THREE DIMENSIONAL OBJECT REPRESENTATION REVISITED

Ruzena Bajcsy and Franc Solina

GRASP Laboratory
Department of Computer and Information Sciences
University of Pennsylvania
Philadelphia, Pennsylvania 19104-6389, USA

Abstract

Categories and shape prototypes are considered for a class of object recognition problems where rigid and detailed object models are not available or do not apply. We propose a modeling system for generic objects to recognize different objects from the same category with only one generic model. We base our design of the modeling system upon the current psychological theories of categorization and human visual perception. The representation consists of a prototype represented by parts and their configuration. Parts are modeled by superquadric volumetric primitives which can be combined via Boolean operations to form objects. Variations between objects within a category are described by changes in structure and shape deformations of prototypical parts. Recovery of deformed superquadric models from sparse 3-D points is developed and some results are shown.

Introduction

Ten years ago one of us[1] argued for common grounds between computer vision and graphics on the issue of representation of three-dimensional objects. Since then some progress has been made and several different representations have been investigated.[2,3] But, in our opinion, the following questions are still open:

1. What is the role of categorization in computer vision?
2. What are the features and primitives necessary for categorization?
3. How to represent categories?

The rest of the introduction is our attempt to answer these questions.

Categories and prototypical shapes

To answer the first question, we need some background on how people form categories. We could not perceive and recognize if we would not pick out the *essential* and separate it from the *accidental*.[4] This sorting of instances is called *categorization*. Hence categories are not arbitrary collections of objects, but reflect the structure of the world. There are two rival theories of categorization in psychology. The first, or the classical theory, claims that category membership is determined by the satisfaction of a set of properties or features (see i.e. Harnad[5]). Most work in pattern recognition work was done in this framework. According to Rosch,[6] however, categories contain one member that seems most representative. These members are called *prototypes*. Properties or features characterizing a category are held in full extent only by the prototypes. Other members of the category are perceived in terms of, or as deviating from, the prototype. This reference is asymmetrical since the reverse does not occur; for example, line orientations are referenced to the vertical and horizontal but a vertical is not referenced to a diagonal (see also Goldmeier[7]). A function, called cue-validity, evaluates for each category member, the extent to which the member is similar to other members and dissimilar from non-category members. Leyton[8] tried to reconcile the two rival theories of categorization by showing that they are two substructures of a single overall structure. Experiments supporting the prototype theory have shown that prototypes are more rapidly recognized in comparison to other objects in the same category.[6]

Categories most closely linked to the structure of the perceived world are called *basic categories*. Rosch[6] has shown that basic categories have the highest level of abstraction among categories for which a generalized outline form can be recognized and an image generated. Basic categories are also the preferred level of reference; children learn them first, and they are recognized faster then subordinate and superordinate categories.[6] (Example: chair vs. arm chair and furniture). Superordinate categories seem to share primarily functional features - i.e. vehicles are for transportation, tools are for fixing. This is in sharp contrast with basic categories which share both *functional and perceptual* features.[6] However, superordinate categories depend on the prior existence of basic categories on whose description they are based. Harnad[5] calls such dependency "grounding of categories." Because the decisive features of superordinate categories are functional and not iconic, it would be easier to recognize them in action or in use rather then in static scenes. Subordinate categories, on the other hand, subdivide basic

---

Thanks to Sandy Pentland, Michael Leyton, Stevan Harnad, Thomas Henderson and Rachel Gelman for useful comments and to Max Mintz for helping us with the fitting optimization. This work was supported in part by the following contracts and grants: USPS contract 104190-87-H-0001/M-0195, ONR Subcontract SB35919-0, NSF/DCR-8410771, ARMY/DAAG-29-84-K-0061, NSF-CER/DCR82-19196 A02, Airforce/F49616-85-K-0014, and DARPA/ONR.

categories according to one or very few perceptual or functional features (i.e. shape, texture, color). Also basic categories seem to be mutually exclusive which is not the case with either subordinate or superordinate categories. Witgenstein[9] noted that categorical judgements become a problem only if one is concerned with boundaries. Categories are formed by induction on the basis of a finite sample. Because of this approximative rather than "exact" nature there maybe a marginal overlap of basic categories, like cups and mugs. However, both appearance and function are similar for them. Categories converge on these approximations by accumulating more data.

Harnad,[5] on the other hand, claims that categorization is very context dependent and that there is no privileged level of categorization. The basic level is hence just a default-context effect. Following Harnad, consider a simple example on how to distinguish trees and animals. Suppose that all the necessary preprocessing is done, giving smoothed, segmented shapes in standard position and orientation. Counting the number of legs and calling something with one leg a tree and instances with more legs an animal would suffice in most cases. By elaborating the rule, better and better approximations could be achieved but no "essence" of a tree or of an animal can be captured as noted already by Wittgenstein.[9]

In order for computers to achieve the robotic capacity of interaction with objects in the world (manipulating, recognizing, describing, and naming of objects), computer vision systems will have to use category like models. These models should support the above mentioned capacities on the basic category level as the default level of the perceived world structure. Since basic categories are nonoverlapping there would be no redundancy in a model data base. Models for subordinate and superordinate categories could be built on top of these basic category models. At the same time, these models should also support the use of specific features for faster recognition in well-known contexts.

Features and primitives for categorization

All natural sciences strive to find some kind of primitives, elementary particles or building blocks. Computer vision is, in this regard, no exception. One would wish to identify very general primitives that are context, view and scale independent.

There are currently three different classes of tasks studied in computer vision which demonstrate the need for different features: path planning and obstacle avoidance, grasping and manipulation of objects, and object recognition. Obstacle avoidance and path planning require only a representation of space occupancy or the lack of it. Space occupancy can be described by simple uniform building blocks (voxels, elementary cubes, boxes or spheres) or by a 3-D convex hull. Grasping tasks require somewhat more detailed shape information, such as the overall size and shape of those surface patches where a gripper can be positioned for executing the grasp. To this class of tasks belongs also the classification of objects according to some gross shape characteristics, as mail pieces sorting and handling.[10] For object recognition two extreme examples can be considered. On one side of the spectrum are tasks like machine inspection where the models are precisely defined by a CAD system with explicit information about part-whole relationship, surface characteristics, and tolerances.[11] Another, much harder type of recognition task, is to classify biological species. Although all the criteria and necessary features for the classification are precisely defined, due to the intraspecies variations fixed shape models cannot be used. While constraints of rigidity hold for the first example[3] the models for the second task must incorporate some flexibility of shape.

As much as these tasks look different, we believe that they can be all regarded as a type of *categorization*. At one end of the spectrum, we are concerned with very general, superordinate categories, like obstacles, and on the other, we want to classify very specific, subordinate categories, like biological species. These tasks, however, have in common the ordering of features required for categorization. Recent psychological experiments by Novick and Tversky[12] have shown that in constructing solutions to geometric analogy tasks we perform operations in a specific *order:* involving first location, orientation, size, general outline, and addition of small parts and other details at the end. This order parallels the order in which the corresponding information is needed in planning and executing a drawing or during object identification. Leyton,[13] has shown that such layered asymmetric structure underlies not only planning but also perceptual organization, structuring of categories, and grammar. The above mentioned categorization tasks then differ by how far they follow up this sequence of features. For path planning just the locations of obstacles suffice, while for identification of biological species, very fine details must be recognized.

In the context dependent categorization scheme, Harnad[5] believes that humans form *three* kinds of representation when encountering a sensory input. First, an iconic representation is formed which is something like a smoothed, filtered image - an analog to the sensory input. From this first representation, a so called category representation is derived, which is bounded and context dependent in the sense that it reduces the input structure to those invariant features that are sufficient for categorization in a given context. This second representation is still a shape preserving representation at least for the invariant features. It enhances the between-category differences and within-category similarities, known as the cue-validity effect.[14] Finally, a symbolic representation is derived.

