International Journal of Computer Vision, 13, 3, 311–330 (1994)

Integrating Qualitative and Quantitative Shape Recovery

SVEN J. DICKINSON
*Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A4**

DIMITRI METAXAS
Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104-6389

Received March 9, 1993; Revised July 16, 1993 and December 22, 1993

Abstract. Recent work in qualitative shape recovery and object recognition has focused on solving the "what is it" problem, while avoiding the "where is it" problem. In contrast, typical CAD-based recognition systems have focused on the "where is it" problem, while assuming they know what the object is. Although each approach addresses an important aspect of the 3-D object recognition problem, each falls short in addressing the complete problem of recognizing and localizing 3-D objects from a large database. In this paper, we first synthesize a new approach to shape recovery for 3-D object recognition that decouples recognition from localization by combining basic elements from these two approaches. Specifically, we use qualitative shape recovery and recognition techniques to provide strong fitting constraints on physics-based deformable model recovery techniques. Secondly, we extend our previously developed technique of fitting deformable models to occluding image contours to the case of image data captured under general orthographic, perspective, and stereo projections. On one hand, integrating qualitative knowledge of the object being fitted to the data, along with knowledge of occlusion supports a much more robust and accurate quantitative fitting. On the other hand, recovering object pose and quantitative surface shape not only provides a richer description for indexing, but supports interaction with the world when object manipulation is required. This paper presents the approach in detail and applies it to real imagery.

1 Introduction

Since the introduction of a class of qualitatively-defined volumetric primitives, called *geons* (Biederman 1985), interest has been growing in building 3-D object recognition systems based on qualitative shape (Bergevin and Levine 1989; Biederman et al. 1992; Dickinson et al. 1992a; Dickinson et al. 1992b; Jacot-Descombes and Pun 1992; Fairwood 1991; Raja and Jain 1992). One of the primary motivations in these systems is that, as stated by Biederman (1985), the task of recognizing (or identifying) an object should be separated from the task of locating it. Furthermore, the exact shape of the object need not be recovered to facilitate recognition; a coarse-level description of shape is sufficient to distinguish between different classes of objects. The above systems, therefore, address only the task of recognizing the object. This is in contrast to classical 3-D object recognition systems, in which exact viewpoint is required to verify typically weak object hypotheses, while the object models capture the exact geometry of the object (Clemens 1991; Huttenlocher 1988; Lowe 1985; Thompson and Mundy 1987). Determining the pose of the object is a critical component of these approaches.

Each of the above recognition schools addresses an important requirement of recognition systems: coarse object identification and object localization. However, there has been little effort to combine them into a single paradigm. In cases where detailed object localization and shape recovery is required, the qualitative shape recovery methods fall short, while in cases where

*Current address: Center for Cognitive Science and Department of Computer Science, Rutgers University, P.O. Box 1179, Piscataway, NJ 08855-1179.

there are large object databases whose models are invariant to minor changes in shape, the quantitative recognition/localization methods fall short. In this paper, we unify these two schools into a single approach which first recovers qualitative shape from an image, and then uses that shape to constrain a quantitative recovery of the object's shape and pose.

Physics-based modeling (Pentland and Sclaroff 1991; Terzopoulos et al. 1988; Terzopoulos and Metaxas 1991; D. Metaxas and D. Terzopoulos 1991; Metaxas 1992; Metaxas and Terzopoulos 1993) provides a very powerful mechanism for quantitatively modeling an object's shape for recognition. In a typical geometry-based model-driven recovery process, image features are matched to a set of rigid, a priori object models which dictate the exact geometry of an object and offer few degrees of freedom. In contrast, deformable models offer a less constrained, data-driven recovery process, in which forces derived from the image deform the model until it fits the data. In previous work, we developed a physics-based framework for recovering shape and nonrigid motion from both 2-D and 3-D data using a new class of deformable part models (Terzopoulos and Metaxas 1991, Metaxas and Terzopoulos 1991; Metaxas 1992). These models incorporate global deformation parameters which represent the salient shape features of natural parts, and local deformation parameters which capture shape details. Thus, unlike previous physics-based techniques (Terzopoulos, Witkin, and Kass 1988), the shape of an object can both be abstracted or represented in detail. The framework also develops physics-based constraints to recover complex articulated models from deformable parts, and provides force-based techniques for fitting such models to sparse, noise-corrupted 2-D and 3-D visual data. These techniques lead to estimators that exploit the dynamic formulation of deformable models to track moving 3-D objects from time-varying observations.

As powerful as these and other active, deformable model recovery techniques are, they have some serious limitations. Their success relies on both the accuracy of initial image segmentation and initial placement of the model given the segmented data. For example, such techniques often assume that the bounding contour of a region belongs to the object, a problem when the object is occluded. In addition, such techniques often require a manual segmentation of an object into parts. For these techniques to be successful in an autonomous recognition system, it is imperative that more attention be given to the initial segmentation of an image into parts. Furthermore, since deformable model recovery techniques are sensitive to initial model position and orientation, the segmentation procedure should provide at least a coarse estimate of position and orientation. Finally, to constrain the process of recovering a deformable model from an image, the segmentation process should extract the qualitative shape of the part, e.g., how many surfaces does it have?; what shape are the surfaces?; etc.

In recent work, we presented an approach to the representation, recovery, and recognition of qualitative 3-D objects from a single 2-D image (Dickinson et al. 1990; Dickinson et al. 1992a, 1992b). In that approach, an object is modeled using a set of object-centered 3-D volumetric modeling primitives; the primitives, in turn, are mapped to a set of viewer-centered aspects. Unlike typical aspect-based recognition systems that use aspects to model entire objects, the approach uses aspects to model the finite set of *parts* from which the objects are constructed; the resulting aspect set is fixed and *independent* of the size of the object database. To accommodate the matching of partial aspects due to primitive occlusion, a hierarchical aspect representation was introduced, called the *aspect hierarchy*, based on the projected surfaces of the primitives; a set of conditional probabilities captures the ambiguity of mappings between the levels of the hierarchy. Once an aspect is recovered from an image, a qualitative volumetric shape primitive is inferred from the aspect. However, a limitation of the recovered primitive is that it simply encodes a shape class; no quantitative shape information such as size and curvature, or accurate position and orientation is specified. The inclusion of such information would not only enhance the descriptive power of the recovered primitives, thereby increasing their indexing power for recognition, it is essential for object manipulation.

In this paper, we integrate qualitative and quantitative shape recovery from 2-D images. In particular, we use knowledge of both a primitive's qualitative shape and its orientation (encoded by its aspect) to provide strong constraints in fitting a deformable model to the contour data. Since the qualitative primitive recovery technique supports primitive occlusion through a hierarchical aspect representation, it can selectively pass to the model fitting stage only those contours belonging to the object. In addition, the correspondence between image faces and model surfaces encoded in the recovered qualitative primitive can be exploited to provide strong constraints on the initial placement of the deformable model. Furthermore, this correspondence allows us to extend our previously developed technique for deformable model fitting, which is limited to orthographic projection of occluding boundaries (Terzopoulos and Metaxas 1991), to the case of more general objects with internal surface discontinuities under general orthographic, perspective, and stereo projections.

The paper is organized as follows. First we describe both the qualitative and quantitative object modeling paradigms. Next, we describe both the qualitative and quantitative shape recovery processes. Finally, we evaluate the performance of the approach on a series of real images of objects under orthographic, perspective, and stereo projection.

2 Related Work

Recently, several researchers have proposed various segmentation techniques to partition image or range data, in order to automate the process of fitting volumetric primitives to the data. Most of those approaches are applied to range data only (e.g., Solina and Bajcsy 1990; Gupta 1991), while Pentland (1990) describes a two-stage algorithm to fit superquadrics to image data. In the first stage, he segments the image using a filtering operation to produce a large set of potential object "parts", followed by a quadratic optimization procedure that searches among these part hypotheses to produce a maximum likelihood estimate of the image's part structure. In the second stage, he fits superquadrics to the segmented data using a least squares algorithm. Pentland's approach is only applicable to the case of occluding boundary data under simple orthographic projection, as is true of earlier work of Terzopoulos et al. (1988), Terzopoulos and Metaxas (1991), and Pentland and Sclaroff (1991), which address only the problem of model fitting.

The fundamental difference between our approach and the above approaches is that we use a qualitative segmentation of the image to provide sufficient constraints to our deformable model fitting procedure. In addition, we generalize our deformable model fitting technique to accommodate orthographic, perspective, and stereo projections.

3 Object Modeling

3.1 Qualitative Shape Modeling

In this section, we briefly review the qualitative shape modeling technique described in (Dickinson et al. 1990, 1992a, 1992b).

3.1.1 Object-Centered Models. Given a database of object models representing the domain of a recognition task, we seek a set of three-dimensional volumetric primitives that, when assembled together, can be used to construct the object models. Many 3-D object recognition systems have successfully employed 3-D volumetric primitives to construct objects. Commonly used classes of volumetric primitives include polyhedra (e.g., Lowe 1985), generalized cylinders (e.g., Brooks 1983), and superquadrics (e.g., Pentland 1986). Whichever set of volumetric modeling primitives is chosen, they will be mapped to a set of viewer-centered aspects.

To demonstrate our approach to object recognition, we have selected an object representation similar to that used by Biederman (1985), in which the Cartesian product of contractive shape properties gives rise to a set of volumetric primitives called geons. For our investigation, we have chosen three properties including cross-section shape, axis shape, and cross-section size variation (Dickinson et al. 1990). The values of these properties give rise to a set of ten primitives (a subset of Biederman's geons), modeled using Pentland's SuperSketch 3-D modeling tool (Pentland 1986), and illustrated in Figure 1(a). To construct objects, the primitives are attached to one another with the restriction that any junction of two primitives involves exactly one distinct surface from each primitive.

