

# 3-D to 2-D Recognition with Regions\*

David Jacobs  
NEC Research Institute  
4 Independence Way  
Princeton, NJ 08540, USA  
dwj@research.nj.nec.com

Ronen Basri<sup>†</sup>  
Department of Applied Math.  
The Weizmann Inst. of Science  
Rehovot, 76100, Israel  
ronen@wisdom.weizmann.ac.il

## Abstract

This paper presents a novel approach to parts-based object recognition in the presence of occlusion. We focus on the problem of determining the pose of a 3-D object from a single 2-D image when convex parts of the object have been matched to corresponding regions in the image. We consider three types of occlusions: self-occlusion, occlusions whose locus is identified in the image, and completely arbitrary occlusions. We derive efficient algorithms for the first two cases, and characterize their performance. For the last case, we prove that the problem of finding valid poses is computationally hard, but provide an efficient, approximate algorithm. This work generalizes our previous work on region-based object recognition, which focused on the case of planar models.

---

\*A preliminary version of this paper has appeared in [29] A brief overview of these and related results has appeared in [8]

<sup>†</sup>This research was supported by the Unites States-Israel Binational Science Foundation, Grant No. 94-100. The vision group at the Weizmann Inst. is supported in part by the Israeli Ministry of Science, Grant No. 8504. Ronen Basri is an incumbent of Arye Dissentshik Career Development Chair at the Weizmann Institute.

# 1 Introduction

Recognizing a known 3-D object using a single 2-D image is a central and difficult problem in visual recognition. One of the key issues is developing adequate representations to support flexible recognition of general objects. Existing approaches are often well-suited to only a small class of objects (eg. polyhedra, rotationally symmetric objects, low-order algebraic surfaces). In this paper we show how to make use of a very simple and general representation of the parts of 3-D objects to determine their pose. At the same time, we also provide results on the fundamental computational complexity of determining the pose of 3-D objects in the presence of occlusion.

A good representation for 3-D recognition should: 1) be rich enough to describe the shape of 3-D objects; 2) have a 2-D analog that can be reliably computed from an image; and 3) allow us to understand the relationship between the 3-D representation and its 2-D projection to perform useful recognition tasks. In this paper, we focus on the capability of a representation to support pose determination, one of the most basic problems faced by a complete recognition system.

Existing representations for 3-D recognition in 2-D images have significant limitations. Many previous approaches have relied on finding a correspondence between simple geometric features, such as points or lines. Lowe[33] and Clemens[14], for example, determine pose based on a match between line segments, while Fischler and Bolles[18], Huttenlocher and Ullman[26], Horaud[22], Ullman and Basri[57], Jacobs[27], Rothwell et al.[47], and Alter and Jacobs[1] use point features to determine pose, and Thompson and Mundy[54] make use of vertices. It is fairly well understood how to use local features for pose determination or indexing, but they have significant weaknesses. Local features often do not capture the shape of complex, curved 3-D objects. And it may be quite difficult to locate 2-D image features that correspond to the local features of a non-polyhedral 3-D object, since the contour generator of such objects is completely viewpoint dependent.

There has been some recent work that extracts and matches point features from the outlines of smooth objects. Forsyth et al.[19] use points derived from the bitangents of objects to derive an invariant description of the contour. This work, however, is limited to rotationally symmetric objects. Vijayakumar, Kriegman and Ponce[58] show how to build an indexing function using bitangents for more general curved 3-D objects. They show the surprising result that a description based on bitangents can be represented as 1-D curves in a lookup table. These approaches use only a limited amount of the structural information available, however.

Another approach to recognizing smooth 3-D objects involves describing the 3-D object and 2-D image with algebraic surfaces and curves, and then registering these algebraic descriptions. Kriegman and Ponce[31] have taken this approach, using elimination methods to solve for object pose. While this approach has provided significant insight into how the overall problem may be solved, it has the disadvantage of requiring a somewhat complex, iterative solution method. Specifically, their method requires a good estimate of pose to begin with, and then uses a variation of Newton's method to converge to the locally optimal pose. Forsyth[20] has shown

how to use an algebraic description of an image contour to determine the projective shape of the algebraic surface that produced it. This result is not practical, however, as it is extremely sensitive to noise. In general, while algebraic descriptions may be used to accurately represent a 3-D model, it is extremely difficult to derive a corresponding description of an image, since such descriptions may be very sensitive to noise.