For our goals, we must know that features are structured and must be recovered in a certain order: position (center of a part or of the whole object) followed by orientation (assuming that a principal axis exists), size (along the principal axis), and some shape descriptions. The first two primitives determine the transformation between the world or observer coordinate system and the object centered coordinate system. While there is not much dispute on how to represent rigid transformations, the best way of encoding shape information is still an open question and addressed in the next section. However, to continue the above sequence, the recovery of gross shape features should precede the recovery of shape details. The second important point is that a picture is always worth more than a symbolic description, since symbols can leave out some critical feature that would be necessary to sort out some future instance. Eventually any symbolic system must be grounded in perceptual acquaintance.[5] Computer vision systems that rely only on symbolic representations are doomed to fail in real world applications.

Shape representation of categories

We pointed out in the previous sections that basic categories are most closely related to the structure of the perceived world and we should be able to model them. Shape representation of basic categories is difficult because the representation must allow for variations of shape within a category and yet differentiate between categories. In recognition we compare the iconic representation of the category prototype with the shape of the object under consideration. The relation that can establish equality among the model prototype and the object is *shape* deformation.

Deformation of shape is a process that affects both natural and man-made objects. It is a highly intuitive way of describing and thinking about objects. Deformation is not just any alteration of shape. It conveys the impression that the shape as a whole has undergone a change that can be modeled by some physical process. Thompson[15] pointed out that several natural forms are easily explained when they are regarded as deformations of a simple structure. A deformation is highly intuitive and easily visualized process which helps not only to explain natural forms, but also simulates some manufacturing processes for fabrication of objects. The attributes responsible for the subdivision of a basic category into subordinate categories often correspond to some deformation. It is not surprising, then, that for verbal descriptions of objects we often use adjectives that reflect some underlying deformation of a simpler, prototypical shape.[16] The deformation terms are often perceived as manifestations of physical processes acting on an imagined physical object with a prototypical shape. These ideas about prototypes and deformation were taken up for shape representation in computer graphics.[17,18,19]

But what are the primitives that undergo deformation? We get the answer by finding the right scale. The right level of granularity for representing basic categories seem to be primitives that correspond to the human notion of parts. Building object representations by putting together parts is a common practice in computer vision, computer graphics and design. While, in general, almost any set of primitive building blocks will do, as long as they enable easy manipulation, we are interested here in primitives that have a perceptual salience and hence reflect the structure in the world. The apparent complexity of our environment is produced from a limited vocabulary of parts by applying a small set of generic processes over and over again.[19,20,21] Pentland[19] in particular, has shown that such part based models can describe a large class of man-made and natural objects and advocated their immediate recognition in images instead of building point-by-point descriptions of surfaces and volumes.

This change of focus to the structure in images[22,23,19,20,24] was due also to the largely unsuccessful point-based quantitative methods which use overconstrained assumptions about illumination, surfaces etc. in real world applications. The task of recovering structure, also known as perceptual organization, was first investigated by the Gestalt school in psychology.[25] People are able to perceive structure and pick out parts in images from various sources (i.e., line drawings, photographs, x-rays, etc.) apart from recognizing familiar objects.[19] The addition of semantic context rarely affects this spontaneous, pre-attentive organization of images into parts. The basis of this remarkable capability may be the fact that regular relationships like parallel lines, curvilinearity of arcs, symmetry, two or more terminations of vertices at a common point and so on, are very *unlikely to arise by chance.*[22]

When looking for a partitioning rule, two possibilities exist. One is to formalize the decomposition of objects into parts by defining part boundaries in terms of differential geometry.[26,23] The other way of defining parts is to describe the shape of possible parts. Pentland[19] proposed that by applying simple deformations and boolean combination to superquadric primitives one could capture much of the human notion of "parts." This theory of parts satisfies the topological constraints on part structure demonstrated by Hoffman and Richards.[23]

Shape representation for object recognition

Object recognition is a problem of search. Most current vision systems have a small number of rigid models. The basic procedure they use is to find corresponding features in image and model and perform a least-squares fit or subgraph isomorphism. There are two unknowns in this model-based search: the viewpoint and the right model. Some systems concentrate just on efficient solving of the first unknown, assuming identical models like the bin picking problem.[27] The solution is normally done in a "predict, match and back project" paradigm. Given a small number of rigid models like those found in industrial applications, linear search or heuristic feature-indexing techniques are possible.[28] But a vision system that has to function in an unrestricted environment cannot handle the visual input with a limited set of rigid and precise models. It has to navigate, locate and recognize objects. All these tasks should be preferably solved in the same framework by refining the information in each step. Using the analogy of search again, search can be sped up by using "bucket" search or hash functions. As we argued before, the most convenient "buckets" for visual perception are basic categories because basic categories reflect the perceived structure of the world. Hence, another way to speed up visual processing is to use those larger "chunks" or parts that are the result of perceived world structure.

Pixel based qualitative methods criticized by researchers in perceptual organization are also very sensitive to noise, missing data, errors and waste resources just trying to account for shape details that are not relevant for most vision tasks anyway. Beside arguments about ordering of features, categorization, perceptual organization, and speed, this is yet another reason to bypass recovery of intrinsic characteristics (i.e., edges, surface patches) and recognize intermediate level models directly.

A popular mid-grain primitive for shape representation is generalized cylinders,[29] their most notable application being Brook's ACRONYM system.[30] However, generalized cylinders were originally not intended as part representation. Generalized cylinders have large expressive power, but in practice only some subsets of them get used (i.e., linear strait homogeneous generalized cylinders[31]). After defining superquadrics in the next section, we argue why superquadrics are more suitable for shape representation on the basic category level.

Superquadrics

Superquadrics can be compared to lumps of clay that can be deformed and glued together into very realistic looking models (see Pentland's *SuperSketch* graphics system[19]). Superquadrics are a family of parametric shapes that were invented by the Danish designer Peit Hein[32] as an extension of basic quadric surfaces and solids (see also Barr[17]). Superquadric surface is defined by the following column vector:

$$\vec{x}(\eta, \omega) = \begin{bmatrix} a_1 C_\eta^{\varepsilon_1} C_\omega^{\varepsilon_2} \\ a_2 C_\eta^{\varepsilon_1} S_\omega^{\varepsilon_2} \\ a_3 S_\eta^{\varepsilon_1} \end{bmatrix} \qquad (1)$$

where $-\pi/2 \leq \eta \leq \pi/2$ and $-\pi \leq \omega < \pi$. C_η stands for $cos(\eta)$ and S_η for $sin(\eta)$. Parameters η and ω correspond to latitude and longitude angles of vector $\vec{x}$ expressed in spherical coordinates. Angle ω lies in the x-y plane while η corresponds to the angle between $\vec{x}$ and its projection in x-y plane. Scale parameters a_1, a_2, a_3 define the size of superquadrics in x, y and z directions, respectively. ε_1 is the squareness parameter along the z axis and ε_2 is the squareness parameter in the x-y plane. Superquadrics can model a large set of standard building blocks, like ellipsoids, cylinders, and parallelopipeds (Figure 1).