3.1.2 Viewer-Centered Models. To recover the volumetric primitives from an image, we need some way of modeling their appearance in the image. Traditional aspect graph representations of 3-D objects model an entire object with a set of aspects, each defining a topologically distinct view of an object in terms of its visible surfaces (Koenderink and van Doorn 1979). Our approach differs in that we use aspects to represent a (typically small) set of volumetric primitives from which each object in our database is constructed, rather than representing an entire object directly. Consequently, our goal is to use aspects to recover the 3-D primitives that make up the object in order to carry out a recognition-by-parts procedure, rather than attempting to use aspects to recognize entire objects. The advantage of this approach is that since the number of qualitatively different primitives is generally small, the number of possible aspects is limited and, more important, *independent* of the number of objects in the database. The disadvantage is that if a primitive is occluded from a given 3-D viewpoint, its projected aspect in the image will also be occluded. Thus we must accommodate the matching of occluded aspects, which we accomplish by use of a hierarchical representation we call the *aspect hierarchy.*

The aspect hierarchy consists of three levels, consisting of the set of *aspects* that model the chosen primitives, the set of component *faces* of the aspects, and the set of *boundary groups* representing all subsets of contours bounding the faces. Figure 1(b) illustrates a portion of the aspect hierarchy. The ambiguous mappings between the levels of the aspect hierarchy are captured in a set of conditional probabilities

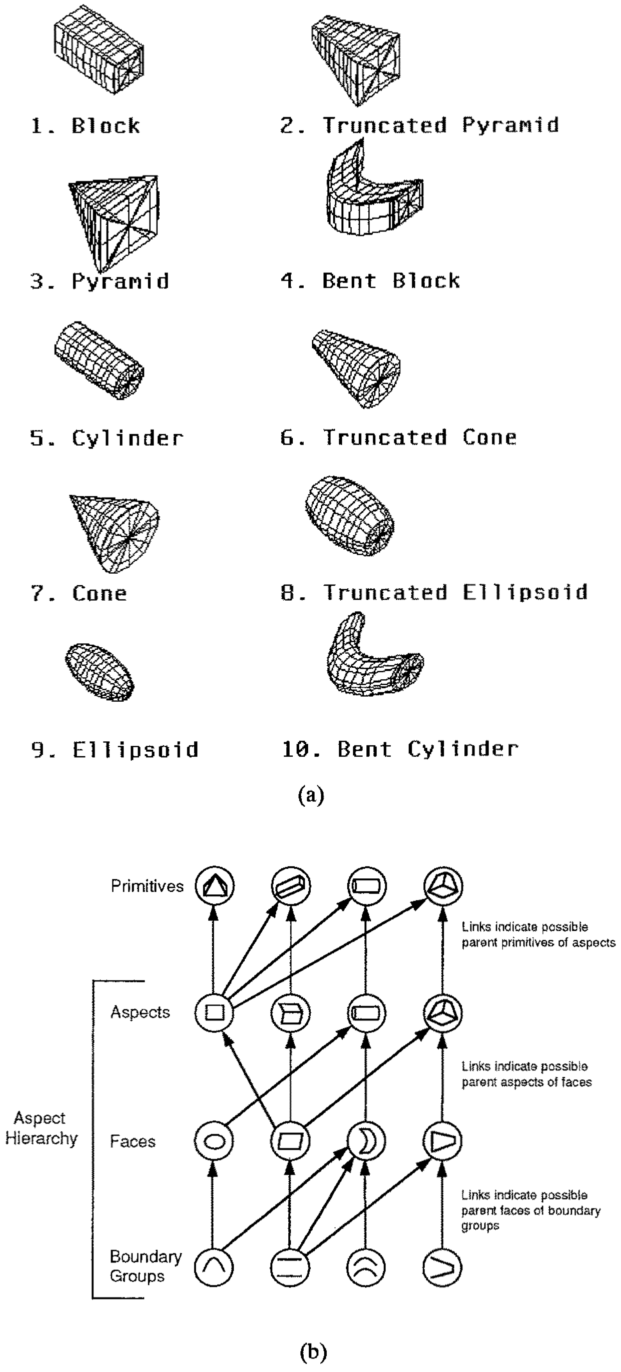


Fig. 1.(a) The ten modeling primitives, (b) the aspect hierarchy.

(Dickinson et al. 1990, 1992b), mapping boundary groups to faces, faces to aspects, and aspects to primitives. These conditional probabilities result from a statistical analysis of a set of images approximating the set of *all* views of *all* the primitives.

3.2 Quantitative Shape Modeling

In this section we first briefly review the general formulation of deformable models; further detail can be found in (Terzopoulos and Metaxas 1991, Metaxas and Terzopoulos 1991). We then extend the formulation to the case of orthographic, perspective, and stereo projections.

3.2.1 Geometry. Geometrically, the models developed in this paper are closed surfaces in space whose intrinsic (material) coordinates are $\mathrm{u} = (u, v)$, defined on a domain Ω. The positions of points on the model relative to an inertial frame of reference Φ in space are given by a vector-valued, time varying function of u:

$$\mathbf{x}(\mathrm{u}, t) = (x_1(\mathrm{u}, t), x_2(\mathrm{u}, t), x_3(\mathrm{u}, t))^\top, \quad (1)$$

where $^\top$ is the transpose operator. We set up a noninertial, model-centered reference frame ϕ, as shown in Fig. 2, and express these positions as:

$$\mathbf{x} = \mathbf{c} + \mathbf{R}\mathbf{p}, \quad (2)$$

where $\mathbf{c}(t)$ is the origin of ϕ at the center of the model, with the orientation of ϕ given by the rotation matrix $\mathbf{R}(t)$. Thus, $\mathbf{p}(\mathrm{u}, t)$ denotes the canonical positions of points on the model relative to the model frame. In Terzopoulos and Metaxas (1991), we further express $\mathbf{p}$ as the sum of a reference shape $\mathrm{s}(\mathrm{u}, t)$ (global deformation) and a displacement function $\mathbf{d}(\mathrm{u}, t)$ (local deformation):

$$\mathbf{p} = \mathbf{s} + \mathbf{d}. \quad (3)$$

However, since computing 3-D local deformations from 2-D contour data is underconstrained, we will consider only global deformations, $\mathbf{s}$, since they are sufficient to represent the shapes of the ten volumetric primitives shown in Figure 1(a). Thus, we have:

$$\mathbf{p} = \mathbf{s}. \quad (4)$$

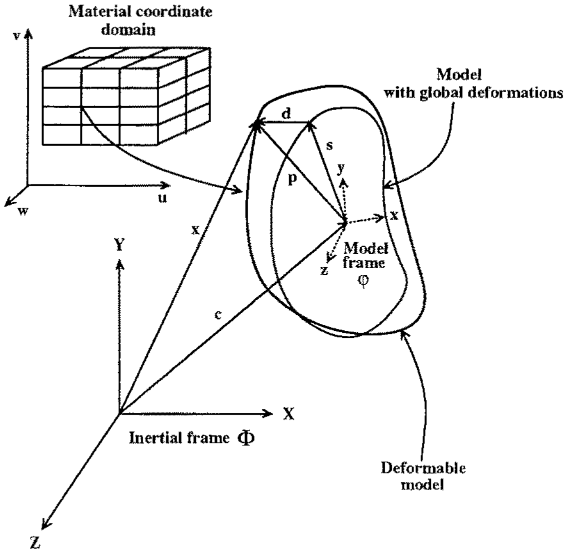


Fig. 2. Geometry of deformable model.

The ensuing formulation can be carried out for any reference shape given as a parameterized function of u. Based on the shapes we want to recover, we first consider the case of superquadric ellipsoids (Barr 1981), which are given by the following formula:

$$\mathbf{e} = a \begin{pmatrix} a_1 C_u^{\epsilon_1} C_v^{\epsilon_2} \\ a_2 C_u^{\epsilon_1} S_v^{\epsilon_2} \\ a_3 S_u^{\epsilon_1} \end{pmatrix}, \quad (5)$$

where $-\pi/2 \leq u \leq \pi/2$ and $-\pi \leq v < \pi$, and where $S_w^{\epsilon} = \mathrm{sgn}(\sin w)|\sin w|^{\epsilon}$ and $C_w^{\epsilon} = \mathrm{sgn}(\cos w)|\cos w|^{\epsilon}$, respectively. Here, $a \geq 0$ is a scale parameter, $0 \leq a_1, a_2, a_3 \leq 1$ are aspect ratio parameters, and $\epsilon_1, \epsilon_2 \geq 0$ are "squareness" parameters.

We then combine linear tapering along principal axes 1 and 2, and bending along principal axis 3 of the superquadric $\mathbf{e}$ into a single parameterized deformation $\mathbf{T}$, and express the reference shape as:

$$\begin{aligned} \mathbf{s} &= \mathbf{T}(\mathbf{e}, t_1, t_2, b_1, b_2, b_3) \\ &= \begin{pmatrix} \left(\frac{t_1 e_3}{a a_3 w} + 1\right) e_1 + b_1 \cos\left(\frac{e_3 + b_2}{a a_3 w} \pi b_3\right) \\ \left(\frac{t_2 e_3}{a a_3 w} + 1\right) e_2 \\ e_3 \end{pmatrix}, \quad (6) \end{aligned}$$

where $-1 \leq t_1, t_2 \leq 1$ are the tapering parameters in principal axes 1 and 2, respectively; where b_1 defines the magnitude of the bending and can be positive or negative; $-1 \leq b_2 \leq 1$ defines the location on axis 3 where bending is applied; and $0 < b_3 \leq 1$ defines the region of influence of bending. Our method for incorporating global deformations is not restricted to only tapering and bending deformations. Any other deformation that can be expressed as a continuous parameterized function can be incorporated as our global deformation in a similar way.

We collect the parameters in $\mathbf{s}$ into the parameter vector:

$$\mathbf{q}_s = (a, a_1, a_2, a_3, \epsilon_1, \epsilon_2, t_1, t_2, b_1, b_2, b_3)^\top. \tag{7}$$

The above global deformation parameters are adequate for quantitatively describing the ten modeling primitives shown in Figure 1(a).

3.2.2 Kinematics. The velocity of points on the model is given by:

$$\begin{aligned}\dot{\mathbf{x}} &= \dot{\mathbf{c}} + \dot{\mathbf{R}}\mathbf{p} + \mathbf{R}\dot{\mathbf{p}} \\ &= \dot{\mathbf{c}} + \mathbf{B}\dot{\theta} + \mathbf{R}\dot{\mathbf{s}},\end{aligned} \tag{8}$$

where $\theta = (\ldots, \theta_i, \ldots)^\top$ is the vector of rotational coordinates of the model, and $\mathbf{B} = [\ldots \partial(\mathbf{Rp})/\partial\theta_i \ldots]$. Furthermore,

$$\dot{\mathbf{s}} = \left[\frac{\partial \mathbf{s}}{\partial \mathbf{q}_s}\right] \dot{\mathbf{q}}_s = \mathbf{J}\dot{\mathbf{q}}_s, \tag{9}$$

where $\mathbf{J}$ is the Jacobian of the deformable superquadric model with respect to the global degrees of freedom.

