PARTS: STRUCTURED DESCRIPTIONS OF SHAPE

Alex P. Pentland

Artificial Intelligence Center, SRI International
Menlo Park, California
and
Center for the Study of Language and Information
Stanford University

ABSTRACT [1]

A shape representation is presented that has been shown competent to accurately describe an extensive variety of natural forms (e g., people, mountains, clouds, trees), as well as man-made forms, in a succinct and natural manner. The approach taken in this representational system is to describe scene structure at a scale that is similar to our naive perceptual notion of "a part," by use of descriptions that reflect a possible formative history of the object, e.g., how the object might have been constructed from lumps of clay. For this representation to be useful it must be possible to recover such descriptions from image data; we show that the primitive elements of this representation may be recovered in an overconstrained and therefore potentially reliable manner.

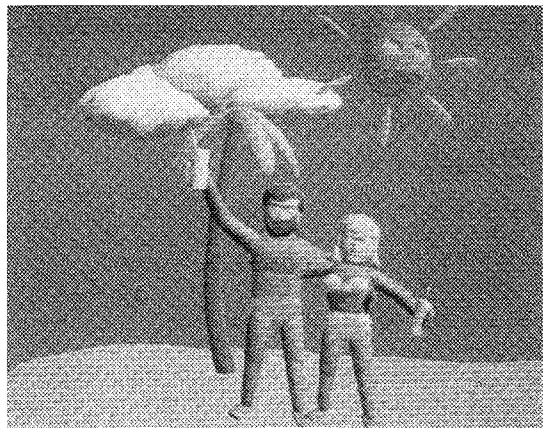

Figure 1: A scene constructed of 100 primitives, less than 1k bytes of information.

1 Introduction

Most models used in vision and reasoning tasks have been of been of only two kinds: high-level, *specific* models, e.g., of people or houses, and low-level models of, e.g., edges. The reason research has almost exclusively focused on these two types of model is a result more of historical accident than conscious decision. The well-developed fields of optics, material science and physics (especially photometry) have provided well worked out and easily adaptable models of image formation, while engineering, especially recent work in computer aided design, have provided standard ways of modeling industrial parts, airplanes and so forth.

Both the use of image formation models and specialized models has been heavily investigated. It appears to us that both types of models, although useful for many applications, encounter insuperable difficulties when applied to the problems faced by, for instance, a general purpose robot. In the next two subsections we will examine both types of models and outline their advantages and disadvantages for recovering and reasoning about important scene information. In the remainder of this section we will then motivate, develop and investigate an alternative category of models.

[1] This research was made possible by National Science Foundation, Grant No. DCR-85-19283, by Defense Advanced Research Projects Agency contract no. MDA 903-83-C-0027, and by a grant from the Systems Development Foundation. I wish to thank Marty Fischler, Ruzena Bajcsy and Andy Witkin for their comments and insights.

1.1 Models of Image Formation

Most recent research in computational vision has focused on using point-wise models borrowed from optics, material science and physics. This research has been pursued within the general framework originally suggested by Marr [1] and by Barrow and Tenenbaum [2], in which vision proceeds through a succession of levels of representation. Processing is primarily data-driven (bottom-up), i.e., the initial level is computed directly from local image features, and higher levels are then computed from the information contained in small regions of the preceding levels.

Problems for vision. Despite its prevalence, there are serious problems that seem to be inherent to this research paradigm. Because scene structure is underdetermined by the local image data [3], researchers have been forced to make *unverifiable* assumptions about large-scale structure (e.g., smoothness, isotropy) in order to derive useful information from their local analyses of the image. In the real world, unfortunately, such assumptions are often seriously in error: in natural scenes the image formation parameters change in fairly arbitrary ways from point to point, making any assumption about local context quite doubtful. As a result, those techniques that rely on strong assumptions such as isotropy or smoothness have proved fragile and error-prone; they are simply not useful for many natural scenes.

That such difficulties have been encountered should not, perhaps, be too surprising. It is easily demonstrated (by looking through a viewing or reduction tube) that people can obtain little information about the world from a local image patch taken out of its context. It is also clear that detailed, analytic models of the image formation process are not essential to human

perception; humans function quite well with range finder images (where brightness is proportional to distance rather than a function of surface orientation), electron microscope images (which are approximately the reverse of normal images), and distorted and noisy images of all kinds — not to mention drawings [4].

Problems for reasoning. Perhaps even more fundamentally, however, even if depth maps and other maps of intrinsic surface properties could be reliably and densely computed, how useful would they be in reasoning tasks? Industrial vision work using laser range data has demonstrated that the depth maps, reflectance maps and the other maps of the 2-1/2D sketch are still basically just images. Although useful for obstacle avoidance and other very simple tasks, they still must be segmented, interpreted and so forth before it can be used for any more sophisticated task [5].