The surface normal vector can be computed anywhere on the surface from the cross product of two tangential vectors and expressed as a function of components of the surface position vector $\vec{x}$:

$$\vec{n}(\eta, \omega) = \begin{bmatrix} \frac{1}{a_1} C_\eta^{2-\varepsilon_1} C_\omega^{2-\varepsilon_2} \\ \frac{1}{a_2} C_\eta^{2-\varepsilon_1} S_\omega^{2-\varepsilon_2} \\ \frac{1}{a_3} S_\eta^{2-\varepsilon_1} \end{bmatrix} = \begin{bmatrix} \frac{1}{x_s} C_\eta^{2} C_\omega^{2} \\ \frac{1}{y_s} C_\eta^{2} S_\omega^{2} \\ \frac{1}{z_s} S_\eta^{2} \end{bmatrix} \qquad (2)$$

This relation between the surface position vector $\vec{x}$ and the surface normal vector $\vec{n}$ enables straightforward rendering even of deformed superquadrics (Figure 1).

Interpretation of sparse 3-D points with superquadrics

Superquadrics are appropriate models not only for computer graphics or CAD design,[17,33,34] but also for analysis of scenes in computer vision as suggested first by Pentland.[19] We do not believe that superquadrics are the best solution for all problems in computer vision. We believe for several reasons, however, that they are appropriate as part-based models for the class of basic categories. On this level, very detailed shape descriptions are not necessary. Variations of shape for the prototype and deformation paradigm are easy to describe either with the two shape parameters or with global deformations. With a *small* set of parameters a *large* set of shape primitives can be *uniformly* handled. Besides, superquadrics model the whole object, including the occluded sides, by assuming global symmetry.

Superquadrics are suitable models for computer vision because we can form *overconstrained* estimates of their parameters. This overconstraint comes from using models defined by a few parameters to describe a large number of 3-D points. This enables us to verify our estimated models and measure the "goodness of fit." For a superquadric in general position we have to estimate 11 parameters: location in space (3 par.), orientation in space (3 par.), size (3 par.), and two shape parameters, ε_1 and ε_2. On the other hand, many more 3-D points are typically available on the surface of the modeled object from either range imaging or passive stereo. To find the parameters so that the model best fits the data is called an overdetermined optimization problem.

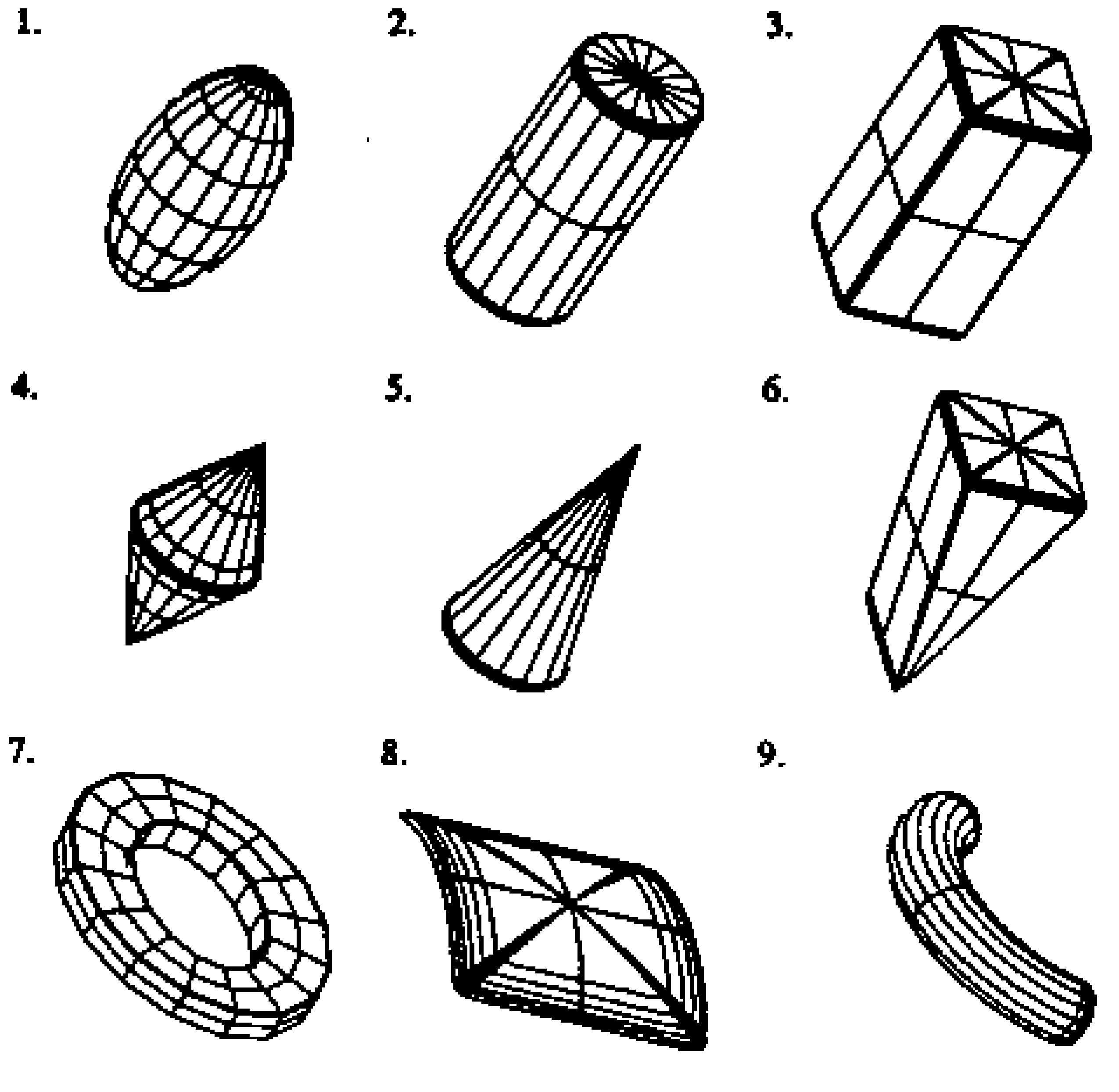

Figure 1: Examples of superquadrics: 1. superellipsoid ($\varepsilon_1, \varepsilon_2 = 1$), 2. cylinder ($\varepsilon_1 = 0.1, \varepsilon_2 = 1$), 3. parallelopiped ($\varepsilon_1 = 0.1, \varepsilon_2 = 0.1$), 4. double cone ($\varepsilon_1 = 2, \varepsilon_2 = 1$), 5. cone formed by tapering a cylinder, 6. wedge formed by tapering a parallelopiped, 7. toroid superquadric (requires two additional parameters[17]), 8. and 9. bent parallelopipeds.

Pentland[19] first suggested an analytic solution for all independent parameters using parametric equations (1) and (2), adjusted for general position. As input he proposed information from 2-D contours and shading. Using linear regression, one could compute parameter values that provide the best fit. Pentland[35] currently recovers superquadrics from range data by computing a heuristic "goodness-of-fit" functional in a coarse grain search over the *entire* parameter space. We believe, due to the complex relationship of the 11 independent parameters which are required for a superquadric in general position, that an analytic solution for superquadric parameters is not practical. A heuristic approach, on the other hand, lacks precision, and global search is computationally too expensive. We introduce here a relatively fast *iterative* fitting procedure based on implicit equations of superquadric surfaces and surface normals. We first eliminate parameters η and ω from eq. (1) by using trigonometric identity $C_x^2 + S_x^2 = 1$ to get an implicit equation of a superquadric surface:

$$\left[\left[\frac{x_s}{a_1} \right]^{\frac{2}{\varepsilon_2}} + \left[\frac{y_s}{a_2} \right]^{\frac{2}{\varepsilon_2}} \right]^{\frac{\varepsilon_2}{\varepsilon_1}} + \left[\frac{z_s}{a_3} \right]^{\frac{2}{\varepsilon_1}} = 1 \qquad (3)$$

