism for Relating Lexical and Pragmatic Information: Its Relevance to Recognition and Generation", Proc. of TINLAP Workshop, Cambridge 1975.

27. Rosenschein, S.J., "Structuring a Pattern Space, with Applications to Lexical Information and Event Interpretation", Ph.D. Dissertation, University of Pennsylvania, Philadelphia, PA 1975.