Defining $r = \frac{e_3+b_2}{aa_3}\pi b_3$, the Jacobian matrix $\mathbf{J}$ is a 3×11 matrix whose non-zero entries are:

$$\begin{aligned}
\mathbf{J}_{11} &= (t_1 S_u{}^{\epsilon_1} + 1) a_1 C_u{}^{\epsilon_1} C_v{}^{\epsilon_2} + \frac{b_1 b_2 b_3}{a^2 a_3}\pi \sin(r) \\
\mathbf{J}_{21} &= (t_2 S_u{}^{\epsilon_1} + 1) a_2 C_u{}^{\epsilon_1} S_v{}^{\epsilon_2} \\
\mathbf{J}_{31} &= a_3 S_u{}^{\epsilon_1} \\
\mathbf{J}_{12} &= (t_1 S_u{}^{\epsilon_1} + 1) a C_u{}^{\epsilon_1} C_v{}^{\epsilon_2} \\
\mathbf{J}_{23} &= (t_2 S_u{}^{\epsilon_1} + 1) a C_u{}^{\epsilon_1} S_v{}^{\epsilon_2} \\
\mathbf{J}_{14} &= \frac{b_1 b_2 b_3}{a {a_3}^2}\pi \sin(r) \\
\mathbf{J}_{34} &= a S_u{}^{\epsilon_1}
\end{aligned}$$

$$\begin{aligned}
\mathbf{J}_{15} &= t_1 \ln(|\sin u|) S_u{}^{\epsilon_1} a a_1 C_u{}^{\epsilon_1} C_v{}^{\epsilon_2} \\
&\quad + (t_1 S_u{}^{\epsilon_1} + 1) a a_1 \ln(|\cos u|) C_u{}^{\epsilon_1} C_v{}^{\epsilon_2} \\
&\quad - b_1 b_3 \pi \ln(|\sin u|) S_u{}^{\epsilon_1} \sin(r) \\
\mathbf{J}_{25} &= t_2 \ln(|\sin u|) S_u{}^{\epsilon_1} a a_2 C_u{}^{\epsilon_1} S_v{}^{\epsilon_2} \\
&\quad + (t_2 S_u{}^{\epsilon_1} + 1) a a_2 \ln(|\cos u|) C_u{}^{\epsilon_1} S_v{}^{\epsilon_2} \\
\mathbf{J}_{35} &= a a_3 \ln(|\sin u|) S_u{}^{\epsilon_1} \\
\mathbf{J}_{16} &= (t_1 S_u{}^{\epsilon_1} + 1) a a_1 \ln(|\cos v|) C_u{}^{\epsilon_1} C_v{}^{\epsilon_2} \\
\mathbf{J}_{26} &= (t_2 S_u{}^{\epsilon_1} + 1) a a_2 \ln(|\sin v|) C_u{}^{\epsilon_1} S_v{}^{\epsilon_2} \\
\mathbf{J}_{17} &= S_u{}^{\epsilon_1} a a_1 C_u{}^{\epsilon_1} C_v{}^{\epsilon_2} \\
\mathbf{J}_{28} &= S_u{}^{\epsilon_1} a a_2 C_u{}^{\epsilon_1} S_v{}^{\epsilon_2} \\
\mathbf{J}_{19} &= \cos(r) \\
\mathbf{J}_{1\,10} &= -\frac{b_1 b_3}{a a_3}\pi \sin(r) \\
\mathbf{J}_{1\,11} &= -b_1 \pi \sin(r)\ r,
\end{aligned} \tag{10}$$

where $S_\theta{}^\epsilon = \operatorname{sgn}(\sin\theta)\ |\sin\theta|^\epsilon$ and $C_\theta{}^\epsilon = \operatorname{sgn}(\cos\theta)\ |\cos\theta|^\epsilon$.

We can therefore write:

$$\dot{\mathbf{x}} = [\mathbf{I}\ \mathbf{B}\ \mathbf{RJ}]\dot{\mathbf{q}} = \mathbf{L}\dot{\mathbf{q}}, \tag{11}$$

where $\mathbf{L}$ is the Jacobian of the superquadric model, $\mathbf{q} = (\mathbf{q}_c^\top, \mathbf{q}_\theta^\top, \mathbf{q}_s^\top)^\top$, with $\mathbf{q}_c = \mathbf{c}$ and $\mathbf{q}_\theta = \theta$.

3.2.3 Dynamics. When fitting the model to visual data, our goal is to recover $\mathbf{q}$, the vector of degrees of freedom of the model. The components $\mathbf{q}_c$ and $\mathbf{q}_\theta$ are the global rigid motion coordinates and $\mathbf{q}_s$ are the global deformation coordinates. Our approach carries out the coordinate fitting procedure in a physically-based way. We make our model dynamic in $\mathbf{q}$ by introducing mass, damping, and a deformation strain energy. This allows us, through the apparatus of Lagrangian dynamics, to arrive at a set of equations of motion governing the behavior of our model under the action of externally applied forces.

The Lagrange equations of motion take the form (Terzopoulos and Metaxas 1991):

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{D}\dot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{g}_q + \mathbf{f}_q, \tag{12}$$

where $\mathbf{M}$, $\mathbf{D}$, and $\mathbf{K}$ are the mass, damping, and stiffness matrices, respectively, where $\mathbf{g}_q$ are inertial (centrifugal and Coriolis) forces arising from the dynamic coupling between the local and global degrees of freedom, and where $\mathbf{f}_q(\mathrm{u}, t)$ are the generalized external forces associated with the degrees of freedom of the model. The generalized external forces will be discussed in detail in Section 4.2.2.

3.2.4 Orthographic Projection. In the case of orthographic projection, the points on the model $\mathbf{x} = (x, y, z)$ project to the image points x_p and y_p as follows:

$$\begin{aligned} x_p &= x \\ y_p &= y. \end{aligned} \tag{13}$$

By taking the derivative of the above equation (13) with respect to time, we arrive at the following formulas:

$$\begin{aligned} \dot{x}_p &= \dot{x} \\ \dot{y}_p &= \dot{y}. \end{aligned} \tag{14}$$

Rewriting (14) in matrix form and using (11), we arrive at the following matrix equations:

$$\begin{bmatrix} \dot{x}_p \\ \dot{y}_p \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} \tag{15}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{L}\dot{\mathbf{q}}. \tag{16}$$

If we rewrite (16) in compact form, we get

$$\begin{bmatrix} \dot{x}_p \\ \dot{y}_p \end{bmatrix} = \mathbf{L}_o \dot{\mathbf{q}}, \tag{17}$$

where

$$\mathbf{L}_o = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{L}. \tag{18}$$

3.2.5 Perspective Projection. In the case of perspective projection, points on the model $\mathbf{x} = (x, y, z)$ project into image points, x_p and y_p, based on the formula:

$$\begin{aligned} x_p &= \frac{x}{z} f \\ y_p &= \frac{y}{z} f, \end{aligned} \tag{19}$$

where f is the focal length.

By taking the derivative of the above equation (19) with respect to time, we arrive at the following formulas:

$$\begin{aligned} \dot{x}_p &= \dot{x}\frac{f}{z} - \frac{x}{z^2} f \dot{z} \\ \dot{y}_p &= \dot{y}\frac{f}{z} - \frac{y}{z^2} f \dot{z}. \end{aligned} \tag{20}$$

Rewriting (20) in matrix form and using (11), we arrive at the following matrix equations

$$\begin{bmatrix} \dot{x}_p \\ \dot{y}_p \end{bmatrix} = \begin{bmatrix} f/z & 0 & -x/z^2 f \\ 0 & f/z & -y/z^2 f \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} \tag{21}$$

$$= \begin{bmatrix} f/z & 0 & -x/z^2 f \\ 0 & f/z & -y/z^2 f \end{bmatrix} \mathbf{L}\dot{\mathbf{q}}. \tag{22}$$

If we rewrite (22) in compact form, we get

$$\begin{bmatrix} \dot{x}_p \\ \dot{y}_p \end{bmatrix} = \mathbf{L}_p \dot{\mathbf{q}}, \tag{23}$$

where

$$\mathbf{L}_p = \begin{bmatrix} f/z & 0 & -x/z^2 f \\ 0 & f/z & -y/z^2 f \end{bmatrix} \mathbf{L}. \tag{24}$$

The above two Jacobian matrices, $\mathbf{L}_o$ and $\mathbf{L}_p$, will be used in the calculation of the generalized external forces $\mathbf{f}_q$ from two-dimensional external forces $\mathbf{f}$ that the data exert on the model.

3.2.6 Stereo Projection. In the case of stereo projection, we assume two parallel cameras, each under perspective projection, resulting in two images, L and R. The model points $\mathbf{x}$ project on each of the images based on (19) and the corresponding Jacobian matrices $\mathbf{L}_{PL}$ and $\mathbf{L}_{PR}$ are calculated using (24).

To recover the exact location of the model frame $\mathbf{c}$, we apply the following procedure:

- We first independently fit the model to the left and right image data. This results in two model instances, m_L and m_R, one per image, having the same scale.
- Choosing one of the images, say R, we project the locations $\mathbf{c}_L$ and $\mathbf{c}_R$, of the left and right model frames of the two model instances m_L and m_R, into R. Let the locations of the

projected model centers be $\mathbf{c}_{LI}$ and $\mathbf{c}_{RI}$, respectively.

- We then map the difference in the x coordinates of $\mathbf{c}_{LI}$ and $\mathbf{c}_{RI}$ into a force that modifies $\mathbf{c}_L$ and $\mathbf{c}_R$ in the direction of $\mathbf{c}_L$ and $\mathbf{c}_R$, respectively, according to the following formula:
$$\dot{\mathbf{c}}_k = s|c_{LIx} - c_{RIx}|\frac{\mathbf{c}_k}{\|\mathbf{c}_k\|} \tag{25}$$
where $k = L$ or $k = R$, $s = 1$ if $c_{LIx} < c_{RIx}$, and $s = -1$ otherwise.
- Once $\mathbf{c}_{LI} = \mathbf{c}_{RI}$, we first sum the forces that the left and right image data exert on the model. From their sum, we then compute the generalized force f_{q_a} that corresponds to the scaling parameter a (5), and using (12), we modify a.

4 Shape Recovery

4.1 Qualitative Shape Recovery

Primitive recovery consists of the following three steps, resulting in a graph representation of the image in which nodes represent recovered 3-D primitives, and arcs represent hypothesized connections between the primitives; details of the complete recovery process, including algorithms to handle various segmentation errors, can be found in Dickinson et al. (1992b). In the following subsections, we briefly review the approach to recovering qualitative shape.

4.1.1 Face Recovery. The first step to recovering a set of faces is a region segmentation of the input image. We begin by applying Saint-Marc, Chen, and Medioni's edge-preserving adaptive smoothing filter to the image (Saint-Marc et al. 1991), followed by a morphological gradient operator (Lee et al. 1987). A hysteresis thresholding operation is then applied to produce a binary image from which a set of connected components is extracted. Edge regions are then burned away, resulting in a *region topology graph* in which nodes represent regions and arcs specify region adjacencies.

