Decoupling Recognition and Localization in CAD-Based Vision

Sven J. Dickinson

Department of Computer Science
University of Toronto
6 King's College Rd
Toronto, Ontario
Canada M5S 1A4

Dimitri Metaxas

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

Abstract

Many CAD-based recognition systems have relied on accurate pose estimation and back-projection in order to verify weak correspondences between simple image and model features. This coupling of recognition and localization requires object models which capture the exact geometry of the object, precluding the recognition of generic objects in less restricted domains. In this paper, we synthesize a new approach to 3-D object shape recovery which decouples the processes of recognizing and localizing objects. We first use qualitative shape recovery techniques to recognize objects. If and only if detailed shape or pose is required, we then use the recovered qualitative shape to provide strong fitting constraints on physics-based deformable model recovery techniques.

1 Introduction

Typical CAD-based vision systems for 3-D object recognition from 2-D images have adopted an approach resembling that shown in Figure 1 (e.g., [13, 26, 10]). Small groups of three or more simple image features such as high curvature points, corners, or lines are paired with corresponding features on some model which captures the *exact* geometry of the object. Next, a transformation is computed which brings the model features into alignment with the image features. Finally, the correspondence between image and model features is verified by back-projecting other model features into the image and searching for image features at those locations. If enough image features appear at their expected locations, the object's identity and pose are confirmed.

Despite the popularity of this approach, it is not without its limitations. First of all, since simple image features such as corners and lines are abundant in any CAD model, there may exist many possible correspondences between image and model features that must be hypothesized. As object complexity increases, or as the size of the object database grows, this problem becomes even more acute. Another limitation of the approach is that verifying the position and/or orientation of simple image features requires that the pose of the model be accurately estimated with respect to the image. Furthermore, such geometric verification means that the object models are not invariant to minor changes in shape. For example, if the curvature of a coffee cup handle changes slightly, or the cup's dimensions change, an entirely new CAD model may be needed. Although this approach has proven effective for manufacturing domains where the exact shape of the object is known and the domain of the recognition system is typically a single object, a new paradigm is needed to address less restrictive domains.

Consider, for example, the domain in which a mobile robot, equipped with a vision system, moves about a household retrieving objects for a handicapped person. Not only must the system be aware of a large domain of objects, but it is impractical to expect detailed CAD models to exist for these objects. Rather, the system must have knowledge of generic object classes like cup, book, glass, lamp, etc., which it can use to identify the objects in cluttered scenes. For example, the model of a cup should be some symbolic description such as "bent cylinder attached at its ends to the side of a cylinder." In this manner, everyday objects, or at least those which can be described as catenations of simple volumetric parts, can be easily acquired by the recognition system without complex geometric models.

Since the introduction of a class of generic or qualitatively-defined volumetric primitives, called *geons* [3], interest has been growing in building 3-D

0-8186-5310-8/94 $03.00 © 1994 IEEE

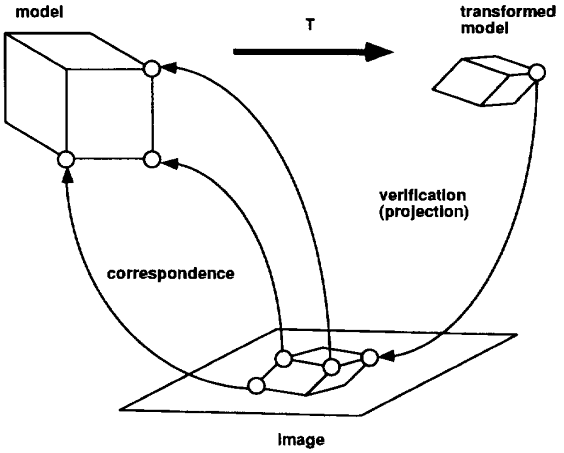


Figure 1: Typical CAD-Based Vision Approach to Object Recognition

object recognition systems based on qualitative shape [2, 7, 8, 21]. One of the primary motivations in these systems is that, as stated by Biederman [3], the task of recognizing (or identifying) an object should be separated from the task of locating it, i.e., determining its pose. Furthermore, the exact shape of the object need not be recovered to facilitate recognition; a coarse-level description of an object in terms of its parts is not only sufficient to distinguish between different classes of objects, but provides an efficient indexing mechanism for recognition from large object databases.