We define the "inside-outside" function as:

$$F\,(x_S, y_S, z_S) \; = \tag{4}$$

$$\left[\left(\left(\frac{x_S}{a_1} \right)^{\frac{2}{\varepsilon_2}} + \left(\frac{y_S}{a_2} \right)^{\frac{2}{\varepsilon_2}} \right)^{\frac{\varepsilon_2}{\varepsilon_1}} + \left(\frac{z_S}{a_3} \right)^{\frac{2}{\varepsilon_1}} \right]^{\varepsilon_1}$$

With the outermost exponent ε_1 we force F to grow quadratically instead of exponentially. This is critical to ensure proper convergence during model recovery. When $F\,(x_S, y_S, z_S) = 1$, the point (x_S, y_S, z_S) is on the surface of the superquadric. If $F\,(x_S, y_S, z_S) > 1$, the corresponding point lies outside and if $F\,(x_S, y_S, z_S) < 1$ inside the superquadric. Equation (3) defines the surface in a superquadric centered coordinate system (x_S, y_S, z_S) but 3-D points from passive stereo or range imaging are given in a world coordinate system (x_W, y_W, z_W). We have to express these 3-D points in the superquadric centered coordinate system which is done by a translation and a sequence of rotations. A 4×4 matrix T is a convenient way of expressing such transformation in homogeneous coordinates:

$$\begin{bmatrix} x_S \\ y_S \\ z_S \\ 1 \end{bmatrix} = T \begin{bmatrix} x_W \\ y_W \\ z_W \\ 1 \end{bmatrix} \tag{5}$$

where

$$T = Trans\,(p_x, p_y, p_z) \cdot Rot\,(\phi, \theta, \psi) \tag{6}$$

We use Euler angles to express the orientation in terms of rotation ϕ about the z axis, followed by a rotation θ about the new y axis, and finally, a rotation about ψ the new z axis. Using equations (5) and (6) we can rewrite equation (4) to get a superquadric in general position and orientation:

$$F\,(x_W, y_W, z_W) \; = \tag{7}$$

$$F\,(x_W, y_W, z_W, a_1, a_2, a_3, \varepsilon_1, \varepsilon_2, \phi, \theta, \psi, p_x, p_y, p_z\,)$$

The independent parameters expressed in vector notation are: $\vec{a} = [a_1, a_2, \cdots, a_{11}]^T$. Suppose we have N 3-D surface points (x_W, y_W, z_W) which we want to model with a superquadric. We want to vary the 11 adjustable parameters a_j, j = 1, ... , 11 in eq. (7) to fit a superquadric surface to these 3-D points. The superquadric model predicts the position of a point (x_W, y_W, z_W) relative to the surface of the model. For points on the surface of a superquadric:

$$F\,(x_W, y_W, z_W; a_1, \cdots, a_{11}\,) = 1 \tag{8}$$

We want to get such values for a_j's that most of the 3-D points will lay on or close to the model's surface. We achieve this by minimizing:

$$\sum_{i=0}^{N} [1 - F\,(x_{W_i}, y_{W_i}, z_{W_i}; a_1, \cdots, a_{11})\,]^2 \tag{9}$$

Since the model in eq. (7) is a nonlinear function of 11 parameters a_j, j = 1, ... , 11, the minimization must proceed iteratively.[36] Given trial values for $\vec{a}$, we evaluate eq. (7) and employ a procedure that improves the trial solution. The procedure is then repeated with new trial values until the sum of least squares (eq. 9) stops decreasing or the changes are statistically meaningless. Since first derivatives $\partial F / \partial a_i$ for $i = 1, \cdots, 11$ can be computed, we use the Levenberg-Marquardt method for nonlinear least squares[37] This method varies smoothly between the extremes of the inverse-Hessian method and the steepest descent method. The first trial set of parameters, $\vec{a}$, must be set experimentally to some initial estimates $\vec{a}_i$. We compute the center of gravity of given 3-D points which corresponds roughly to the coordinate center of the superquadric and moments of inertia to estimate the orientation in space. The size of the superquadric is estimated along the axes of the local coordinate system. The default estimate for ε_1 and ε_2 is 1. During the fitting procedure we introduce "jitter" by adding Poisson distributed noise to the evaluation of function F. Local minima in the 11 parameter space are thus avoided and a global convergence assured.[35]

We first tested the fitting procedure on synthetic range data. We randomly chose points on a superquadric surface and added gaussian noise to the point coordinates. By changing the initial estimates we investigated the robustness of the minimization procedure. We were able to fit *simultaneously* all 11 parameters for initial estimates that differed up to 50% from the actual values. However, a problem shows up due to self-occlusion. For example, when using function F for fitting a model to the points that correspond to a cylinder in Figure 2, an infinite number of cylinders of different length satisfy eq.(9). Obviously only the model with the *smallest* possible volume is the desired solution. We solved the problem by multiplying F with a factor corresponding to the volume of the model: $a_1 a_2 a_3 F$. This new criteria function has a minimum that corresponds to the smallest superquadric that fits a set of 3-D points. However, now we have to know beforehand the correct values of parameters a_1, a_2, and a_3. This is fixed by defining the model as:

$$R = a_1\, a_2\, a_3\, (F - 1) \tag{10}$$

Now the value of function R for points on the surface of a superquadric is 0 and we must minimize:

$$\sum_{i=0}^{N} [R\,(x_{W_i}, y_{W_i}, z_{W_i}; a_1, \cdots, a_{11})\,]^2 \tag{11}$$

Function R also has a minimum for $a_1, a_2, a_3 = 0$. In practice, this was not a problem as long as the initial estimates for a_1, a_2, and a_3 were not too small (Figure 2).

Deformed superquadrics can be recovered using the same technique. Global deformations like tapering, bending, and twisting require just a few additional parameters. Consider, for example, tapering along the z axis:

$$X = f_x(z)\; x \tag{12}$$

$$Y = f_y(z)\; y$$

$$Z = z$$

where X, Y, Z are the surface vector components of the deformed superquadric, and f_x and f_y are the tapering functions in x and y axis. The corresponding "inside-outside" function of a tapered superquadric is:

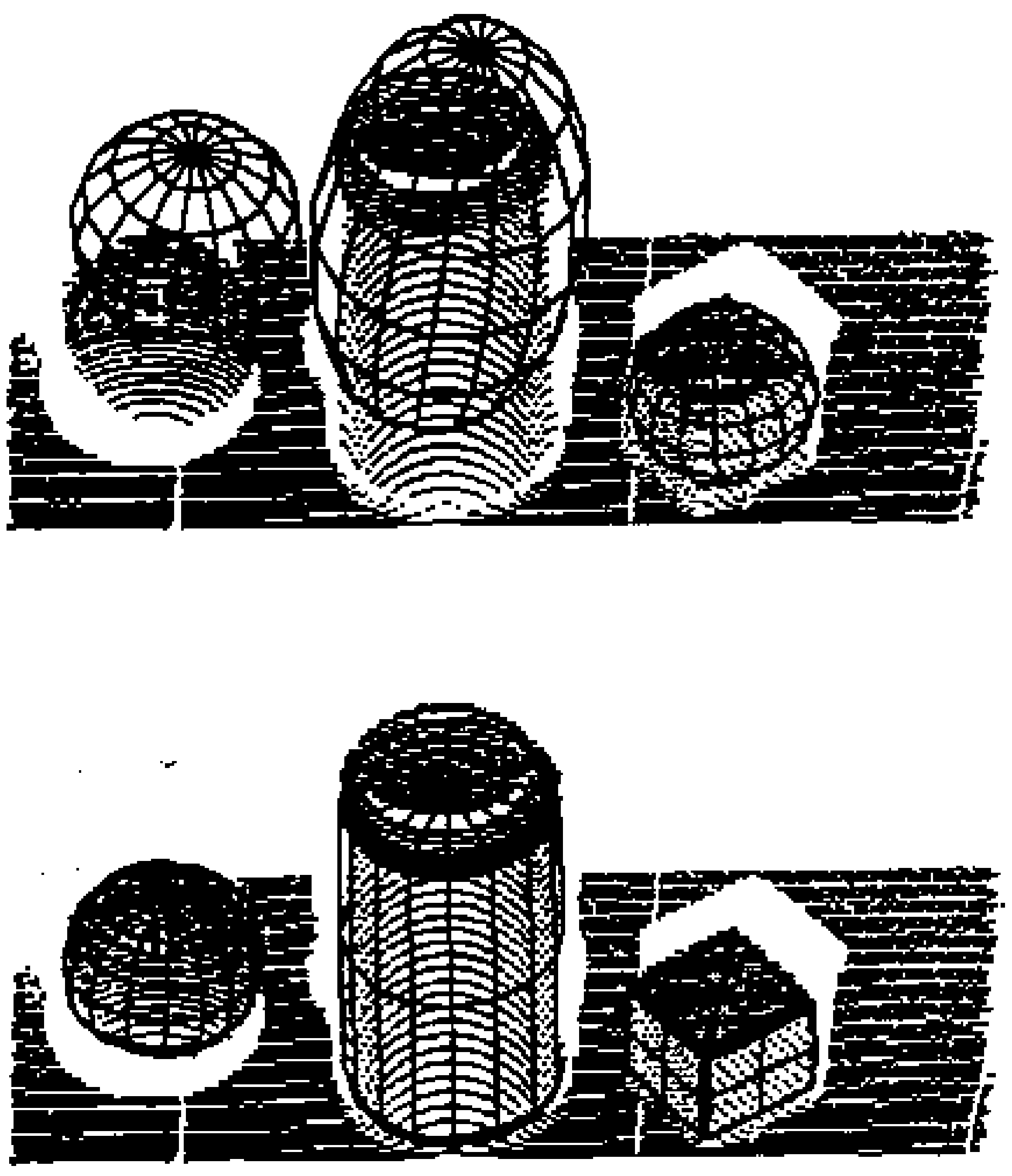

Figure 2: Interpretation of real range images [38] with superquadric models. In the top image are the initial estimated models. Bellow are the recovered models after about 20 iterations. All 11 parameters were adjusted simultaneously to achieve a least squares fit. Initial segmentation of the image into supporting surface and objects was done by hand. The small indentation on the top of the cylinder was averaged out by the model.

$$F(X_S, Y_S, Z_S) = \tag{13}$$

$$\left[\left(\left(\frac{X_S}{f_x(Z_S)\, a_1} \right)^{\frac{2}{\varepsilon_2}} + \left(\frac{Y_S}{f_y(Z_S)\, a_2} \right)^{\frac{2}{\varepsilon_2}} \right)^{\frac{\varepsilon_2}{\varepsilon_1}} + \left(\frac{Z_S}{a_3} \right)^{\frac{2}{\varepsilon_1}} \right]^{\varepsilon_1}$$

For linear tapering:

$$f_x(Z_S) = \frac{K_x}{a_3} Z_S + 1 \quad \text{and} \quad f_y(Z_S) = \frac{K_y}{a_3} Z_S + 1 \tag{14}$$

where $-1 \leq K \leq 1$. In Figure 3, the initial estimated model is not tapered. Parameters K_x and K_y are adjusted simultaneously with the other 11 parameters. *Any* shape deformation can be recovered in this way as long as the inverse transformation is available such that x, y, z components of the non-deformed superquadric can be expressed in terms of X, Y, Z components of the deformed superquadric and the necessary deformation parameters.

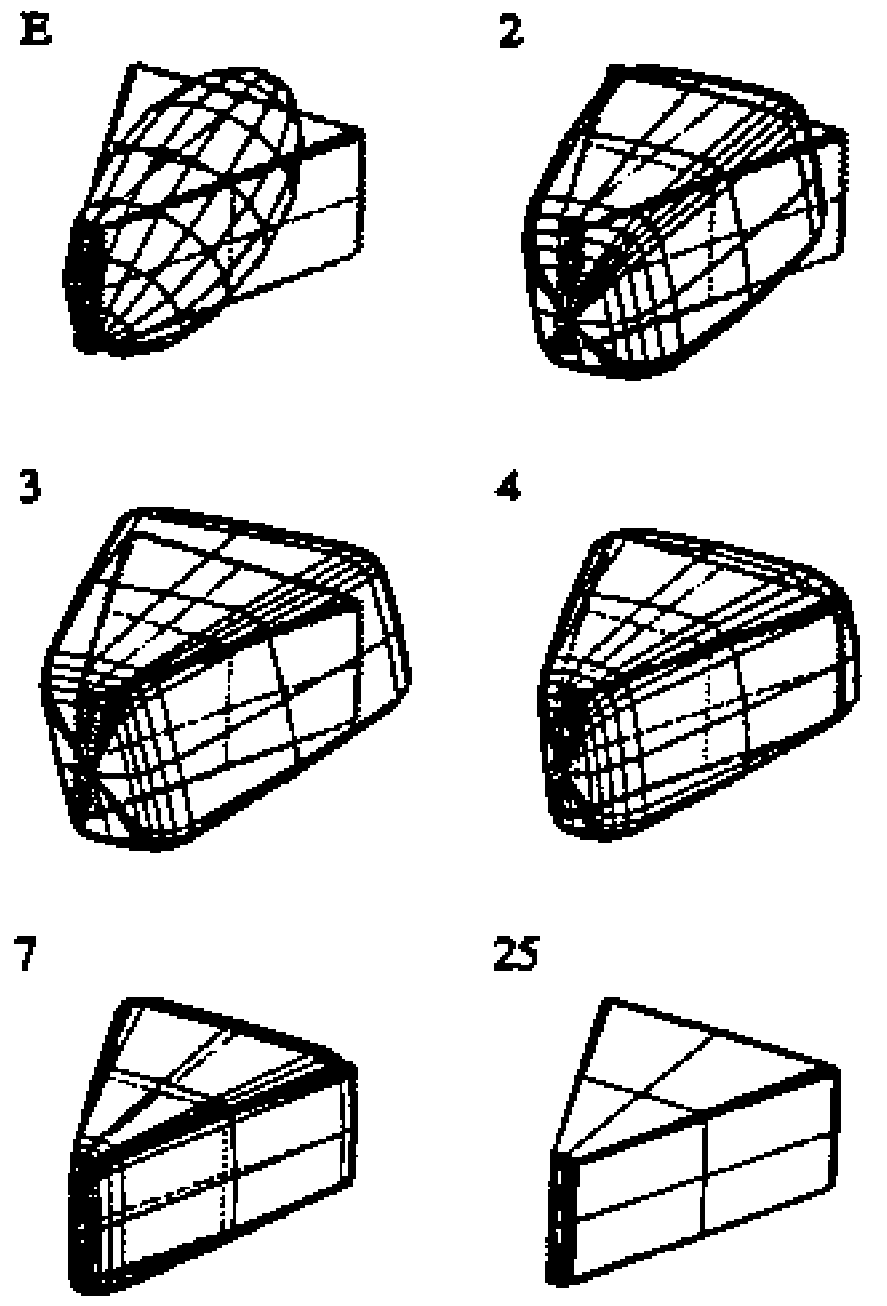

Figure 3: Recovery of a tapered superquadric. The initial estimate and some of the following iterations are shown superimposed on the light grey superquadric representing the input data (100 3-D points). All 13 model parameters were adjusted *simultaneously* to achieve a least squares fit. The whole fitting procedure took only about one minute on a VAX 785.