1.2 Specialized Models

The alternative to models of image formation has been engineering-style representations; e.g., CAD-CAM models of specific objects that are to be identified and located. Such detailed, specific models have provided virtually all of the success stories in machine vision; nonetheless, such models have important inherent limitations.

Problems for vision. As the object's orientation varies these models produce a *very* large number of different pixel configurations. The large number of possible appearances for such models makes the problem of recognizing them very difficult — unless an extremely simplified representation is employed. The most common type of simplified representation is that of a wireframe model whose components correspond to the imaged edges. The use of an impoverished representation, however, generally means that the flexibility, reliability and discriminablity of the recognition process is limited. Thus research efforts employing specific object models have floundered whenever the number of objects to be recognized becomes large, when the objects may be largely obscured, or when there are many unknown objects also present in the scene.

Problems for reasoning. An even more substantive limitation of systems that employ *only* high-level, specific models is that there is no way to learn new objects: new models must be specially entered, usually by hand, into the database of known models. This is a significant limitation, because the ability to encounter a new object, enter it into a catalog of known objects, and thereafter recognize it is an absolute requirement of truly general purpose robot.

1.3 Part and Process Models

In response to these difficult problems some researchers have begun to search for a third type of model, one with a grain size intermediate between the point-wise models of image formation and the complex, specific models of particular objects [6,7].

Recent research in graphics, biology, and physics has provided us with good reason to believe that it may be possible to accurately describe our world by means of a few, commonly-occuring types of formative processes [1,8,9,10]; i.e., that our world can be modeled as a relatively small set of generic processes — for instance, bending, twisting, or interpenetration — that occur again and again, with the apparent complexity of our environment being produced from this limited vocabulary by compounding these basic forms in myriad different combinations.

Moreover, some modern psychologists [18,19,20], as well as the psychologists of the classic Gestalt movement, have argued that the initial stages of human perception function primarily to discover image features that indicate the presence of these generic categories of shape structure. They have presented strong evidence showing that we conceive of the world in terms of *parts*, and that the first stages of human perception are primarily concerned with detecting features that indicate the structure of those parts. This part-structure, then, seems to form the building blocks upon which we build the rest of our perceptual interpretation.

Such part-and-process models offer considerable potential for reasoning tasks, because they describe the world in something like "natural kind" terms: they speak qualitatively of whole forms and of relations between parts of objects, rather than of local surface patches or of particular instances of objects. It seems, for instance, that we employ such intermediate-grain descriptions in commonsense reasoning, learning, and analogical reasoning [13,14,15].

The problem with forming such "parts" models is that they must be complex enough to be reliably recognizable, and yet simple enough to reasonably serve as building blocks for specific object models. Current 3-D machine vision systems, for instance, typically use "parts" consisting of rectangular solids and cylinders. Unfortunately, such a representation is only capable of an extremely abstracted description of most natural and biological forms. It cannot accurately and succinctly describe most natural animate forms or produce a succinct description of complex inanimate forms such as clouds or mountains. If we retreat from cylinders to generalized cylinders we can, of course, describe such shapes accurately. The cost of such retreat is that we must introduce several 1-D functions describing the axis and cross-section shape; this makes the representation neither succinct nor intuitively attractive.

2 A Representation

The idea behind this representational system is to provide a vocabulary of models and operations that will allow us to model our world as the relatively simple composition of component "parts," parts that are reliably recognizable from image data.

The most primitive notion in this represention is analogous to a "lump of clay," a modeling primitive that may be deformed and shaped, but which is intended to correspond roughly to our naive perceptual notion of "a part." For this basic modeling element we use a parameterized family of shapes known as a *superquadrics* [10,11], which are described (adopting the notation $\cos\eta = C_\eta$, $\sin\omega = S_\omega$) by the following equation:

$$\vec{X}(\eta,\omega) = \begin{pmatrix} C_\eta^{\epsilon_1} C_\omega^{\epsilon_2} \\ C_\eta^{\epsilon_1} S_\omega^{\epsilon_2} \\ S_\eta^{\epsilon_1} \end{pmatrix}$$

where $\chi(\eta,\omega)$ is a three-dimensional vector that sweeps out a surface parameterized in latitude η and longitude ω, with the surface's shape controlled by the parameters e_1 and e_2. This family of functions includes cubes, cylinders, spheres, diamonds