While addressing the problem of identifying or recognizing generic objects from a large database, the above systems have not addressed the problems of either detailed shape extraction or shape localization. Referring back to our robot vision system, its task may be to retrieve the big cup as opposed to the small cup, in which case more detailed geometry may need to be recovered. Furthermore, since the cup is to be grasped, its dimensions and 3-D location must also be recovered.

Physics-based modeling [20, 25, 24, 15, 14, 16] 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. However, 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 entire 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. Clearly, a more robust recovery would require more knowledge of the object's position, orientation, and shape.

In this paper, we propose a two-step recovery process that decouples the tasks of recognizing an object and localizing the object. The first step recovers the qualitative shape of an object in terms of its volumetric parts [5, 7, 6]. If detailed shape or localization is needed to manipulate the object, for example, we then use knowledge of a part's qualitative shape and orientation to provide strong constraints in fitting a deformable model to the part. In tandem, these two steps provide a coarse-to-fine approach to object recognition, which is essential for less structured domains with large numbers of objects.

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 [23, 9], while Pentland [19] 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 in case of occluding boundary data under simple orthographic projection, as is true of earlier work of Terzopoulos et al. [25], Terzopoulos and Metaxas [24], Metaxas and Terzopoulos [17], and Pentland and Sclaroff [20], which address only the problem of model fitting. Taking a different approach, Raja and Jain [21] segment a range image into parts corresponding to geons, and then fit a superquadric to the part to determine geon orientation.

The fundamental difference between our approach and the above approaches is that we use a qualitative segmentation of the image to provide sufficient constraints on 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 [5, 7, 6].

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, generalized cylinders, and superquadrics.

To demonstrate our approach to object recognition, we have selected an object representation similar to that used by Biederman [3], in which the Cartesian product of contrastive 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. [7]). The values of these properties give rise to a set of ten primitives (a subset of Biederman's geons), shown in Figure 2(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

Traditional aspect graph representations of 3-D objects model an entire object with a set of aspects, each defining a topologically distinct view of the object in terms of its visible surfaces (Koenderink and van Doorn [11]). 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 2(b) illustrates a portion of the aspect hierarchy, along with a few of the primitives. The ambiguous mappings between the levels of the aspect hierarchy are captured in a set of conditional probabilities, 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 [24, 14]. We then extend the formulation to the case of orthographic, perspective, and stereo projections.

3.2.1 Geometry

Geometrically, the models used 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: $\mathbf{x}(\mathrm{u}, t) = (x_1(\mathrm{u}, t), x_2(\mathrm{u}, t), x_3(\mathrm{u}, t))^\top$, where $^\top$ is the transpose operator. We set up a noninertial, model-centered reference frame ϕ [14] and express these positions as:

$$\mathbf{x} = \mathbf{c} + \mathbf{R}\mathbf{p}, \tag{1}$$

where $\mathbf{c}(t)$ is the origin of ϕ at the center of the model, and the orientation of ϕ is 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. We further express $\mathbf{p}$ as the sum of a reference shape $\mathbf{s}(\mathrm{u}, t)$ (global deformation) and a displacement function $\mathbf{d}(\mathrm{u}, t)$ (local deformation):