Use of contour information. When only sparse and very unevenly distributed 3-D points are available as is the case with passive stereo, *contour* information can be used. Visual contours of objects provide strong clues about the local shape of surface regions.[39] Lowe,[24] for example, used only 2-D contours for recognition of objects in general position. From eq. (2) an implicit equation for surface normal vector components can be derived:

$$E(n_x, n_y, n_z) = \tag{15}$$

$$\left[\left((n_x\, a_1)^{\frac{2}{2-\varepsilon_2}} + (n_y\, a_2)^{\frac{2}{2-\varepsilon_2}} \right)^{\frac{2-\varepsilon_2}{2-\varepsilon_1}} + (n_z\, a_3)^{\frac{2}{2-\varepsilon_1}} \right]^{2-\varepsilon_1}$$

Given that the viewing direction is parallel to the z axis of the world coordinate system, the surface normal at the occluding contour has just two components, n_{xw} and n_{yw} ($n_{zw} = 0$), since the surface normal at the occluding contour is perpendicular to the viewing direction. The equation for superquadric surface normals in general orientation is:

$$E\,(n_{xw}, n_{yw}) = \tag{16}$$

$$E\,(n_{xw}, n_{yw}, a_1, a_2, a_3, \varepsilon_1, \varepsilon_2, \phi, \theta, \psi)$$

Note that, since normal vectors represent direction, the addition of any other finite vector does not change their value in any way. Hence surface normals for superquadrics in general position do not depend on the translational part of the homogeneous transformation T of eq. (6). Note also that the magnitude of the vector must correspond to the area of the corresponding local surface patch. This area can be estimated from the length of the contour segment that corresponds to the surface normal. Since, for normals on the occluding contour, $E\,(n_{xw}, n_{yw}; a_1, \cdots, a_8) = 1$, fitting can be achieved by minimizing:

$$\sum_{i=0}^{N} [R\,(x_{W_i}, y_{W_i}, z_{W_i}; a_1, \cdots, a_{11})]^2 + \tag{17}$$

$$\sum_{i=0}^{M} [1 - E\,(n_{xw_i}, n_{yw_1}; a_1, \cdots, a_8)]^2$$

where N is the number of 3-D points, and M is the number of normal vector estimates. This criterion turns out to be similar to Dane's[40] criterion for fitting range data to general quadric surfaces.

Prototype model

The proposed category model consists of prototypical part models with possible variations in structure and deformations of parts to account for the variation inside the category (Figure 4). Winston et al.[41] considered a model with similar flavor for learning physical descriptions from functional definitions based on the ACRONYM modeling system.[30]

An object model in the proposed scheme is a hierarchical structure, where the position of each part prototype is defined in the coordinate system of a part on the preceding higher level, like in the Marr-Nishihara model.[42] Allowable changes in relative position among parts to account for differences in a category are specified for each joint by giving a range of possible values for each parameter. Individual parts are modeled with superquadric primitives. Variations are described by specifying a range of values for each superquadric parameter including deformation parameters. Isomorphism of sets of superquadric parameters must be resolved when matching. Clearly, changes in structure and shape of different parts are interrelated, i.e., little cups don't have huge handles although little cups and large handles by themselves legal. One should somehow represent this interrelationship. However, even people have problems categorizing border cases (see Witgenstein[9]), especially if puzzling objects are purposely constructed.

Modeling of cavities deserves a special note. Cavities can be modeled either by Boolean difference of two primitives or by bending a thin plate along two axes. For recovery it is better to use bending, since when Boolean difference is used, it is difficult to recover parameters of a primitive that is not there in the first place. Valuable insight on how to form models can be gained by examining how real objects are manufactured. A container would be made using deformation, while a hole small in comparison to the whole object is made by Boolean difference.

The proposed category model is part-based, but it also accommodates features. Each part or relation has a set of associated features and properties - tags. In well specified contexts, feature indexing enables faster recognition. Object recognition by satisfying a set of features is standard in pattern recognition. All features in the proposed model must be universal enough to be representative for all objects that belong to a particular category. Hence, allowable deformations must not distort any of the features.

Recognition procedure

The category model from the previous section allows for flexibility in recognition. The category model can be accessed either through superquadric parameters or in a known context more quickly through features. The input to the recognition procedure is, in each case, sparse 3-D points from passive stereo or range imagery. The task of the low level vision is to segment the image into parts so that superquadrics can be recovered or some features can be extracted (i.e. holes, cavities, handles, parallel surfaces, parallel edges, rotational symmetries, number of parts, and relative measures like ratio of width/height, or ranges with lower and upper bounds of object dimensions). What features can be detected depends on the sophistication of the low level vision and the geometric reasoner which combines the results of low level vision into features.

When superquadric parameters are recovered, verification consists of checking if the parameters and part structure are in a specified range. Recovery can be speeded up by not only making better estimates for position, orientation, and size, but also to detecting the presence of different types of deformation and hence the necessity to recover their parameters. Bending, for example, can be inferred from the medial axis transform. For modeling containers other default parameter values can be selected, closer to the expected ones. Since deformations are not commutative we have to adopt a predefined sequence of deformations during recovery. Leyton[43] demonstrated strong differential-geometric constraints on the decomposition of prototypification for 2-D shapes when regarding prototypes as shapes with global symmetry structure. This work might suggest the optimal type and sequence of deformations.

Another strategy is necessary for feature based recognition. For a recovered feature, all objects that have that feature can be selected and their position and orientation hypothesized.[28] Two techniques are used for comparison, one looks for similarity or common features, the other for differences or distinctive features to set two objects or concepts apart. Similarity seems a good mechanism for initial hypothesis generation while differences seem suited for selecting the best hypothesis among the best few. Models that are not consistent with all extracted features are eliminated from the set of possible models. Since features alone are not sufficient for identification, (different arrangements of the same features may correspond to different objects) hypothesis verification has to measure how well the possible models fit the data. For verification, shape comparison on the geometric level is more reliable then with symbolic representation. Instead of discarding geometric information in favor of symbolic we suggest to keep the *geometric* information and work with it whenever it is more convenient. Verification with the