A number of other approaches attempt to represent 3-D models using a specific vocabulary of shapes, typically based on an algebraic description, such as generalized cylinders [11, 12, 21, 34, 42, 50, 56, 59], superquadrics [39, 45, 51, 53], and geons [9, 10]. Often these approaches handle only a limited class of objects. For example, when generalized cylinders are used a major difficulty lies in computing the 2-D projection of the 3-D axis and sweeping rule. Image occlusion and noise make this problem especially difficult. Considerable effort has led to solutions of this problem for only some restricted classes of shapes.

Moment-based methods are somewhat related to ours, in that they compute a description of image regions to match to model volumes. These methods might align regions based on their center of mass, or on higher order moments. Examples of this approach can be found in [17, 23, 37, 36, 40, 43, 44, 48]. These approaches do not extend to the recognition of a 3-D object from a single 2-D image, however. First of all, volumes of 3-D points always produce self-occlusion, since different subsets of the surface of the volume are visible from different viewpoints. Therefore, the center of mass of the projection of a model volume will not be the projection of the center of mass of the 3-D point set. Second, the center of mass of a surface that is curved in 3-D does not project to the center of mass of its image, even when there is no self-occlusion. This is because the extent to which different portions of the 3-D surface are foreshortened will depend on the viewing direction. A final disadvantage of methods based on moments is that they are sensitive to occlusion by other objects in the scene.

We address the problem of 3-D recognition from a 2-D image in a parts-based framework. That is, like most work on the recognition of general, curved 3-D objects, we divide a 3-D object into its component parts, and expect that a bottom-up grouping system will identify image regions that are candidate matches for these parts. We use a simple, direct representation of an object's parts as general volumes in 3-D, using 2-D areas to represent their image projection. This representation can be applied to any 3-D shape (we discuss restrictions concerning convexity later), and derived from a 2-D image without any need to fit algebraic constructs (eg., conics, lines, corners) to parts located in an image, or to compute other intermediate properties of parts such as their axes. Therefore, our first two goals for a representation are satisfied very generally, with no assumptions except those imposed by a parts-based approach to recognition.

The bulk of this paper will attack the third goal of a representation by showing how we can relate 3-D volumes to 2-D regions for the important problem of pose determination. This extends our previous results, which focused on planar objects ([7]). At the same time, we use this general representation to show novel results about the fundamental complexity of pose determination for general 3-D objects. To do this, we divide the possible types of occlusions that may make pose determination difficult into three conditions. First, we consider self-occlusions. We provide a simple algorithm that uses linear programming to find pose, given correspondences

between 3-D volumes and 2-D regions. This is the same method that we applied to planar objects in [7], although we provide some new results demonstrating when this algorithm will produce correct results. Second, we consider occlusions whose position has been identified in the image. That is, we assume that each edge bounding an image region is labeled as either a region boundary, or an occlusion boundary. We show that a variation on our original algorithm can handle this case too. Third, we consider arbitrary occlusions of unknown location. That is, we assume that a 2-D region is a subset of the projection of the corresponding 3-D volume, but that any of the boundaries of the region may be due to occlusion. We show that the problem of pose determination in this case is fundamentally harder, by showing that this is a superset of a problem in computational geometry that is known to be hard. We stress that this result is not specific to our algorithms. It shows that for parts-based object recognition, the problem of pose determination is provably more difficult when occlusions are of unknown location in an image than when their position is known. However, we then provide an approximate algorithm, which is computationally efficient, and show that in a number of cases this leads to accurate results. Finally, we show how to handle degenerate solutions, which can especially occur with this approximate solution.

While we focus on the pose determination problem, we expect these results to fit into a complete recognition system as follows. At compile time, we divide an object up into component parts, preferably convex. At run time, we use a grouping system to identify candidate image groups. One might use intensity-based segmentation or a system that finds salient convex sets of edges ([28]). We then consider matches between image groups and model parts, with a search that may be directed with the addition of cues such as color, as was done by Nayar and Bolle[37]. Pose is determined using our current work, and then a hypothetical projection of the model may be confirmed or rejected using additional cues. The steps of this process are illustrated in numerous experiments described in [27]. That system robustly matched convex object parts, but used a feature-based indexing system not suitable for non-polyhedral 3-D objects. In our current paper, all of these steps are also implemented, except we do not experiment with search to match object parts, focusing instead on determining the capability of our system to produce accurate poses once the correct match is found.