$$\mathbf{p} = \mathbf{s} + \mathbf{d}. \tag{2}$$

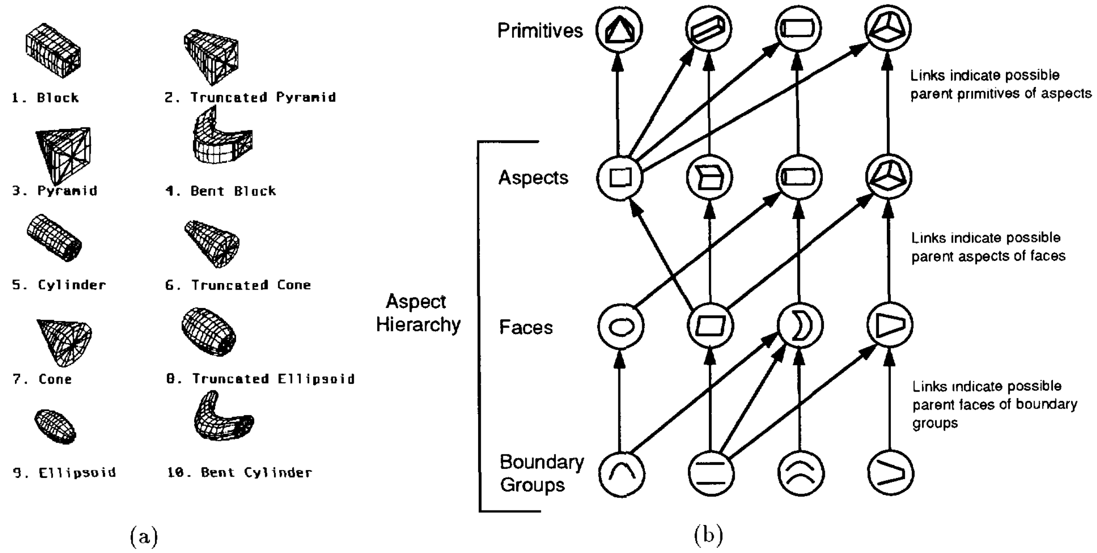


Figure 2: (a) The Ten Modeling Primitives, (b) The Aspect Hierarchy

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

$$\mathbf{p} = \mathbf{s}. \tag{3}$$

Based on the shapes we want to recover, we first consider the case of superquadric ellipsoids [1], which are given by the following formula:

$$\mathbf{e} = a \begin{pmatrix} a_1 C_u{}^{\epsilon_1} C_v{}^{\epsilon_2} \\ a_2 C_u{}^{\epsilon_1} S_v{}^{\epsilon_2} \\ a_3 S_u{}^{\epsilon_1} \end{pmatrix}, \tag{4}$$

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 e[1] into a single parameterized deformation $\mathbf{T}$, and express the reference shape as:

$$\begin{aligned} \mathbf{s} &= \mathbf{T}(\mathbf{e}, t_1, t_2, b_1, b_2, b_3) \\ &= \begin{pmatrix} (\frac{t_1 e_3}{a a_3 w} + 1)\ e_1 + b_1\ cos(\frac{e_3 + b_2}{a a_3 w} \pi b_3) \\ (\frac{t_2 e_3}{a a_3 w} + 1)\ e_2 \\ e_3 \end{pmatrix}, \end{aligned} \tag{5}$$

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 of s into the parameter vector:

$$\mathbf{q}_s = (a, a_1, a_2, a_3, \epsilon_1, \epsilon_2, t_1, t_2, b_1, b_2, b_3)^{\top}. \tag{6}$$

The above global deformation parameters are adequate for quantitatively describing the ten modeling primitives shown in Figure 2(a).

3.2.2 Kinematics and Dynamics

The velocity of points on the model is given by:

$$\dot{\mathbf{x}} = \dot{\mathbf{c}} + \mathbf{B}\dot{\boldsymbol{\theta}} + \mathbf{R}\dot{\mathbf{s}}, \tag{7}$$

[1] These coincide with the model frame axes x, y and z respectively.