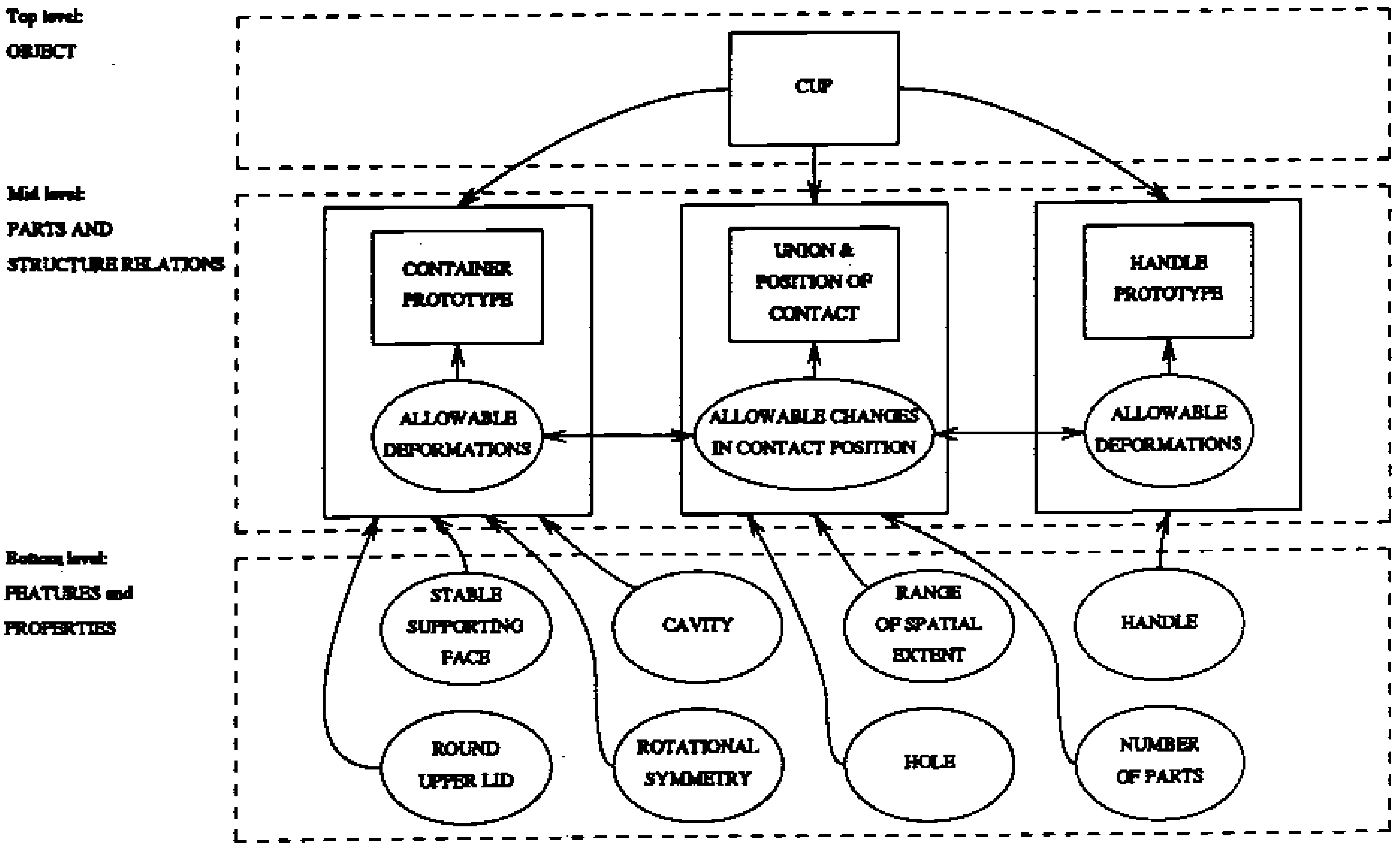

Figure 4: A category model for cups. Category models consist of three levels. The upper level is the category label that applies to all objects in the category. The middle level are the prototypical parts and their structural relations together with allowable changes in structure and allowable shape deformations of part primitives to accommodate variations inside the category. The bottom level specifies the associated features and properties that point to parts or their relation/structure on the second level.

proposed model then consists of checking how the parts are put together and deforming each part model in allowable limits to match it to the input data.

If more than one category model matches the data, either the data is not sufficient to discriminate among the possible categories or the object is a borderline case (since some objects may actually match two different categories). The more a prototype must be deformed, the less typical is the object for the category. By comparing two competing models one can find out whether the information hidden due to self-occlusion can resolve the ambiguity. Changing viewing direction or rotating the object to get information hidden by occlusion may help.[44]

Discussion

The most important assumption we adopted for shape representation is that objects can be modeled as a prototype and a set of shape deformations that account for variations inside an object category. We based our work on Pentland's[19] ideas on applying deformations and boolean operations on superquadric models to capture the human notion of parts. We showed how these ideas on shape representation connect to recent psychological research on categorization; in particular to basic categories.[6] Sets of similar objects can also be modeled with general cylinders by defining a range of values for each parameter.[30] However, this approach has difficulties describing shape deformations or to recover model parameters in a direct way. Stansfield[45] constructed a type of generic models for object recognition in an active multisensor system.

We also argued that recovery of shape information should be done in a structured manner, from general to more specific.[8] Then different vision tasks, like path planning and object recognition, can be studied in the same framework.

A large part of computer vision community has so far rejected superquadric models under assumption that, due to the two exponents ε_1 and ε_2, large errors or instability in estimation are inevitable. We believe that the new recovery procedure introduced in this paper will prove to the contrary. Recovery using the "inside-outside" function is much faster compared to the only other method known to us.[35] Recovery speed depends on the number of 3-D points, the number of model parameters, and the accuracy of the initial estimates. Still, circumstances under which some of the parameters cannot be recovered because of the violation of general view (i.e., when just one face of a cube is visible[46]) should be studied. We also have to address automatic segmentation of scenes into parts appropriate for superquadric modeling. It seems plausible that segmentation should work hand in hand with model recovery.[35] In this paper we presented just one possible way of object recognition - by recovering superquadric parameters. However, feature tags in the proposed category model can serve not only for recognition in well defined contexts, but also, to interface the vision system to a natural language query system or to address the problem of inferring function from shape.

Relation between shape and function. Gombrich[21] noted that for human made symbols (his example is a hobby horse), the common factor is function, rather then form, or just that formal aspect which fulfills the minimum requirement for the performance of the function. Tversky and Hemenway[47] pointed out the particular salience of *parts* on the basic category level. For objects and biological categories, basic category cuts seem to follow natural breaks in the structure of the world and this structure is determined by part configuration. In comparison to the basic level, the proportion of parts of all the common features decreases for both super and subordinate categories. Parts and part configuration seem to form a natural bridge connecting *perception* (appearance) of objects and *behavior* (activity) toward them, and in turn *communication* about them. Perceived part configuration underlies both perceived structure and perceived function, and forms the basis of intuitive causal reasoning and naive induction. The basis of naive induction is that separate parts have separate functions, similar parts have similar functions, and more salient parts have more important functions. This close relation opens up possibilities to infer function from shape. Feature tags in the proposed basic category model could describe also the corresponding function.

Function and functionality in this context are meant as a proper action or a design which fulfills its purpose. Functions of man-made objects are defined in terms of the user (agent) model and goal model. But parts and function are related also independent of human users when organisms or objects are studied as self-contained systems. The functional basis of shape in nature was first investigated in depth by Thompson.[15] He pointed out that the repetition of shapes is not accidental; the same shapes are encountered across species for the same function since they were molded by the same physical laws. (See also Leyton[43]). The "form follows function" hypothesis carries over also to man-made objects, where despite the large diversity found in design, certain *functional* dimensions must be met for adequate usage. This seems to define objects on the basic category level. The subordinate category level of man-made objects is the result of available technologies, skill in manufacturing, habits, and esthetic preference. Since this whole set of often disparate requirements influences the design process, the task of synthesis of form is not easily formalized.[48] We use kitchen objects for investigating the relationship between shape and function since their gross shape corresponds to a variety of classes (functions); like blob-like (for containment), flat (for support), and elongated (for manipulation). They are a challenging set for shape representation since they have holes, handles, cavities, but with a clearly defined function or purpose.

Other work concerned with the relation between shape and function is the naive physics program,[16] work by Davis,[49] and by Winston et al[41] on learning physical descriptions of objects from functional descriptions, examples, and precedents. It is interesting to note that archeologists also work on inverse design. Given excavated artifacts or some of their parts they want to determine among other facts also their function. In a way, the whole process of visual perception is to the instant visual input the same, as archeology is to the excavated artifacts. Unfortunately much of the machine vision tends to emulate the speed of archeological analysis rather than the instantaneous human vision.