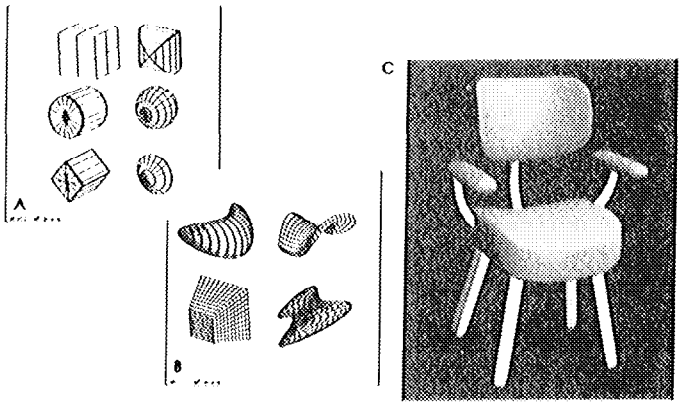

Figure 2: (a) A sampling of the basic forms allowed, (b) deformations of these forms, (c) a chair formed from Boolean combinations of appropriately deformed superquadrics.

and pyramidal shapes as well as the round-edged shapes intermediate between these standard shapes. Some of these shapes are illustrated in Figure 2(a). Superquadrics are, therefore, a superset of the modeling primitives currently in common use.

These basic "lumps of clay" (with various symmetries and profiles) are used as prototypes that are then deformed by stretching, bending, twisting or tapering, and then combined using Boolean operations to form new, complex prototypes that may, recursively, again be subjected to deformation and Boolean combination [12]. As an example, the back of a chair is a rounded-edge cube that has been flattened along one axis, and then bent somewhat to accommodate the rounded human form. The bottom of the chair is a similar object, but rotated 90°, and by "or-ing" these two parts together with elongated rectangular primitives describing the chair legs we obtain a complete description of the chair, as illustrated in Figure 2(c). We have found that this representational system has a surprizingly powerful generative power that allows the creation of a tremendous variety of form, such as is illustrated by Figure 1.

This descriptive language is designed to describe shapes in a manner that corresponds to a possible formative history, e.g., how one would create a given shape by combining lumps of clay. Thus the description provides us with an explanation of the image data in terms of the interaction of generic formative processes. This primitive explanation can then be refined by application of specific world knowledge and context, eventually deriving causal connections, affordances, and all of the other information that makes our perceptual experience appear so rich and varied.

For instance, if we have parsed the chair in Figure 2(c) into its constituent parts we could deduce that the bottom of the chair is a stable platform and thus might be useful as a seat, or we might hypothesize that the back of the chair can rigidly move relative to the supporting rod, given the evidence that they are separate "parts" and thus likely separately formed. We believe that this process-oriented, possible-history form of representation will prove to be extremely useful for commonsense reasoning tasks.

2.1 Building 3-D models

This type of representation seems to produce models that represent the shape "naturally." We have, for instance, performed a protocol analysis in which we found [14] that when adult human subjects are required to verbally describe imagery with completely novel content, their typical spontaneous strategy is to employ a descriptive system analogous to this one — i.e., form is described by modifying and combining prototypes. Moveover, the non-proper-noun terms used were limited and stereotyped: they resorted largely to terms indicating interpenetration (boolean combination), squareness-roundness, bending, tapering, and stretching.

We have also investigated the psychological reality of this descriptive framework using the psychophysical techniques developed by Treisman [17]. Using this experimental paradigm and employing monocular imagery depicting shaded, perspective views of three-dimensional forms, we have collected experimental evidence indicating [21] that convexity-concavity (equivalent to boolean combination), squareness-roundness, bending, tapering and relative axis size (streching) may all be preattentively perceived, that is, there appear to be parallel "detectors" that search for the presence (but not absence) of these features within a 3-D scene.

We have also attempted to verify this psychological evidence in a more practical manner. The fact that "natural" man-machine interaction requires that the machine uses a representation that closely matches that of the human operator provides a practical test for our descriptive framework. That is, if an interface based on this representation appears "natural" to users, then we can conclude that the representation must closely match at least one way that people think about 3-D shapes.

We have, therefore, constructed a 3-D modeling system called "SuperSketch," that employs the shape representation described here. This real-time, interactive [2] modeling system is implemented on the Symbolics 3600, and allows users to interactively create "lumps," change their squareness/roundness, stretch, bend, and taper them, and finally to combine them using Boolean operations. This system was used to make the images in this paper.

We have found that interaction is surprisingly effortless: it took less than a half-hour to assemble the faces in Figure 1, and about four hours total to make the complete Figure 1. This is in rather stark contrast to more traditional 3-D modeling systems. It thus appears that the primitives, operations and combining rules used by the computer closely match the way that the human operators think about 3-D shape.