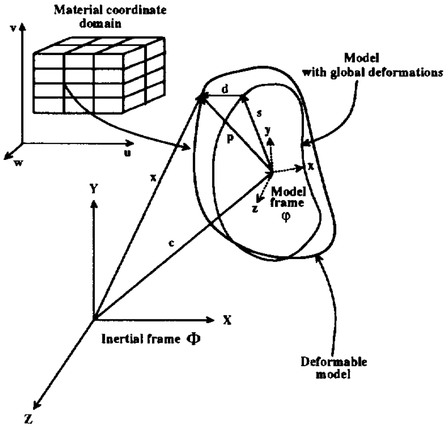


Figure 3: Geometry of Deformable Models

where $\boldsymbol{\theta}$ is the vector of rotational coordinates of the model, and $\mathbf{B} = \partial(\mathbf{Rp})/\partial\boldsymbol{\theta}$. Furthermore, $\dot{\mathbf{s}} = \mathbf{J}\dot{\mathbf{q}}_s$, where $\mathbf{J}$ is the Jacobian of the deformable superquadric model with respect to the global degrees of freedom $\mathbf{q}_s$ [14]. We can therefore write:

$$\dot{\mathbf{x}} = [\mathbf{I}\ \mathbf{B}\ \mathbf{RJ}]\dot{\mathbf{q}} = \mathbf{L}\dot{\mathbf{q}}, \tag{8}$$

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 $\mathbf{q}_\theta = \boldsymbol{\theta}$.

When fitting the model to visual data, our goal is to recover $\mathbf{q}$, the vector of degrees of freedom of the model. 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. In the absence of local deformations, the Lagrange equations of motion take the form [24]:

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{D}\dot{\mathbf{q}} = \mathbf{g}_q + \mathbf{f}_q, \tag{9}$$

where $\mathbf{M}$ and $\mathbf{D}$ are the mass and damping matrices, respectively, $\mathbf{g}_q$ are inertial forces arising from the dynamic coupling between the local and global degrees of freedom, and $\mathbf{f}_q(\mathrm{u}, t)$ 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.

3.2.3 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:

$$x_p = x, \qquad y_p = y. \tag{10}$$

By taking the derivative of the above equation (10) with respect to time, we arrive at the following formulas:

$$\dot{x}_p = \dot{x}, \qquad \dot{y}_p = \dot{y}. \tag{11}$$

Rewriting (11) in matrix form and using (8), we arrive at the following matrix equations:

$$\begin{bmatrix} \dot{x}_p \\ \dot{y}_p \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{L}\dot{\mathbf{q}}. \tag{12}$$

If we rewrite (12) in compact form, we get

$$\begin{bmatrix} \dot{x}_p \\ \dot{y}_p \end{bmatrix} = \mathbf{L}_o\dot{\mathbf{q}}, \tag{13}$$

where

$$\mathbf{L}_o = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{L}. \tag{14}$$

3.2.4 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:

$$x_p = \frac{x}{z}f, \qquad y_p = \frac{y}{z}f, \tag{15}$$

where f is the focal length.

By taking the derivative of the above equation (15) with respect to time, we arrive at the following formulas:

$$\dot{x}_p = \dot{x}\frac{f}{z} - \frac{x}{z^2}f\dot{z}, \qquad \dot{y}_p = \dot{y}\frac{f}{z} - \frac{y}{z^2}f\dot{z}. \tag{16}$$

Rewriting (16) in matrix form and using (8), we arrive at the following matrix equations

$$\begin{bmatrix} \dot{x}_p \\ \dot{y}_p \end{bmatrix} = \begin{bmatrix} f/z & 0 & -x/z^2 f \\ 0 & f/z & -y/z^2 f \end{bmatrix} \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \tag{17}$$

$$\begin{bmatrix} f/z & 0 & -x/z^2 f \\ 0 & f/z & -y/z^2 f \end{bmatrix} \mathbf{L}\dot{\mathbf{q}}. \tag{18}$$

If we rewrite (18) in compact form, we get

$$\begin{bmatrix} \dot{x}_p \\ \dot{y}_p \end{bmatrix} = \mathbf{L}_p\dot{\mathbf{q}}, \tag{19}$$

where

$$\mathbf{L}_p = \begin{bmatrix} f/z & 0 & -x/z^2 f \\ 0 & f/z & -y/z^2 f \end{bmatrix} \mathbf{L}. \qquad (20)$$

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.