From a region topology graph, each region is characterized according to the qualitative shapes of its bounding contours. First, the bounding contour of each region is partitioned at curvature extrema using Saint-Marc, Chen, and Medioni's adaptive smoothing curve partitioning technique (Saint-Marc et al. 1991). Next, each bounding contour is classified as straight, convex, or concave, by comparing the contour to a fitted line. Finally, each pair of bounding contours is checked for cotermination, parallelism, or symmetry. The result is a *region boundary graph* representation for a region in which nodes represent bounding contours, and arcs represent pairwise nonaccidental relations between the contours.

Face labeling consists of matching a region boundary graph to the graphs representing the model faces in the aspect hierarchy. Region boundary graphs that exactly match a face in the aspect hierarchy will be given a single label with probability 1.0. For region boundary graphs that do not match due to occlusion, segmentation errors, or errors in computing their graphs, we descend to an analysis at the boundary group level and match subgraphs of the region boundary graph to the graphs representing the boundary groups in the aspect hierarchy. Each subgraph that matches a boundary group generates a set of possible face interpretations (labels), each with a corresponding probability. The result is a *face topology graph* in which each node contains a set of face labels (sorted by decreasing order of probability) associated with a given region.

4.1.2 Aspect Recovery. In an unexpected object recognition domain in which there is no a priori knowledge of scene content, we can formulate the problem of extracting aspects as follows: Given a face topology graph with a set of face hypotheses (labels) at each node (region), find an *aspect covering* of the face topology graph using aspects in the aspect hierarchy, such that no region is left uncovered and each region is covered by only one aspect. Or, more formally: Given an input face topology graph, FTG, partition the nodes (regions) of FTG into disjoint sets, $S_1, S_2, S_3, \ldots, S_k$, such that the graph induced by each set, S_i, is isomorphic to the graph representing some aspect, A_j, from a fixed set of aspects, $A_1, A_2, A_3, \ldots, A_n$.

There is no known polynomial time algorithm to solve this problem (see Dickinson et al. 1992 for a discussion on the problem's computational complexity); however, the conditional probability matrices embedded in the aspect hierarchy provide a powerful constraint that can make the problem tractable. For each face hypothesis (for a given region), we can use the face to aspect mapping to generate the possible *aspect hypotheses* that might encompass that face. At each face, we collect all the aspect hypotheses (corresponding to all face hypotheses) and rank them in decreasing order of probability.

We can now reformulate our bottom-up aspect recovery problem as a search through the space of aspect labelings of the regions (nodes) in the face topology graph. In other words, we wish to choose one aspect hypothesis from the list at each node, such that the instantiated aspects completely cover the face topology graph. For our search through the possible aspect labelings of the face topology graph, we employ Algorithm A (Nilsson 1980) with a heuristic designed to meet three objectives. First, we favor selections of aspects instantiated from higher probability aspect hypotheses. Second, we favor selections whose aspects have fewer occluded faces, since we are more confident of their labels. Finally, we favor those aspects covering more faces in the image; we seek the minimal aspect covering of the face topology graph. Since there may be many labelings which satisfy this constraint, and since we cannot guarantee that a given aspect covering represents a correct interpretation of the scene, we must be able to enumerate, in decreasing order of likelihood, all aspect coverings until the objects in the scene are recognized.

In an expected object recognition domain in which we are searching for a particular object or part, we use the aspect hierarchy as an attention mechanism to focus the search for an aspect at appropriate regions in the image. Moving down the aspect hierarchy, target objects map to target volumes which, in turn, map to target aspect predictions which, in turn, map to target face predictions. Verification of the target aspect prediction occurs at those faces in the face topology graph whose labels match the target face prediction. The scores of the matching faces are used to order the recovery process which attends first to high-quality faces. This attention mechanism has been used to drive an active recognition system which moves the cameras to obtain either a more likely or unambiguous view of an object's part (Dickinson et al. 1993).

4.1.3 Primitive Recovery. In the expected object recognition approach described above, primitive recovery consists of mapping the recovered aspect directly to the target primitive prediction. Primitive recovery for the unexpected object recognition case is more complex. From an *aspect covering* of the regions in the image, a set of primitive labels and their corresponding probabilities is inferred (using the aspect hierarchy) from each aspect. Primitive recovery is formulated as a search through the space of primitive labelings of the aspects in the aspect covering, guided by a heuristic based on the probabilities of the primitive labels. Each solution, or *primitive covering*, found by the search is a valid primitive interpretation of the input image. Encoded in each recovered primitive is the aspect in which it is viewed; the aspect, in turn, encodes the faces that were used in instantiating the aspect, while each face specifies those contours in the image used to instantiate the face.

4.1.4 Stereo Correspondence. In the case of stereo projection, we independently apply the qualitative shape recovery process to the left and right images. The correspondence problem then consists of matching qualitative primitive descriptions in the two images. A pair of volumes represents a correspondence if: (i) the volumes have the same label, (ii) their aspects have the same label, and (iii) the ratio of the vertical intersection of the bounding rectangles of the two volumes to the vertical size of each bounding rectangle exceeds some threshold (epipolar constraint). Intuitively, volumes from the left and right image are said to correspond if they are of the same type, they are viewed in roughly the same orientation, and their vertical disparity is small. Note that this provides only a coarse correspondence; dimensions, orientation, and curvature of the volumes may be disparate. During the independent quantitative shape recovery of the left and right models, additional shape information can be used to prune weak correspondences, providing a coarse-to-fine stereo correspondence scheme.

4.2 Quantitative Shape Recovery

4.2.1 Simplified Numerical Simulation. Equations (12) give the general equations of motion for a dynamic model with deformations. A full implementation and simulation of the general equations would be appropriate for physically-based animation where realistic motion is important (Terzopooulos and Witkin 1988). However, in computer vision and geometric design applications involving the fitting of models to data, models governed by simplified equations of motion suffice, as the experiments in Section 5 will demonstrate.

We can simplify the equations while preserving useful dynamics by setting the mass density $\mu(\mathrm{u})$ to zero to obtain (Terzopoulos and Metaxas 1991):

$$\mathbf{D}\dot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{f}_q. \tag{26}$$

These equations yield a model which has no inertia and comes to rest as soon as all the applied forces vanish or equilibrate. Equation (26) is discretized in material coordinates u using nodal finite element basis functions. We carry out the discretization by tessellating the surface of the model into linear triangular elements.

The formulation of our model yields numerically stable equations of motion that may be integrated forward through time using explicit procedures. For fast interactive response, we employ a first-order Euler method to integrate (26). The Euler procedure updates the degrees of freedom $\mathbf{q}$ of the model at time $t + \Delta t$ according to the formula:

$$\mathbf{q}^{(t+\Delta t)} = \mathbf{q}^{(t)} + \Delta t\ (\mathbf{D}^{(t)})^{-1}(\mathbf{f}_q^{(t)} - \mathbf{K}\mathbf{q}^{(t)}), \tag{27}$$

where Δt is the time step size.

Taking time steps in $\mathbf{q}$ is straightforward, but the rotation component $\mathbf{q}_\theta$ is a little delicate. We represent $\mathbf{q}_\theta$ using quaternions. Updating quaternions at each time step is easier than directly updating a rotation matrix and ensuring that it remains orthogonal.

A quaternion $[s, \mathbf{v}]$ with unit magnitude, $\|[s, \mathbf{v}]\| = s^2 + \mathbf{v}^\top\mathbf{v} = 1$, specifies a rotation of the model from its reference position through an angle $\theta = 2\cos^{-1} s$ around an axis aligned with vector $\mathbf{v} = [v_1, v_2, v_3]^\top$. The rotation matrix corresponding to $[s, \mathbf{v}]$ is:

$$\mathbf{R} = \begin{bmatrix} 1 - 2(v_2^2 + v_3^2) & 2(v_1v_2 - sv_3) & 2(v_1v_3 + sv_2) \\ 2(v_1v_2 + sv_3) & 1 - 2(v_1^2 + v_3^2) & 2(v_2v_3 - sv_1) \\ 2(v_1v_3 - sv_2) & 2(v_2v_3 + sv_1) & 1 - 2(v_1^2 + v_2^2) \end{bmatrix}. \tag{28}$$

To obtain $\mathbf{q}_\theta$ from (27), we use the generalized torque at time t given by $\mathbf{f}_\theta^\top = \int \mathbf{f}^\top\mathbf{B}du$, with $\mathbf{B}$ (Shabana 1989; Terzopoulos and Metaxas 1991):

$$\mathbf{B}(u) = -\mathbf{R}\ \tilde{\mathbf{p}}(u)\ \mathbf{G}, \tag{29}$$

where $\mathbf{R}$ represents the rotation matrix at time t, where $\tilde{\mathbf{p}}(u)$ is the dual 3×3 matrix of the position vector $\mathbf{p}(u) = (p_1, p_2, p_3)^\top$ (see (4)) defined as:

$$\tilde{\mathbf{p}}(u) = \begin{bmatrix} 0 & -p_3 & p_2 \\ p_3 & 0 & -p_1 \\ -p_2 & p_1 & 0 \end{bmatrix} \tag{30}$$

and where $\mathbf{G}$ is a 3×4 matrix whose definition is based on the value of the quaternion $\mathbf{q}_\theta = [s, \mathbf{v}]$ representing the rotation at time t:

$$\mathbf{G} = 2\begin{bmatrix} -v_1 & s & v_3 & -v_2 \\ -v_2 & -v_3 & s & v_1 \\ -v_3 & v_2 & -v_1 & s \end{bmatrix}. \tag{31}$$

4.2.2 Applied Forces. In the dynamic model fitting process, the data are transformed into an externally applied force distribution $\mathbf{f}(\mathrm{u}, t)$. We convert the external forces to generalized forces $\mathbf{f}_q$ which act on the generalized coordinates of the model (Terzopoulos and Metaxas 1991). We apply forces to the models based on differences between the model's projection in the image and the image data. Each of these forces corresponds to the appropriate generalized coordinate that has to be adapted so that the model fits the data. Given that our vocabulary of primitives is limited, we devise a systematic way of computing the generalized forces for each primitive. The computation depends on the influence of particular parts of the projected image on the model degrees of freedom. Such parts correspond to the image faces (grouped to form an aspect) provided by the qualitative shape extraction. In the case of occluded primitives, resulting in both occluded aspects and occluded faces, only those portions (boundary groups) of the faces used to define the faces exert external forces on the models.