2.2 Biological forms

In Figure 1 (as in all cases examined to date) when we try to model a particular 3-D form we find that we are able to describe — indeed, it is quite natural to describe — the shape in a manner that corresponds to the organization our perceptual apparatus imposes upon the image. That is, the components of the

[2]Because these forms have an underlying analytical form, we can use fast, qualitative approximations to accomplish hidden surface removal, intersection and image intensity calculations in "real time," e g , a "lump" can be moved, hidden surface removal accomplished, and drawn as a 200 polygon line drawing approximation in 1/8th of a second.

description match one-to-one with our naive perceptual notion of the "parts" in the figure, e g., the face in Figure 1 is composed of primitives that correspond exactly to the cheeks, chin, nose, forehead, ears, and so forth.

This correspondence indicates that we are on the right track; e g , that this representation will be useful in understanding commonsense reasoning tasks. Similarly, the ability to make the right "part" distinctions offers hope that we can form qualitative descriptions of specific objects ("Ted's face") or of classes of objects ("a long, thin face") by specifying constraints on part parameters and on relations between parts, in the manner of Winston [15].

Finally, we note that the extreme brevity of these descriptions makes many otherwise difficult reasoning tasks relatively simple, e g , even NP-complete problems can be easily solved when the size of the problem is small enough. The human bodies shown in Figure 1, for instance, require combining only 45 primitives, or approximately 450 bytes of information (these informational requirements are not a function of body position). Similarly, the description for the face requires the combination of only 18 primitives, or fewer than 200 bytes of information.

2.3 Complex inanimate forms

This method for representing the three-dimensional world, although excellent for biological and man-made forms, becomes awkward when applied to complex natural surfaces such as mountains or clouds The most pronounced difficulty is that, like previously proposed representations, our superquadric lumps-of-clay representation becomes implausably complex when confronted with the problem of representing, e.g., a mountain, a crumpled newspaper, a bush or a field of grass.

Why do such introspectively simple shapes turn out to be so hard to represent? Intuitively, the main source of difficulty is that there is too much information to deal with. Such objects are amazingly bumpy and detailed; there is simply too much detail, and it is too variable.

People escape this overwhelming complexity by varying the level of descriptive abstraction — the amount of detail captured — depending on the task. In cases like the crumpled newspaper, or when recognizing classes of objects such as "a mountain" or "a cloud," the level of abstraction is very high. Almost no specific detail is required, only that the crumpledness of the form comply with the general physical properties characteristic of that type of object. In recognizing a *specific* mountain, however, people will require that all of the major features be identical, although they typically ignore smaller details. Even though these details are "ignored," however, they must still conform to the constraints characteristic of that type of object: we would never mistake a smooth cone for a rough-surfaced mountain even if it had a generally conical shape.

Our previous work with fractal models of natural surfaces [16] allows us to duplicate this sort of physically-meaningful abstraction from the morass of details encountered in natural scenes. It lets us describe a crumpled newspaper by specifying certain structural regularities — its crumpledness, in effect — and leave the rest as variable detail. It lets us specify the qualitative shape — i.e , the surface's roughness — without (necessarily) worrying about the details.

We may construct fractal surfaces by using our superquadric

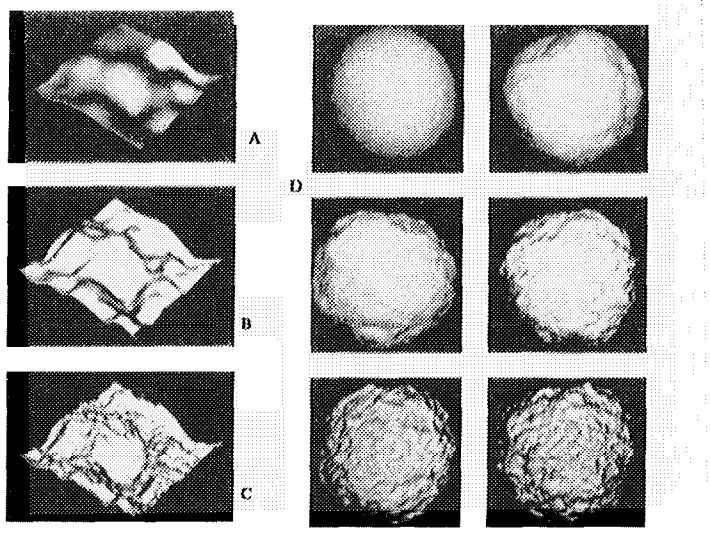

Figure 3: (a) - (c) show the construction of a fractal shape by successive addition of smaller and smaller features with number of features and amplitudes described by the ratio $1/r$, (d) shows spherical shapes with surface crenulations ranging from smooth ($r \approx 0$) to rough ($r \approx 1$).