Conclusions

We motivate modeling of basic categories for general purpose computer vision. The representation is part based but feature tags can support also object recognition in well defined contexts. We present a new way of recovering superquadric models by iterative minimization of the inside-outside function. The method seems to be fast and robust and is not affected by noise, and partial or missing data.

References

[1] N. Badler and R. Bajcsy , "Three-dimensional representation for computer graphics and computer vision," *ACM Computer Graphics (Proceedings Siggraph 78)* 12(3) pp. 153-160 (1978).

[2] M. Brady and H. Asada, "Smoothed local symmetries and their implementation," *International Journal of Robotics Research* 3(3) pp. 36-61 (1984).

[3] O. D. Faugeras and M. Hebert, "The representation, recognition, and locating of 3-D objects," *International Journal of Robotics Research* 5(3) pp. 27-52 (1986).

[4] E. H. Gombrich, J. Hochberg, and M. Black, *Art, perception, and reality,* Johns Hopkins University Press, Baltimore (1972).

[5] S. Harnad, "Category Induction and Representation," in *Categorical Perception,* ed. S. Harnad, Cambridge University Press (1987).

[6] E. Rosch, "Principles of Categorization," in *Cognition and categorization,* ed. E. Rosch and B. Lloyd, Erlbaum, Hillsdale, NJ (1978).

[7] E. Goldmeier, *Similarity in Visually Perceived Forms,* International Universities Press, New York (1972).

[8] M. Leyton, "Principles of information structure common in six levels of the human cognitive system," *Information sciences* 38(1) pp. 1-120 (1986).

[9] L. Wittgenstein, *Philosophical Investigations,* Macmillan, New York (1953).

[10] F. Solina and R. Bajcsy, "Modeling of Mail Pieces with Superquadrics," Proceedings USPS Advanced Technology Conference, Washington, DC pp. 472-481 (1986).

[11] T. Henderson, C. Hansen, A. Samai, C. C. Ho, and B. Bhanu, "CAGD-Based 3-D Visual Recognition," Proceedings 8th ICPR, Paris pp. 230-232 (1986).

[12] L. R. Novick and B. Tversky, "Cognitive constructs on ordering operations: The case of geometric analogies," *Journal of experimental psychology: general,* (In press).

[13] M. Leyton, "A theory of information structure I. General principles," *Journal of Mathematical Psychology* 30(2) pp. 103-160 (1986).

[14] E. Rosch and Mervis, "Family resemblance studies in the internal structure of categories," *Cognitive Psychology* 7 pp. 573-605 (1975).

[15] D'Arcy Thompson, *On growth and form,* Cambridge University Press, Cambridge, England (1917). (Abridged edition, Ed. J. T. Bonner, 1961)

[16] J. R. Hobbs, T. Blenko, B. Croft, G. Hager, H. A. Kautz, P. Kube, and Y. Shoham , "Commonsense Summer: Final Report," Report No. CSLI-85-35, Center for the Study of Language and Information, Stanford University (1985).

[17] A. H. Barr, "Superquadrics and angle-preserving transformations," *IEEE Computer Graphics and Applications* 1 pp. 11-23 (1981).

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

[19] A. P. Pentland, "Perceptual Organization and the Representation of Natural Form," *Artificial Intelligence* **28**(3) pp. 293-331 (1986).

[20] I. Biederman, "Human image understanding: Recent research and theory," *Computer Vision, Graphics, and Image Processing* 32 pp. 29-73 (1985).

[21] E. H. Gombrich, *Meditations on a hobby horse and other essays on the theory of art,* Universtiy of Chicago Press, Chicago (1985).

[22] A. P. Witkin and J. M. Tenenbaum , "On the role of structure in vision," in *Human and Machine Vision,* ed. J. Beck, B. Hope, A. Rosenfeld , Academic Press, New York (1983).

[23] D. D. Hoffman and W. A. Richards , "Parts of recognition," *Cognition* **18** pp. 65-96 (1985).

[24] David G. Lowe, *Perceptual Organization and Visual Recognition,* Kluwer, Boston (1985).

[25] M. Wertheimer, "Principles of perceptual organization," in *Source book of Gestalt psychology,* ed. W. H. Ellis, , London and New York (1923).

[26] J. Koenderink and A. van Doorn , "The shape of objects and the way contours end," *Perception* **11** pp. 129-137 (1982).

[27] R. C. Bolles and P. Horaud, "3DPO: A Three-Dimensional Part Orientation System," *International Journal of Robotics Research* **5**(3) pp. 3-26 (1986).

[28] T. F. Knoll and R. C. Jain , "Recognizing partially visible objects using feature indexed hypotheses," *IEEE J. of Robotics and Automation* **RA-2** pp. 3-13 (1986).

[29] T. O. Binford , "Visual perception by computer," *IEEE Conf. on Systems and Control,* (1971).

[30] R. A. Brooks, *Model-Based Computer Vision,* UMI Research Press, Ann Arbor (1984).

[31] K. Rao and R. Nevatia, "From sparse 3-D data directly to volumetric descriptions," Proceedings: DARPA Image Understanding Workshop (February 1987).

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

[33] I. Elliot, "Discussion and Implementation Description of Experimental Interactive SuperQuadric Based 3D Drawing System," Internal Report LP 2/IME4, European Computer-Industry Research Centre GmgH, Muenchen (July 1986).

[34] A. Pentland, "Toward an ideal 3-D CAD system," SPIE Conf. on machine vision and the man-machine interface, Los Angeles, CA (January 1987).

[35] A. Pentland, "Recognition by parts," SRI Technical note No. 406, Menlo Park, CA (1986).

[36] L. E. Scales, *Introduction to Non-Linear Optimization,* Springer, New York (1985).

[37] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering, *Numerical Recipies,* Cambridge University Press, Cambridge, England (1986).

[38] C. Hansen and T. Henderson, "UTAH Range Database," Technical Report UUCS-86-113, Computer Science Department, University of Utah, Salt Lake City (April 1986).

[39] J. Koenderink , "What does the occluding contour tell us about solid shape?," *Perception* **13** pp. 321-330 (1984).

[40] C. A. Dane, "An object-centered three-dimensional model builder," Ph.D. Thesis, University of Pennsylvania, Philadephia (1982).

[41] P. H. Winston, T. O. Binford, B. Katz, and M. Lowry , "Learning physical description from functional descriptions, examples, and precedents," Proceedings AAAI, Washington, DC pp. 433-439 (1983).

[42] D. Marr, *Vision,* Freeman, San Francisco (1982).

[43] M. Leyton, "A process-grammar for representing shape," *Artificial Intelligence,* (). To appear

[44] G. Hager, "Active reduction of uncertainty in multisensor systems," Technical report MS-CIS-86-76, GRASP Lab 79, University of Pennsylvania, Philadelphia (1986).

[45] S. A. Stansfield, "Representation and control within an intelligent, active, multisensor system for object recognition," Proc. 1985 Int. Conf. IEEE Systems, Man, and Cybernetics, Tucson, AR pp. 215-219 (1985).

[46] J. J. Koenderink and A. J. van Doorn, "The singularities of the visual mapping," *Biological Cybernetics* **24** pp. 51-59 (1976).

[47] B. Tversky and K. Hemenway, "Objects, Parts, and Categories," *Journal of Experimental Psychology: General* **113**(2) pp. 169-193 (June 1984).

[48] C. Alexander, *Notes on the Synthesis of Form,* Harvard University Press, Cambridge, MA (1964).

[49] E. Davis, "Shape and function of solid objects: some examples," Technical report No. 137, Courant Institute, New York (1984).