For each of the three projection models, we compute the generalized forces $\mathbf{f}_q$ from 2-D image forces f, using the following formula:

$$\mathbf{f}_q^\top = \int \mathbf{f}^\top \mathbf{L}_k \; d\mathrm{u} = \left(\mathbf{f}_{q_c}^\top, \mathbf{f}_{q_\theta}^\top, \mathbf{f}_{q_s}^\top\right), \qquad (32)$$

where $k = o$ or $k = p$, depending on whether we assume orthographic or perspective projection, respectively. For orthographic projection, we assign forces from image data points to points on the model that lie on a particular region of the model defined by the qualitative shape recovery. For the case of perspective projection, we assign forces from image data points to points on the model that, in addition to satisfying the above property, are near occluding boundaries, thus satisfying the following formula:

$$|\mathbf{i} \cdot \mathbf{n}| < \tau, \qquad (33)$$

where $\mathbf{n}$ is the unit normal at any model point, $\mathbf{i}$ is the unit vector from the focal point to a point on the model, and τ is a small threshold.

4.2.3 Model Initialization. One of the major limitations of previous deformable model fitting approaches is their dependence on model initialization and prior segmentation (Terzopoulos et al. 1988; Terzopoulos and Metaxas 1991, Pentland and Sclaroff 1991). Using the qualitative shape recovery process as a front end, we first segment the image into parts, and for each part, we identify the relevant non-occluded contour data belonging to the part. In addition, the extracted qualitative primitives explicitly define a mapping between the image faces in their projected aspects and the 3-D surfaces on the quantitative models. Finally, although the initial model can be specified at any position and orientation, the aspect that a primitive encodes defines a qualitative orientation that can be exploited to speed up the model fitting process. Sensitivity of the fitting process to model initialization is also overcome by independently solving for the degrees of freedom of the model. By allowing each face in an aspect to exert forces on only one model degree of freedom at a time, we remove local minima from the fitting process and ensure correct convergence of the model.

5 Experiments

To illustrate the shape recovery approach, consider the real image of a toy table lamp, as shown in Figure 3; the results of the bottom-up (unexpected) qualitative shape recovery algorithm are shown in Figure 4. At the top, the image window contains the contours extracted from the image, along with the face numbers. To the left is a window describing the recovered primitives (primitive covering). The mnemonics, PN, PL, and PP, refer to primitive number (simply an enumeration of the primitives in the covering), primitive label (see Figure 1(a)), and primitive

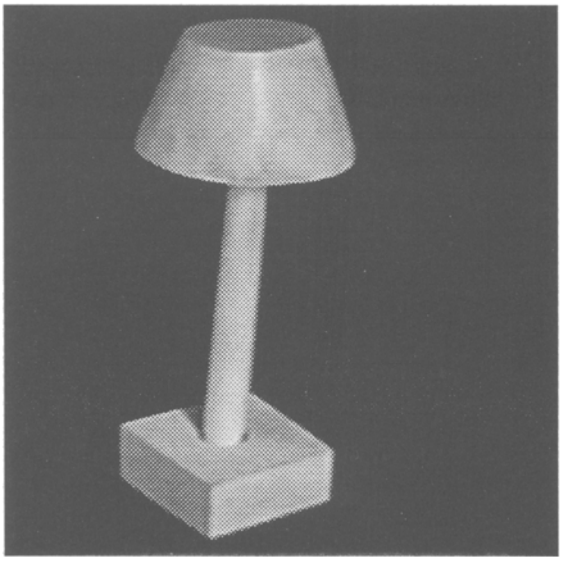
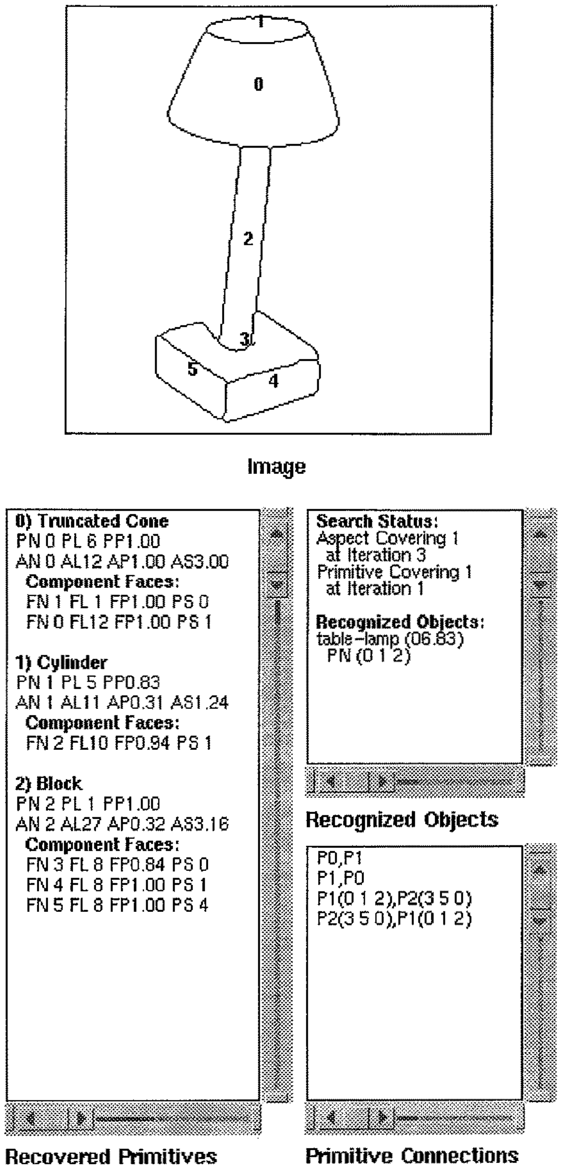


Fig. 3. Image of a table lamp (256 × 256).



```
0) Truncated Cone
PN 0 PL 6 PP1.00
AN 0 AL12 AP1.00 AS3.00
  Component Faces:
  FN 1 FL 1 FP1.00 PS 0
  FN 0 FL12 FP1.00 PS 1

1) Cylinder
PN 1 PL 5 PP0.83
AN 1 AL11 AP0.31 AS1.24
  Component Faces:
  FN 2 FL10 FP0.94 PS 1

2) Block
PN 2 PL 1 PP1.00
AN 2 AL27 AP0.32 AS3.16
  Component Faces:
  FN 3 FL 8 FP0.84 PS 0
  FN 4 FL 8 FP1.00 PS 1
  FN 5 FL 8 FP1.00 PS 4
```

```
Search Status:
Aspect Covering 1
  at Iteration 3
Primitive Covering 1
  at Iteration 1

Recognized Objects:
table-lamp (06.83)
  PN (0 1 2)
```

```
P0,P1
P1,P0
P1(0 1 2),P2(3 5 0)
P2(3 5 0),P1(0 1 2)
```

Fig. 4. Recovered qualitative primitives.

probability, respectively. The mnemonics AN, AL, AP, and AS refer to the aspect number (an enumeration), aspect label (see Dickinson et al. 1992b), aspect probability, and aspect score (how well aspect was verified), respectively. The mnemonics FN, FL, FP, and PS refer to face number (in image window), face label (see Dickinson et al. 1992b), face probability, and corresponding primitive attachment surface (see Dickinson et al. 1992b), respectively, for each component face of the aspect.

To illustrate the fitting stage, consider the contours belonging to the lamp shade (truncated cone). Having determined during the qualitative shape recovery stage that we are trying to fit a deformable superquadric to a truncated cone, we can immediately fix some of the parameters in the model. In addition, the qualitative shape recovery stage provides us with a mapping between faces in the image and physical surfaces on the model. For example, we know that the elliptical face (FN 1) maps to the top of the truncated cone, while the body face (FN 0) maps to the side of the truncated cone. For the case of the truncated cone, we will begin with a cylinder model (superquad) and will compute the forces that will deform the cylinder into the truncated cone appearing in the image. Assuming an orthographic projection, and that the x and y dimensions are equal, we compute the following forces:

1. The cylinder is initially oriented with its z axis orthogonal to the image plane. The first step involves computing the centroid of the elliptical image face (known to correspond to the top of the cylinder). The distance between the centroid and the projected center of the cylinder top is converted to a force which translates the model cylinder. Figure 5(a) shows the image contours corresponding to the lamp shade and the cylinder following application of this force. Figure 5(b) shows a different view of the image plane, providing a better view of the model cylinder.
2. The distance between the two image points corresponding to the extrema of the principal axis of the elliptical image face and two points that lie on a diameter of the top of the cylinder is converted to a force affecting the x and y dimensions with respect to the model cylinder. Figures 5(c) and 5(d) show the image and the cylinder following application of this force.
3. The distance between the projected model contour corresponding to the top of the cylinder and the elliptical image face corresponds