"lumps" to describe the surface's features; specifically, we can use the recursive sum of smaller and smaller superquadric lumps to form a true fractal surface. This construction is illustrated in Figures 3(a) - (c).

We start by specifying the surface's qualitative appearance — it's roughness — by picking a ratio r, $0 \leq r \leq 1$, between the number of features of one size to the number of features that are twice as large. This ratio describes how the surface varies across different scales (resolutions, spatial frequency channels, etc.) and is related to the surface's fractal dimension D by $D = T+r$, where T is the topological dimension of the surface.

We then randomly place n^2 large bumps on a plane, giving the bumps a Gaussian distribution of altitude (with variance σ^2), as seen in Figure 3(a). We then add to that $4n^2$ bumps of half the size, and altitude variance $\sigma^2 r^2$, as shown in Figure 3(b). We continue with $16n^2$ bumps of one quarter the size, and altitude $\sigma^2 r^4$, then $64n^2$ bumps one eighth size, and altitude $\sigma^2 r^6$ and so forth, as shown in Figure 3(c). The final result, shown in Figure 3(c) is a true Brownian fractal shape. Different shaped lumps will produce different textures on the resulting fractal surface.

When the larger components of this sum are matched to a particular object we obtain a description of that object that is exact to the level of detail encompassed by the specified components. This makes it possible to specify a global shape while retaining a qualitative, statistical description at smaller scales: to describe a complex natural form such as a cloud or mountain, we specify the "lumps" down to the desired level of detail by fixing the larger elements of this sum, and then we specify only the fractal statistics of the smaller lumps thus fixing the qualitative appearance of the surface. Figure 3(d) illustrates an example of such description. The overall shape is that of a sphere; to this specified large-scale shape, smaller lumps were added randomly. The smaller lumps were added with six different choices of r (i.e., six different choices of fractal statistics) resulting in

six qualitatively different surfaces — each with the same basic spherical shape. The ability to fix particular "lumps" within a given shape provides an elegant way to pass from a qualitative model of a surface to a quantitative one — or *vice versa*.

3 Recognizing Our Modeling Primitives

The major difficulty in recovering such descriptions is that image data is mostly a function of surface normals, and not directly a function of the surface shape. This is because image intensity, texture anisotropy, contour shape, and the like — the information we have about surface shape — is largely determined by the direction of the surface normal. To recover the shape of a general volumetric primitive, therefore, we must (typically) first compute a dense depth map from information about the surface normals. The computation of such a depth map has been the major focus of effort in vision research over the last decade and, although the final results are not in, the betting is that such depth maps are impossible to obtain in the general, unconstrained situation. Even given such a depth map, the recovery of a shape description has proven extremely difficult, because the parameterization of the surface given in the depth map is generally unrelated to that of the desired description.

Because image information is largely a function of the surface normal, one of the most important properties of superquadrics is the simple "dual" relation between their surface normal and their surface shape. It appears that this dual relationship can allow us to form an overconstrained estimate of the 3-D parameters of such a shape from noisy or partial image data, as outlined by the following equations.

The surface position vector of a superquadric with length, width and breadth a_1, a_2 and a_3 is (again writing $\cos\eta = C_\eta$, $\sin\omega = S_\omega$)

$$\vec{X}(\eta,\omega) = \begin{pmatrix} a_1 C_\eta^{\epsilon_1} C_\omega^{\epsilon_2} \\ a_2 C_\eta^{\epsilon_1} S_\omega^{\epsilon_2} \\ a_3 S_\eta^{\epsilon_1} \end{pmatrix} \tag{1}$$

and the surface normal at that point is

$$\vec{N}(\eta,\omega) = \begin{pmatrix} \frac{1}{x} C_\eta^2 C_\omega^2 \\ \frac{1}{y} C_\eta^2 S_\omega^2 \\ \frac{1}{z} S_\eta^2 \end{pmatrix} \tag{2}$$

Therefore the surface vector $\vec{X} = (x, y, z)$ is dual to the surface normal vector $\vec{N} = (x_n, y_n, z_n)$. From (2), then we have

$$\left(\frac{y y_n}{x x_n}\right)^{1/2} = \tan\omega \tag{3}$$

We may also derive an alternative expression for $\tan\omega$ from (1):

$$\left(\frac{y a_1}{x a_2}\right)^{1/\epsilon_2} = \tan\omega \tag{4}$$

Combining these expressions for $\tan\omega$ and letting $r = y_n/x_n$, $k = (a_1/a_2)^{2/\epsilon_2}$ and $\xi = 2/\epsilon_2 - 1$ we find that

$$r = k\left(\frac{y}{x}\right)^{\xi} \tag{5}$$

$$\frac{dr}{dy} = \frac{k\xi}{x}\left(\frac{y}{x}\right)^{\xi-1} \qquad \frac{dr}{dx} = \frac{-k\xi y}{x^2}\left(\frac{y}{x}\right)^{\xi-1} \tag{6}$$

This gives us two equations relating the unknown shape parameters to image measurable quantities, i.e.,

$$\frac{r}{\frac{dr}{dy}} = \frac{y}{\xi} \qquad \frac{r}{\frac{dr}{dx}} = \frac{-x}{\xi} \tag{7}$$

Thus Equations (7) allow us to construct a linear regression to solve for center and orientation of the form, as well as the shape parameter e_2, given only that we can estimate the surface tilt direction r.