3.2.5 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 (15) and the corresponding Jacobian matrices $\mathbf{L}_{pL}$ and $\mathbf{L}_{pR}$ are calculated using (20).

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[2] 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:

$$\dot{\mathbf{c}}_k = s|c_{LIx} - c_{RIx}|\frac{\mathbf{c}_k}{||\mathbf{c}_k||} \qquad (21)$$

  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 (4), and using (9), we modify a.

4 Shape Recovery

Recovering a volumetric description from the image consists of two steps. First, a qualitative 3-D volume is recovered from the image. Next, the recovered qualitative volume is used to constrain the fitting of a deformable model to the data. In this section, we describe each of these steps in greater detail.

4.1 Qualitative Shape Recovery

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 [22], followed by a morphological gradient operator [12]. 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 [22]. 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

[2] Since the two cameras are parallel, the projections of the two model frame centers differ only in the x direction.

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 [7] 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[3]. At each face, we collect all the aspect hypotheses (corresponding to all face hypotheses) and rank them in decreasing order of probability.[4]

We can now reformulate our bottom-up aspect recovery problem as a search through the space of aspect labelings of the faces in the face topology graph. In other words, we wish to choose one aspect hypothesis from the list at each node in the face topology graph, such that the instantiated aspects completely cover the graph. For our search through the possible aspect labelings of the face topology graph, we employ Algorithm A (Nilsson [18]) 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. Those faces in the face topology graph whose labels match the target face prediction provide an ordered (by decreasing probability) set of ranked search positions at which the target aspect prediction can be verified. If the mapping from a verified aspect to a target volume is ambiguous, this attention mechanism can be used to drive an active recognition system which moves the cameras to obtain a less ambiguous view of an object's part [4].

4.1.3 Primitive Recovery

In the expected object recognition approach described above, volume recovery consists of using the aspect hierarchy to map the recovered aspect directly to the target volume prediction. Volume recovery for the unexpected object recognition case is more complex. From an *aspect covering* of the regions in the image, a set of volume labels and their corresponding probabilities is inferred (using the aspect hierarchy) from each aspect. Volume recovery is formulated as a search through the space of volume labelings of the aspects in the aspect covering, guided by a heuristic based on the probabilities of the volume labels. Each solution, or *volume covering*, found by the search is a valid volumetric part interpretation of the input image. Encoded in each recovered volume is the aspect in which it it 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; di-

[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 [7].

mensions, 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

In computer vision applications [24], we can simplify the equations while preserving useful dynamics by setting the mass density $\mu(\mathrm{u})$ to zero to obtain:

$$\mathbf{D}\dot{\mathbf{q}} = \mathbf{f}_q. \tag{22}$$

These equations yield a model which has no inertia and comes to rest as soon as all the applied forces vanish or equilibrate. Equation (22) is discretized in material coordinates u using nodal finite element basis functions. We carry out the discretization by tesselating the surface of the model into linear triangular elements. Furthermore, for fast interactive response, we employ a first-order Euler method to integrate (22).

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 [24]. We apply forces to the model 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 2D image forces f, using the following formula:

$$\mathbf{f}_q^\top = \int \mathbf{f}^\top \mathbf{L}_k \, du = (\mathbf{f}_{q_c}^\top, \mathbf{f}_{q_\theta}^\top, \mathbf{f}_{q_s}^\top), \tag{23}$$

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:

$$|\mathbf{i} \cdot \mathbf{n}| < \tau, \tag{24}$$

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 [25, 24, 20]. 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 4; the results of the bottom-up (unexpected) qualitative shape recovery algorithm are also shown in Figure 5. 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 2(a)), and primitive probability,