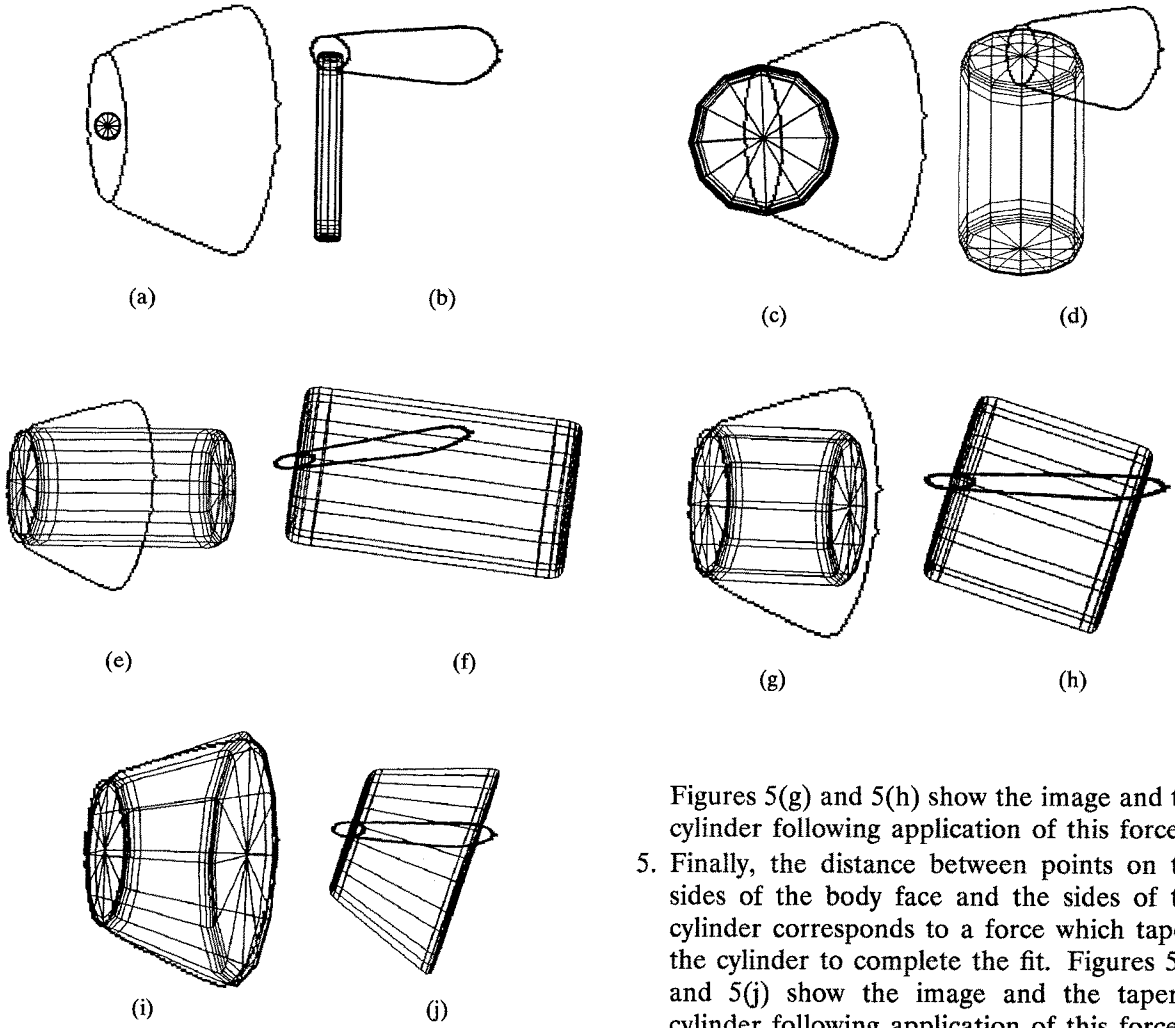


Fig. 5. Quantitative shape recovery for lamp shade.

to a force affecting the orientation of the cylinder. Figures 5(e) and 5(f) show the image and the cylinder following application of this force. This concludes the application of forces arising from the elliptical image face, i.e., top of the truncated cone.

4. Next, we focus on the image face corresponding to the body of the truncated cone to complete the fitting process. The distance between the points along the bottom rim of the body face and the projected bottom rim of the cylinder corresponds to a force affecting the length of the cylinder in the z direction. Figures 5(g) and 5(h) show the image and the cylinder following application of this force.
5. Finally, the distance between points on the sides of the body face and the sides of the cylinder corresponds to a force which tapers the cylinder to complete the fit. Figures 5(i) and 5(j) show the image and the tapered cylinder following application of this force.

Figure 6 shows two views of the initial models for the lamp stem and base, while Figure 7 shows two views of the results of fitting all three parts of the table lamp. Note that with an orthographic projection, we must choose an arbitrary depth for each part; in this case, the models were all initialized with the same depth.

For the case of perspective projection, we apply our top-down (expected) shape recovery technique to the image in Figure 8. A top down search for the best three instances of a qualitative block primitive yields the three primitives shown in Figure 9 (ordered bottom to top by score). Note that due to a large shadow edge that resulted in the undersegmented region 12

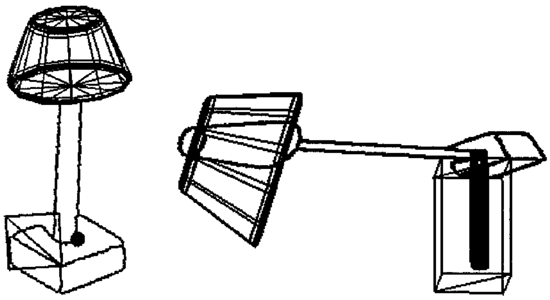


Fig. 6. Initialization of lamp stem and base models.

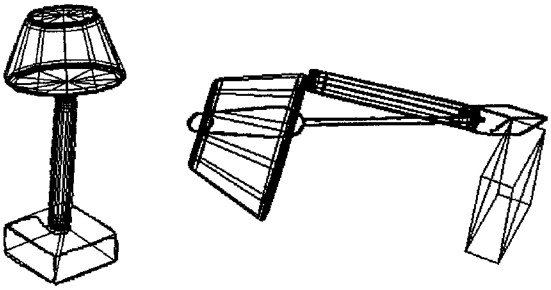


Fig. 7. Final recovery of table lamp (note that depth information is lost in orthographic projection.).

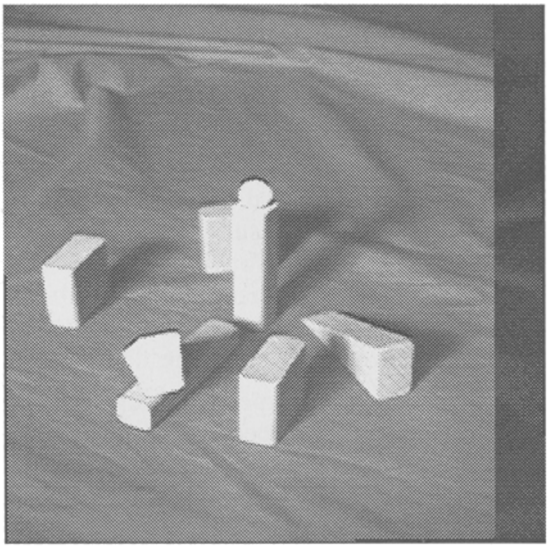
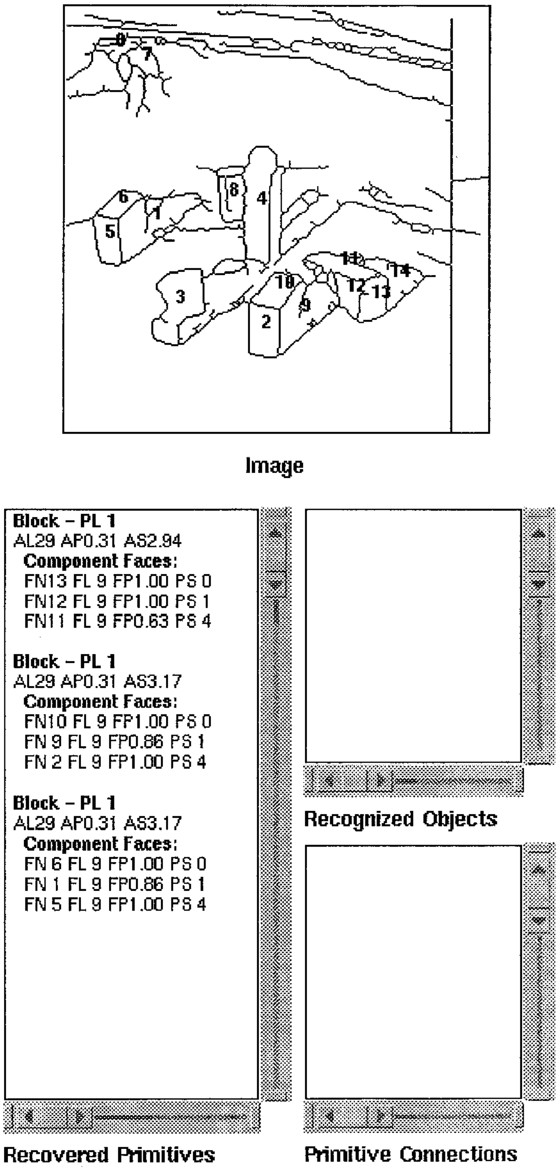


Fig. 8. Image of blocks on a table.



```
Block - PL 1
AL29 AP0.31 AS2.94
  Component Faces:
  FN13 FL 9 FP1.00 PS 0
  FN12 FL 9 FP1.00 PS 1
  FN11 FL 9 FP0.63 PS 4

Block - PL 1
AL29 AP0.31 AS3.17
  Component Faces:
  FN10 FL 9 FP1.00 PS 0
  FN 9 FL 9 FP0.86 PS 1
  FN 2 FL 9 FP1.00 PS 4

Block - PL 1
AL29 AP0.31 AS3.17
  Component Faces:
  FN 6 FL 9 FP1.00 PS 0
  FN 1 FL 9 FP0.86 PS 1
  FN 5 FL 9 FP1.00 PS 4
```

Fig. 9. The best three instances of a qualitative block.

on the triangular face of the wedge, the shape was misclassified as a block since region 12 was classified as having opposites sides parallel. If we apply the quantitative recovery process to these three blocks, we obtain the models depicted in Figures 10, 11, and 12.

Next, we apply the top-down shape recov-

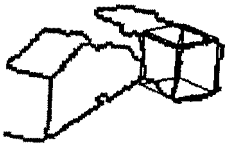


Fig. 10. Model fitted to first block.

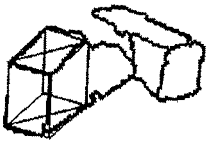


Fig. 11. Model fitted to second block.

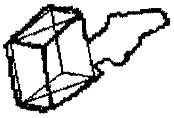


Fig. 12. Model fitted to third block.

ery technique to the stereo pair of an isolated cylinder shown in Figure 13; the system is instructed to search for instances of a cylinder. The results of the qualitative shape recovery are shown in Figure 14, while the initialization of the independent fitting of the model to the left and right images is shown in Figure 15. Following the scaling step, the projection of the final model into the two images is shown in Figure 16. In another example, we apply the top-down object recognition algorithm to the stereo pair shown in Figure 17; the system is instructed to search for high-scoring instances of the block volume, i.e., unoccluded instances appearing as high-probability aspects. Two corresponding pairs were found in each image and are highlighted in Figure 17. Volume score thresholds were set high so that volumes appearing in only the most probable aspect and with little or no occlusion were accepted. Although the top block on the two-block stack to the right was recovered by the algorithm, it was rejected due to the fact that, due to region undersegmentation, one of its faces was merged with a face from the block below, resulting in a lower score. For the smaller block, Figure 18 captures the stage in