When we generalize these equations to include unknown orientation and position parameters for the superquadric shape, we obtain a new set of nonlinear equations that can then be solved (in closed form) for the unknown shape parameters e_1 and e_2, the center position, and the three angles giving the objects orientation. Once these unknowns are obtained the remaining unknowns (a_1, a_2, and a_3, the three dimensions of the object) may be directly obtained.

3.1 Overconstraint and reliability

Perhaps the most important aspect of these equations is that we can form an *overconstrained* estimate of the 3-D parameters: thus we can *check* that our model applies to the situation at hand, and we can *check* that the parameters we estimate are correct. This property of overconstraint comes from using models: when we have used some points on a surface to estimate 3-D parameters, we can check if we are correct by examining additional points. The model predicts what these new points should look like; if they match the predictions then we can be sure that the model applies and that the parameters are correctly estimated. If the predictions do *not* match the new data points, then we know that something is wrong. The ability to check your answer is perhaps the most important property any vision system can have, because only when you can check your answers can you build a reliable vision system. And it is *only* when you have a model that relates many different image points (such as a model of how rigid motion appears in an image sequence, or a CAD-CAM model, or this 3-D shape model) that you can have the overconstraint needed to check your answer.

Another aspect of Equations (7) that deserves special note is that the only image measurement needed to recover 3-D shape is the surface tilt r, the component of shape that is unaffected by projection and, thus, is the most reliably estimated parameter of surface shape It is, for instance, known exactly at smooth occluding contours and both shape-from-shading and shape-from-texture methods produce a more reliable estimate of r than of slant, the other surface shape parameter. That we need only the (relatively) easily estimated tilt to estimate the 3-D shape parameters makes robust recovery of 3-D shape much more likely.

One final note about Equations (7) is that they become singular when superquadric becomes rectangular; i.e., when the sides of the superquadric have zero curvature. This, however, is the case of the blocks world. We may view this work with superquadric shapes, therefore, as a natural extension of the blocks world to a domain that also encompasses smoothly curved shapes

3.2 Recovering Part Descriptions

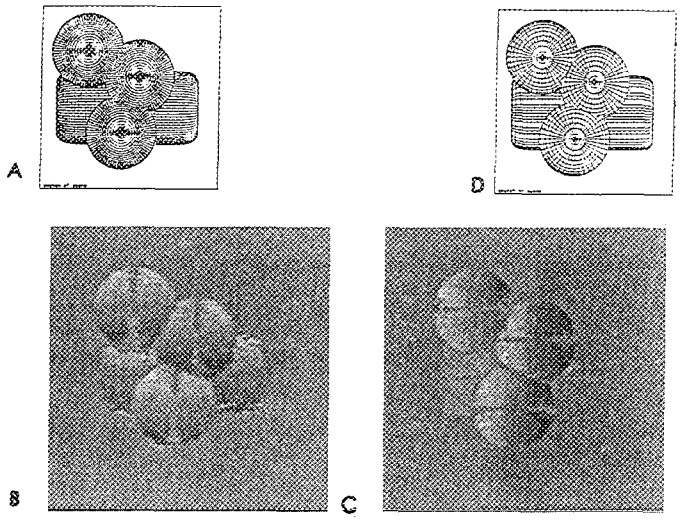

Figure 4: Recovering the part structure of a scene: (a) the original scene, (b) ratio of τ to $d\tau/dy$, (c) ratio of τ to $d\tau/dx$, (d) recovered scene description.

Figure 4 illustrates how we may use Equation (7) for image segmentation and shape recovery. In these examples we will not consider rotation in depth; the extension to three degrees of freedom is straightforeward, although from a numerical view considerably more complex

Figure 4 shows an actual example of recovering a part description from depth information. We started with the complex scene shown in Figure 5(a), and generated a depth array with approximately eight bits of accuracy. We then computed the gradient direction (the tilt τ) over the entire depth array (with about seven bits accuracy), and finally computed the x and y derivatives of the tilt array.