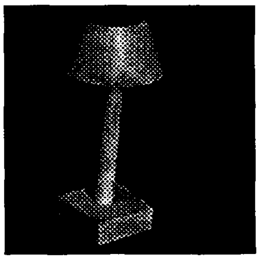
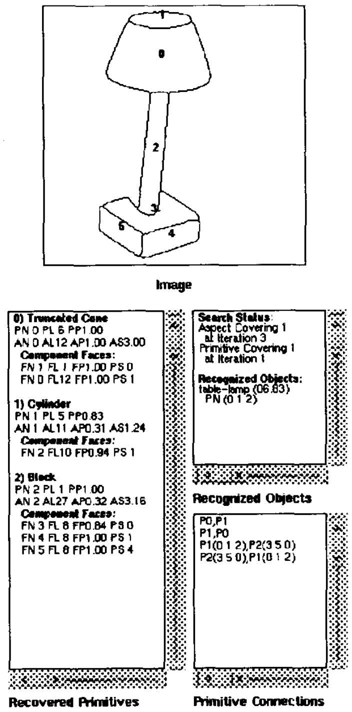


Figure 4: Original Image

respectively. The mnemonics AN, AL, AP, and AS refer to the aspect number (an enumeration), aspect label (see [7]), 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 [7]), face probability, and corresponding primitive attachment surface (see [7]), 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 (superquadric) 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. Fig. 6(a) shows the image contours corresponding to the lamp shade and the cylinder following application of this force. Fig. 6(b) shows a different view of the image plane, providing a better view of the model cylinder.



```
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)
```

Figure 5: Recovered Qualitative Primitives

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. Figs. 6(c) and 6(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 to a force affecting the orientation of the cylinder. Figs. 6(e) and 6(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. Figs. 6(g) and 6(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. Figs. 6(i) and 6(j) show the image and the tapered cylinder following application of this force. The result of fitting all three parts of the lamp is shown in Figure 7, along with a side view in showing how an arbitrary depth must be chosen for each part under an orthographic projection model.

In an example illustrating the stereo recovery technique, we apply the top-down object recognition algorithm to the stereo pair shown in Figure 8; 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 8. Volume score thresholds were set high so that volumes appearing in only the most probable aspect and with little or no occlusion were accepted.[5] 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 9 captures the stage in 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 10, shows the two models converging in depth. Finally, in Figure 11, 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. That is, any set of qualitative volumetric shapes that can be mapped to a recoverable viewer-centered aspect hierarchy, and any quantitative shape