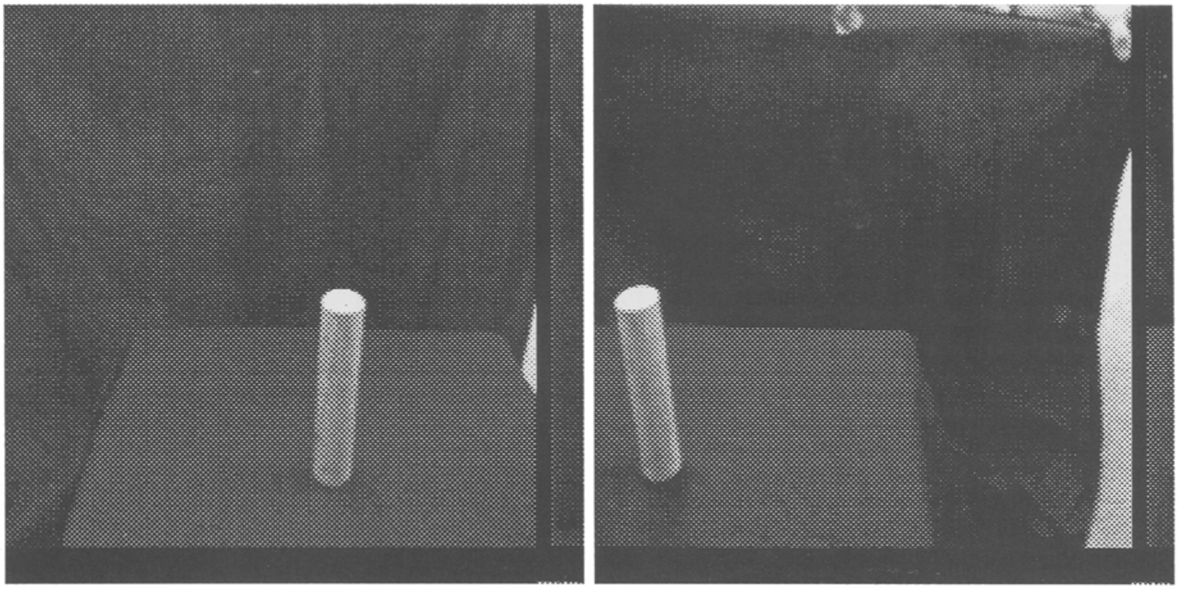
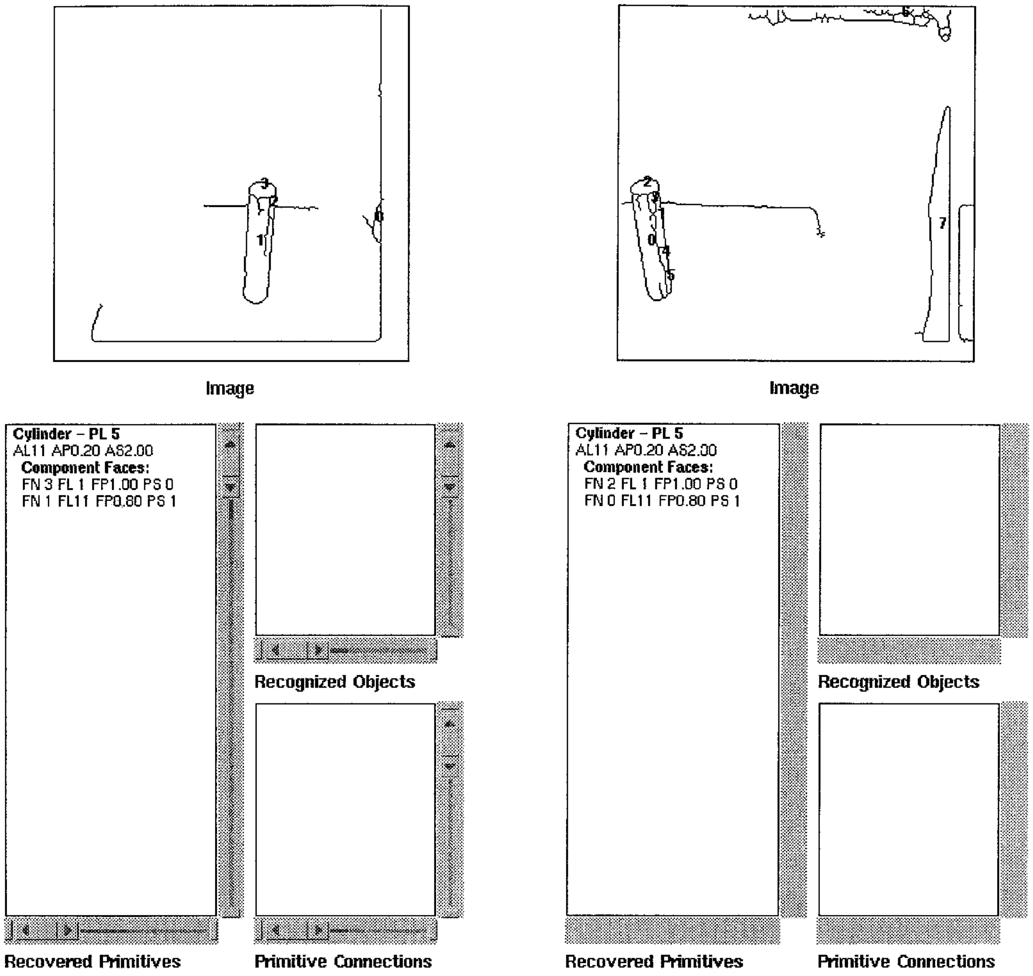


Fig. 13. Left and right stereo images of a cylinder.



```
Cylinder - PL 5
AL11 AP0.20 AS2.00
 Component Faces:
 FN 3 FL 1 FP1.00 PS 0
 FN 1 FL11 FP0.80 PS 1
```

```
Cylinder - PL 5
AL11 AP0.20 AS2.00
 Component Faces:
 FN 2 FL 1 FP1.00 PS 0
 FN 0 FL11 FP0.80 PS 1
```

Fig. 14. Qualitative shape recovery of left and right images.

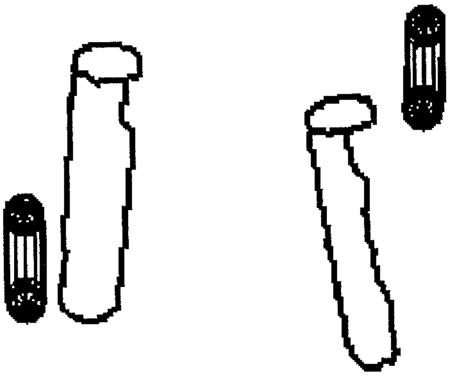


Fig. 15. Model initialization for quantitative shape recovery of left and right images.

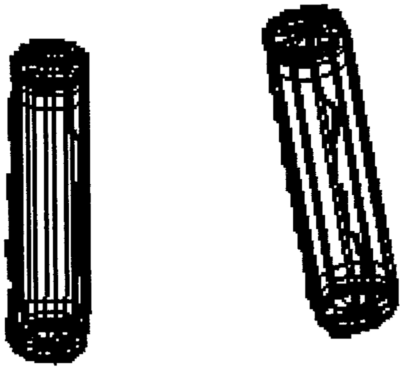


Fig. 16. Unifying the left and right models to determine scale and depth.

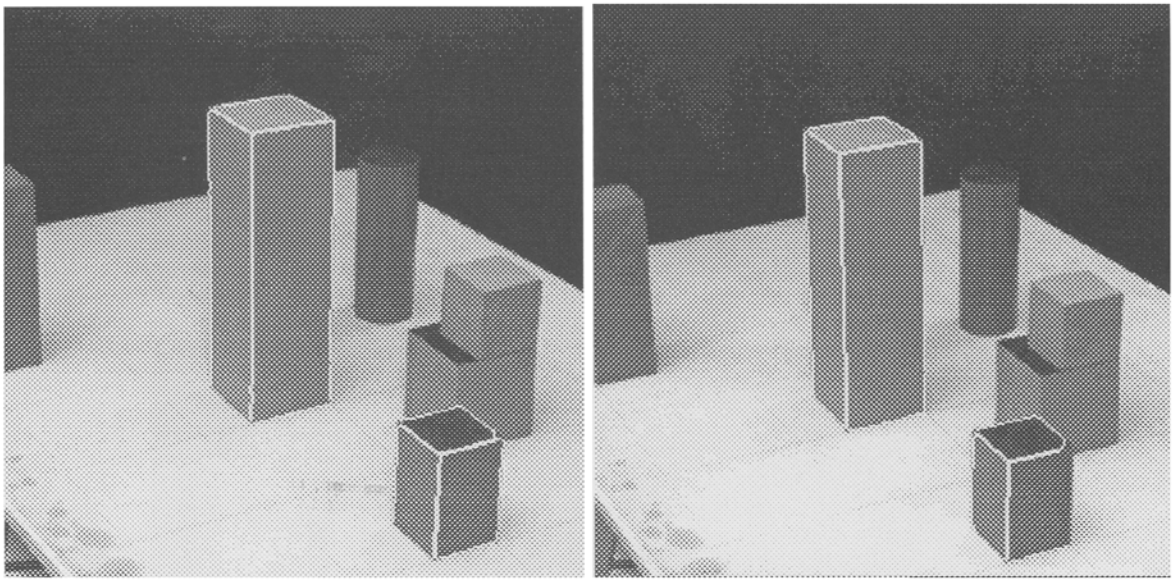


Fig. 17. Left and right stereo images of a cluttered table with corresponding recovered blocks highlighted.

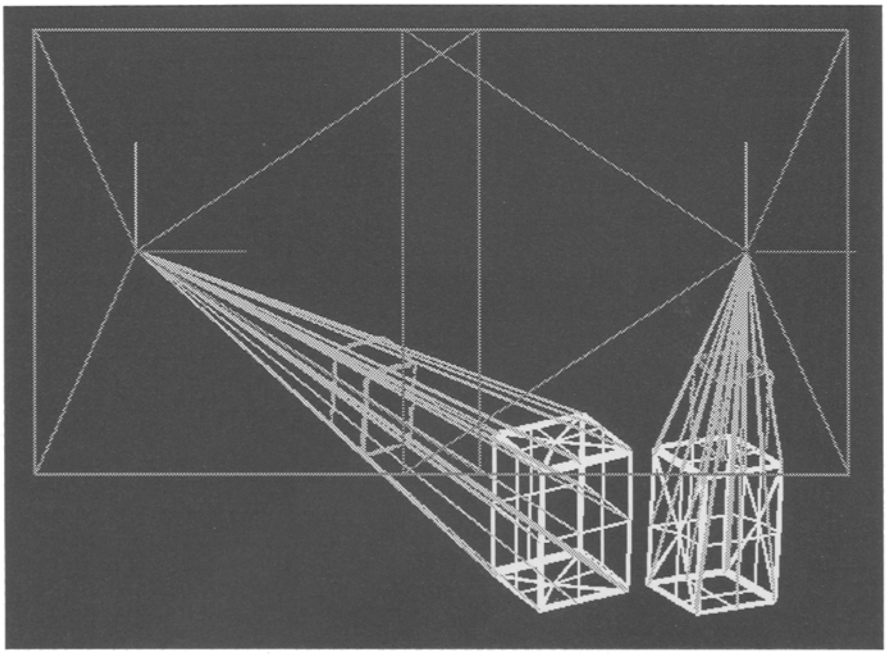


Fig. 18. Independent fitting of models to the smaller block.