From this we calculated the ratios of τ to $d\tau/dy$ (shown in Figure 4(b)) and τ to $d\tau/dx$ (shown in Figure 4(c)). Equation (7) predicts that within each superquadric form: (1) that the value of these ratios should be a linear function of the image y (x) coordinate, (2) that the zero-crossing of this ratio should lie along the x (y) axis of the imaged form, and that (3) the slope of this ratio as a function of y (x) should be proportional to the squareness-roundness of the form along that axis. It can be seen that these relations are in fact obtained, except for a vertical bar caused by the tilt fields' singular transition point.

We may use the image regularity shown in Figures 4(b) and (c) to segment the image: as each imaged superquadric produces a linearly sloping region with a particular orientation and axis direction, we need only segment the ratio of (1) τ to $d\tau/dy$, and (2) τ to $d\tau/dx$ into linearly varying domains in order to completely segment the image into its component parts. We can even use this regularity to "match up" the various portions of a partially occluded object.

It can be seen, for instance, that there are two disjoint areas of the block-like shape. How can we infer that these two visible portions in fact belong to a single whole? From Figures 4(b) and (c) we can observe that the x axes of both visible portions are collinear, and that they have the same slope and zero-crossing when considered as a function of both x and y. This, then, gives us enough information to combine these two separate segments into a single 3-D part: for it is extremely unlikely that two surface segments would be collinear, of the same size and shape, and even share the same centroid, without both being portions of the same object.

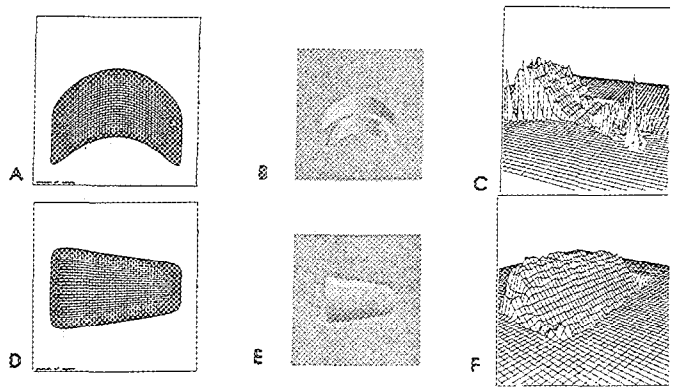

Figure 5: Examples of recovering bent and tapered part descriptions.

Finally, after segmenting the figure into linearly varying domains, the position of their x and y axes and the shape along these axes can be calculated using Equation (7), and then the extent along each axis determined. The resulting recovered description is shown in Figure 4(d); the most pronounced error is that the squareness of the cylindrical shape was somewhat overestimated.

There are two important things to remember about this demonstration: One is that we are recovering a *large-grain, part-by-part* description, rather than simply a surface description. That means that we can predict how the form will look from other views, and reason about the functional aspects of the complete shape. The second is that the estimation process is *over-constrained*, and thus it can be made *reliable*.

3.2.1 Recovering deformed primitives

So far we have not talked about recovering deformed part primitives; deformations, of course, are an important part of our shape representation theory. Figure 5, therefore, shows how we may apply these same ideas to the problem of recovering deformed part primitives.

Figure 5(a) shows a bent cylinder; Figure 5(b) and (c) show the ratio of τ to $d\tau/dx$. It can be seen that the linear relation still holds over most of the form; thus allowing segmentation. Perhaps even more importantly, however, is that the axis of the figure is clearly defined in Figures 5(b) and (c). It appears, therefore, that we may be able to locate the axis of the figure, and then estimate the amount of bending that has occured. Once we know the deformation, we can then undeform the shape, and proceed as before.

Figure 5(d) shows the case of a tapered cylinder. Figures 5(e) and (f) show that a linear ratio of τ to $d\tau/dy$ is still obtained, allowing not only segmentation but also estimation of the shape along the y direction. The amount of tapering can be determined from the tapering extent of the linearly varying region.

4 Summary

We have described a shape representation that is able to accurately describe an wide variety of natural forms (e.g., people, mountains, clouds, trees), as well as man-made forms, in a succinct and natural manner. The approach taken in this representational system is to describe scene structure at a scale that is more like our naive perceptual notion of "a part" than the point-wise descriptions typical of current image understanding research.