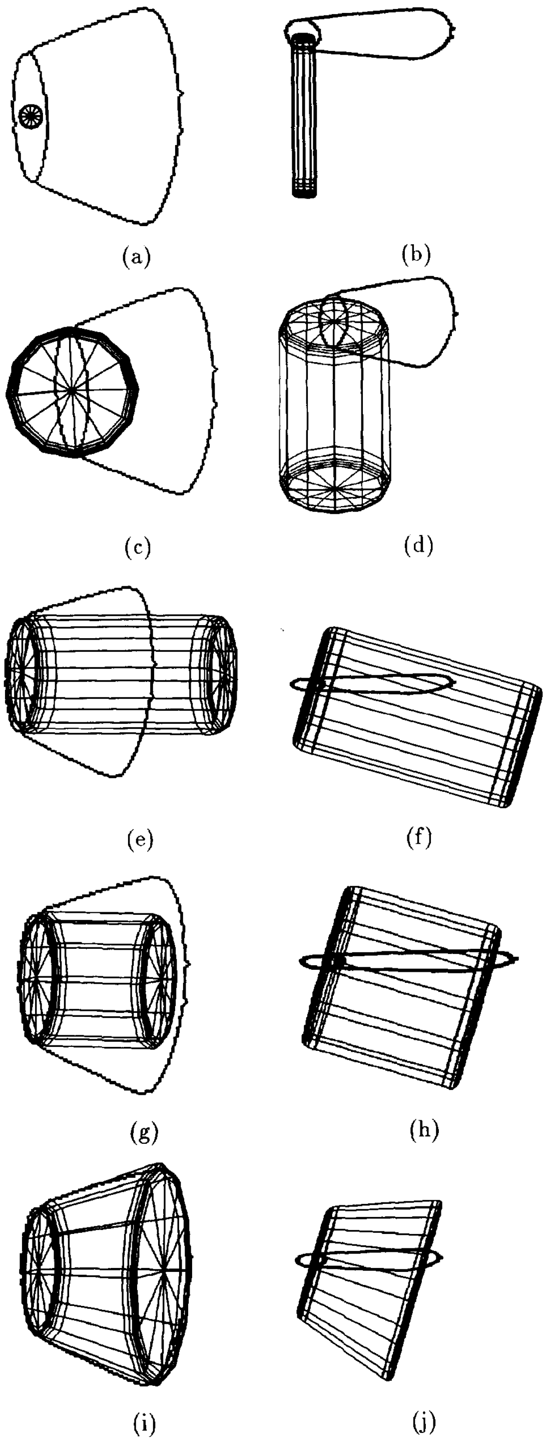


Figure 6: Quantitative Shape Recovery for Lamp Shade

[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.

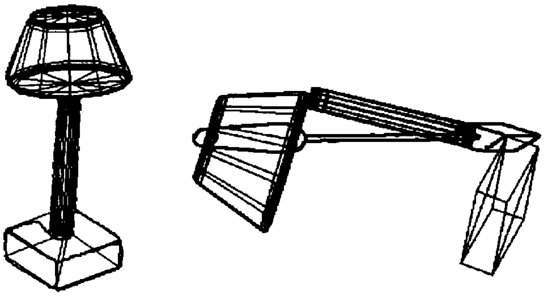


Figure 7: Final Recovery of Table Lamp. Note that depth information is lost in orthographic projection.

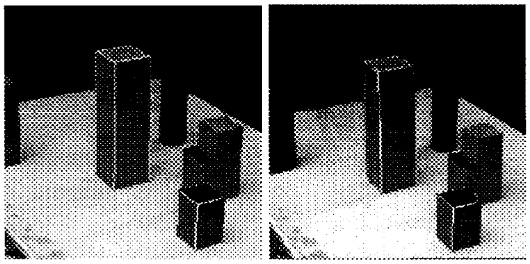


Figure 8: Left and Right Stereo Images of a Cluttered Table with Corresponding Recovered Blocks Highlighted

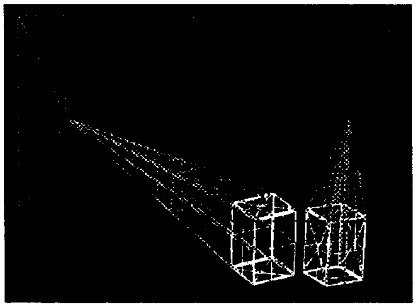


Figure 9: Independent Fitting of Models to the Smaller Block

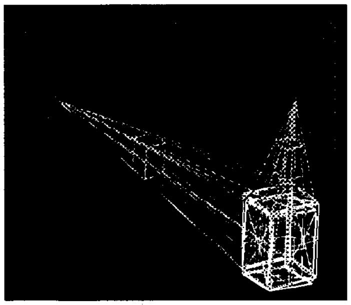


Figure 10: Localization of Model in Depth

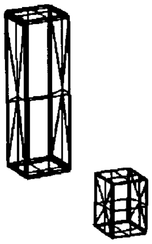


Figure 11: Rendering of Two Fitted Blocks Showing Relative Depth

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.

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 Acknowledgements

The authors gratefully acknowledge the assistance of Michael Chan in the implementation of this work. In addition, the authors would like to thank Gene Amdur for software and systems support.

References

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

[2] 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.

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

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

[5] 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.

[6] 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, 1992.

[7] 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, 1992.

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

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

[10] D. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, 1990.

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

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

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

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

[15] 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.

[16] 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.

[17] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, in press*, 1993.

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

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

[20] 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.

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

[22] 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):514–529, 1991.

[23] F. Solina. Shape recovery and segmentation with deformable part models. Technical Report MS-CIS-87-111, GRASP LAB 128, University of Pennsylvania, Philadelphia, PA, 1987.

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

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

[26] 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.