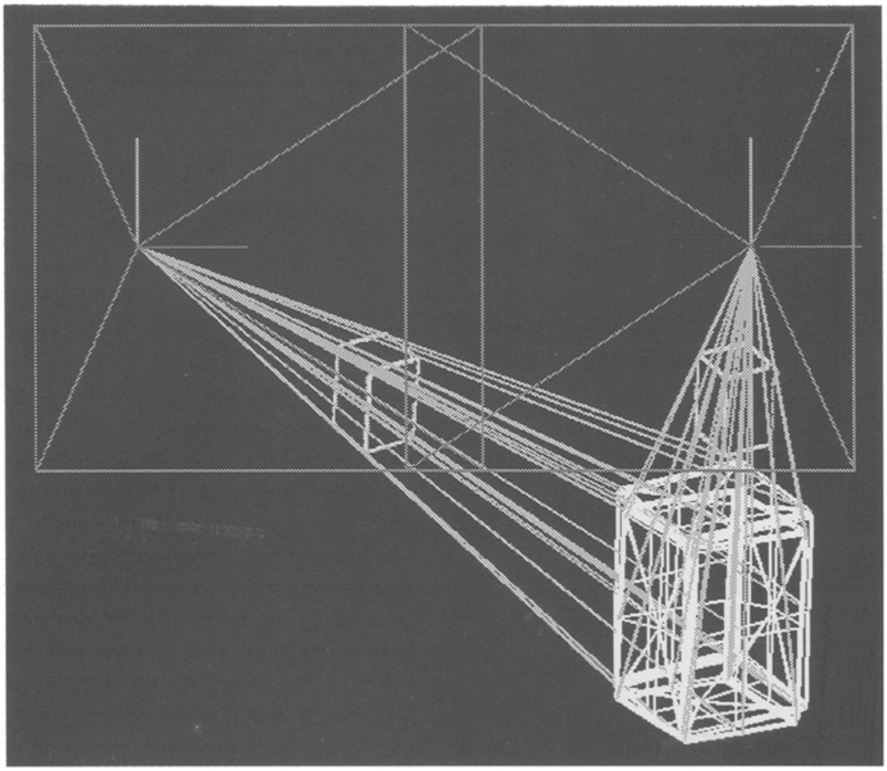


Fig. 19. Localization of model in depth.

the fitting process where each block is being fit independently. Projection rays pass from the camera focal point, through the image contours, and on to the fitted models whose initial depth is chosen arbitrarily. The final step, shown in Figure 19, shows the two models converging in depth. Finally, in Figure 20, we can see both recovered blocks along with their relative depth.

6 Limitations

The approach outlined in this paper is applicable to objects composed of distinct volumetric parts devoid of surface markings or fine structural detail. This is a limitation of the region segmentation scheme, and in order to accommodate more realistic objects, we are currently looking at ways in which salient regions can be abstracted from image detail. Both the qualitative and quantitative shape representation schemes are general. The approach supports any set of qualitative volumetric shapes that can be mapped to a recoverable viewer-centered aspect hierarchy. Moreover, any quantitative shape model that can be defined using our physics-based framework can be deformed by image forces. However, it is important to note that choosing one model will constrain the choice of the other, i.e., a quantitative shape model must be chosen such that it accurately models every possible instance of the qualitative shape model. Finally, it should be noted that the systematic rules that govern the way in which a volume's qualitative shape is used to constrain its quantitative shape recovery are specific to each class of volume. Not only are we exploring how such rules can be automatically extracted through reasoning about the part's shape, but we are also looking at which degrees of freedom of the model can be simultaneously affected by image forces.

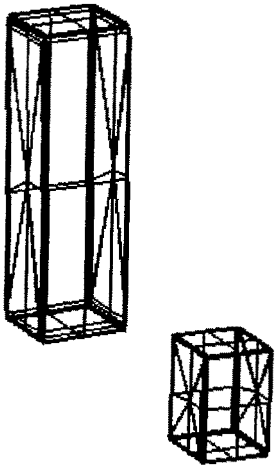


Fig. 20. Rendering of two fitted blocks showing relative depth.

7 Conclusions

The qualitative shape recovery component of the approach is able to capture the coarse shape of objects composed of volumetric primitives *without* solving for exact viewpoint and *without* a precise geometric verification of image features. For many tasks, simply identifying the class of the object is sufficient and there may be no need to either accurately localize the object beyond, for example, "over there", or accurately describe the shape of its components beyond, for example, "cylinder-like". If, however, we need to accurately locate (in order to manipulate) the object once it's been identified, or we need to extract a more detailed shape description in order to distinguish between subclasses of an object, then we can apply the quantitative shape recovery component. The important idea is that the processes of recognizing an object and locating it are decoupled, and that recognition *does not* require accurate localization. In addition, when localization is required, recovered qualitative shape provides strong constraints on the fitting of deformable models, so that the fitting procedure, supporting orthographic, perspective, and stereo projections, is insensitive to both occlusion and initial conditions.

8 Acknowledgments

The authors would like to gratefully acknowledge the assistance of Michael Chan, Gene Amdur, Lars Olsson, Winky Wai, Sean Culhane, James Maclean, Suzanne Stevenson, Yiming Ye, Feng Lu, and Martin Martin in developing some of the software used to generate the results. In addition, Sven Dickinson would like to acknowledge the support of John Tsotsos and of PRECARN Associates' Institute for Robotics and Intelligent Systems, a Government of Canada Network of Centers of Excellence.

Notes

1. These coincide with the model frame axes x, y and x respectively.
2. Since the two cameras are parallel, the projections of the two model frame centers differ only in the x direction.
3. The probability of an aspect hypothesis is the product of the face to aspect mapping and the probability of the face hypothesis from which it was inferred.
4. For a detailed discussion of aspect instantiation and how occluded aspects are instantiated, see Dickinson et al. (1992b).
5. The rules for fitting a superquad to a block assume that the block appears as the most probable aspect, i.e., that aspect which provides the maximum information about the shape of the block.

References

A. Barr. Superquadrics and angle-preserving transformations. *IEEE Computer Graphics and Applications,* 1:11–23, 1981.

R. Bergevin and M. Levine. Generic object recognition: Building coarse 3D descriptions from line drawings. In *Proceedings, IEEE Workshop on Interpretation of 3D Scenes,* pages 68–74, Austin, TX, 1989.

I. Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing,* 32:29–73, 1985.

I. Biederman, J. Hummel, P. Gerhardstein, and E. Cooper. From images edges to geons to viewpoint invariant object models: A neural net implementation. In *Proceedings, SPIE Applications of Artificial Intelligence X: Machine Vision and Robotics,* pages 570–578, Orlando, FL, 1992.

R. Brooks. Model-based 3-D interpretations of 2D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 5(2):140–150, 1983.

D. Clemens. Region-based feature interpretation for recognizing 3-D models in 2-D images. Technical Report AI-TR 1307, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991.

S. Dickinson G. Olofsson, and H. Christensen. Qualitative prediction in active recognition. In *Proceedings, 8th Scandinavian Conference on Image Analysis (SCIA)*, Troms/o, Norway, May 1993.

S. Dickinson, A. Pentland, and A. Rosenfeld, A representation for qualitative 3-D object recognition integrating object-centered and viewer-centered models. In K. Leibovic, editor, *Vision: A Convergence of Disciplines*. Springer Verlag, New York, 1990.

S. Dickinson, A. Pentland, and A. Rosenfeld. From volumes to views: An approach to 3-D object recognition. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 55(2):130–154, 1992a.

S. Dickinson, A. Pentland, and A. Rosenfeld. 3-D shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2): 174–198, 1992b.

R. Fairwood. Recognition of generic components using logic-program relations of image contours. *Image and Vision Computing*, 9(2):113–122, 1991.

A. Gupta. Surface and volumetric segmentation of 3-D objects using parametric shape models. Technical Report MS-CIS-91-45, GRASP LAB 128, University of Pennsylvania, Philadelphia, PA, 1991.

D. Huttenlocher. Three-dimensional recognition of solid objects from a two-dimensional image. Technical Report 1045, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1988.

A. Jacot-Descombes and T. Pun. A probabilistic approach to 3-D inference of geons from a 2-D view. In *Proceedings, SPIE Applications of Artificial Intelligence X: Machine Vision and Robotics*, pages 579–588, Orlando, FL, 1992.

J. Koenderink and A. van. Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32:211-216, 1979.

J. Lee, R. Haralick, and L. Shapiro. Morphologic edge detection. *IEEE Journal of Robotics and Automation*, RA-3(2):142–155, 1987.

D. Lowe. *Perceptual Organization and Visual Recognition.* Kluwer Academic Publishers, Norwell, MA, 1985.

D. Metaxas. Physics-based modeling of nonrigid objects for vision and graphics. *Ph.D. thesis, Dept. of Computer Science, Univ. of Toronto*, 1992.

D. Metaxas and D. Terzopoulos. Constrained deformable superquadrics and nonrigid motion tracking. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 337-343, 1991.

D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):580–591, June 1993.

N. Nilsson. *Principles of Artificial Intelligence*, chapter 2. Morgan Kaufmann Publishers, Inc. Los Altos, CA, 1980.

A. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28:293–331, 1986.

A. Pentland. Automatic extraction of deformable part models. *International Journal of Computer Vision*, 4:107–126, 1990.

A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modeling and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):715–729, 1991.

N. Raja and A. Jain. Recognizing geons from superquadrics fitted to range data. *Image and Vision Computing*, 10(3):179–190, 1992.

P. Saint-Marc, J.-S. Chen, and G. Medioni. Adaptive smoothing: A general tool for early vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):514–526, 1991.

A. Shabana. *Dynamics of Multibody Systems*. Wiley, 1989.

F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(2):131–146, 1990.

D. Terzopoulos and D. Metaxas. Dynamic 3-D models with local and global deformations: Deformable superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):703–714, 1991.

D. Terzopoulos, and A. Witkin. Physically based models with rigid and deformable components. *IEEE Computer Graphics and Applications*, 8(6):41–51, 1988.

D. Terzopoulos, A. Witkin, and M. Kass. Constraints on deformable models: Recovering 3-D shape and nonrigid motion. *Artificial Intelligence*, 36:91–123, 1988.

D. Thompson and J. Mundy. Model-directed object recognition on the connection machine. In *Proceedings, DARPA Image Understanding Workshop*, pages 93–106, Los Angeles, CA, 1987.