We have been able to use this representation to make several interesting points, in particular:

- We have demonstrated that this formative-history-oriented representational system is able to accurately describe a wide range of natural and man-made forms in an extremely simple, and therefore useful, manner.
- We have shown that this approach to perception formulates the problem of recovering shape descriptions as an *overconstrained* problem, thus potentially allowing reliable shape recovery while still providing the flexibility to learn new object descriptions.
- We have collected experimental evidence about the constituent elements of this representation, and have found that (1) evidence from the Triesman paradigm indicates that they are features detected during the early, preattentive stage of human vision, and (2) that evidence from protocol analysis indicates that people standardly make use of these same descriptive elements in generating verbal descriptions, given that there is no similar named object available.
- And finally, we have presented evidence from our 3-D modeling work showing that descriptions framed in the representation give us the right "control knobs" for discussing and manipulating 3-D forms in a graphics environment.

The representational framework presented here is *not* complete. It seems clear that additional process-oriented modeling primitives, such as branching structures or particle systems [22], will be required to accurately represent objects such as trees, hair, fire, or river rapids. Further, it seems clear that domain experts form descriptions differently than naive observers, reflecting their deeper understanding of the domain-specific formative processes and their more specific, limited purposes. Thus, accounting for expert descriptions will require additional, more specialized models. Nonetheless, we believe this descriptive system makes an important contribution to current research by allowing us to describe a wide variety of forms in a surprisingly succinct and natural manner, using a descriptive vocabulary that offers hope for the reliable recovery of shape.

REFERENCES

[1] Marr, D. (1982) *Vision*, San Fransico: W.H. Freeman and Co.

[2] Barrow, H. G., and Tenenbaum, J. M., (1978) Recovering intrinsic scene characteristics from images. In *Computer Vision Systems*, Hanson, A. and Riseman, E. (Ed.) New York: Academic Press.

[3] Pentland, A. Local analysis of the image, (1984) *IEEE Transactions on Pattern Analysis and Machine Recognition, 6* 2, 170-187

[4] Witkin, A. P., and Tenenbaum, J. M., (1985) On perceptual organization. In *From Pixels to Predicates*, Pentland, A. (Ed.) Norwood N.J.: Ablex Publishing Co.

[5] Bolles, B. and Haroud, R., (1985) 3DPO: An inspection system. In *From Pixels to Predicates*, Pentland, A. (Ed.) Norwood N J : Ablex Publishing Co.

[6] Nevatia, R., and Binford, T.O.,(1977) Description and recognition of curved objects. *Artificial Intelligence, 8*, 1, 77-98

[7] A Pentland and A. Witkin, (1984) "On Perceptual Organization," Second Conference on Perceptual Organization, Pajaro Dunes, CA, June 12-15.

[8] Thompson, D'Arcy, (1942) *On Growth and Form*,, 2d Ed., Cambridge, England: The University Press.

[9] Stevens, Peter S., (1974) *Patterns In Nature*, Boston: Atlantic-Little, Brown Books.

[10] Gardiner, M. (1965) The superellipse: a curve that lies between the ellipse and the rectangle, *Scientific American*, September 1965.

[11] Barr, A., (1981) Superquadrics and angle-preserving transformations, *IEEE Computer Graphics and Application, 1* 1-20

[12] Barr, A , (1984) Global and local deformations of solid primitives. *Computer Graphics 18*, 3, 21-30

[13] Hayes, P. (1985) The second naive physics manifesto, In *Formal Theories of the Commonsense World*, Hobbes, J. and Moore, R. (Ed.), Norwood, N.J.: Ablex

[14] Hobbs, J. (1985) Final Report on Commonsense Summer. SRI Artificial Intelligence Center Technical Note 370.

[15] Winston, P., Binford, T., Katz, B., and Lowry, M. (1983) *Proceedings of the National Conference on Artificial Intelligence (AAAI-83)*, pp. 433-439, Washington, D.C., August 22-26.

[16] Pentland, A. (1984a), Fractal-based description of natural scenes, *IEEE Pattern Analysis and Machine Intelligence, 6*, 6, 661-674.

[17] Treisman, A., (1985) Preattentive processing in vision, *Computer Vision, Graphics and Image Processing*, Vol 31, No. 2, pp. 156-177.

[18] Hoffman, D., and Richards, W., (1985) Parts of recognition, In *From Pixels to Predicates*, Pentland, A. (Ed.) Norwood, N.J.: Ablex Publishing Co.

[19] Leyton, M. (1984) Perceptual organization as nested control. *Biological Cybernetics 51*, pp. 141-153.

[20] Beiderman, I., (1985) Human image understanding: recent research and a theory, *Computer Vision, Graphics and Image Processing*, Vol 32, No. 1, pp. 29-73.

[21] Pentland, A. (1986) On perceiving 3-D shape and texture, *Computational models in human vision*, Center for Visual Science, University of Rochester, Rochester. N.Y., June 19-21

[22] Reeves, W. T., (1983) Particle systems - a technique for modeling a class of fuzzy objects, *ACM Transactions on Graphics 2*, 2, 91-108.