## 2 Using Volumes and Regions to Determine Pose

We assume that a hypothesized match exists between a set of model volumes and image regions. Our goal is to use this match to determine the pose of the model.

We assume that the model consists of a set of 3-D volumes denoted:  $V_1, \dots, V_k \subset \mathcal{R}^3$ , where each volume is an arbitrary subset of  $\mathcal{R}^3$ . Similarly, we assume that the image consists of 2-D regions, which are each subsets of  $\mathcal{R}^2$ , and which we denote by:  $R_1, \dots, R_k \subset \mathcal{R}^2$ . Our solution methods will apply to the case where the model and image sets are convex; if we wish to make use of non-convex volumes or regions we should first take their convex hulls. This means that our methods can naturally apply also when some or all of the correspondences are between point features, or (possibly partially occluded) line segments, since these are convex.

Next, we suppose that the image was generated by applying some transformation,  $T$ , that maps points in the 3-D model to points in the image. For the most part we will assume that this is a 3-D to 2-D affine transformation. We denote a point in model space by  $\vec{p} = (x, y, z)$  and in image space by  $\vec{q} = (u, v)$ . If  $\vec{q} = T(\vec{p})$  then we denote  $u = T_u(\vec{p})$  and  $v = T_v(\vec{p})$ . Our goal is to identify the transformation that will best explain the image regions as the product of their corresponding model regions.

As we point out in [7], the problem of finding a transformation that perfectly matches a set of model volumes to their corresponding image regions is a non-convex optimization problem. This follows from the fact that the set of feasible transformations need not be convex, or even connected. Consider for example the case of a model square matched to an identical image square. Matching the model exactly to the image can be performed in four ways (separated from each other by a  $90^\circ$  rotation). Obviously, no intermediate transformation provides a solution to this matching problem. While non-convex optimization problems are often attacked using tools such as gradient descent, we instead take the approach of showing that two different problem formulations can make the matching process convex, and therefore easier to optimize. In section 3 we discuss the conditions under which these formulations will produce the correct model pose.

## 2.1 The Forward Constraints: self-occlusion

First, we show that our problem becomes convex if we merely require that every model point projects inside the corresponding image region, by reviewing results that we have shown in [7]. Formally, the *forward constraints* are satisfied by the transformation,  $T$ , if and only if  $\forall \vec{p} \in V_i$ ,  $T\vec{p} \in R_i$  (that is,  $TV_i \subseteq R_i$ ). These constraints allow a volume to occlude itself (ie. two model points may project to the same image point). They do not capture all possible constraint, however, since they do not require that each image point be explained by a corresponding model point.

We first consider a projection model consisting of a 3-D affine transformation followed by an orthographic projection. [6] shows how to apply our method for perspective projection, as well. Denote the linear part of  $T$  by  $A$ , where  $A$  is a non-singular  $2 \times 3$  matrix with elements  $t_{ij}$ , and the translation part by  $\vec{t} = (t_x, t_y)$ . Then:

$$\begin{aligned} u &= t_{11}x + t_{12}y + t_{13}z + t_x \\ v &= t_{21}x + t_{22}y + t_{23}z + t_y. \end{aligned} \tag{1}$$

This projection model and its equivalent has been recently used by a number of researchers ([32, 57, 30, 55, 27]). It is also equivalent to applying scaled orthographic projection followed by a 2-D affine transformation [27], that is, taking a picture of a picture. Alternately, it is equivalent to a paraperspective projection followed by translation [5], where paraperspective is a first-order approximation to perspective projection [41, 52].

To express the forward constraints, we note that since  $R$  is convex, there exists a set of lines  $L_R$  bounding  $R$  from all directions such that for every point  $\vec{q} \in R$  and for every line  $l \in L_R$

we can write

$$l(\vec{q}) \geq 0. \quad (2)$$

Let the line  $l$  be expressed by the equation:  $Au + Bv + C \geq 0$ , then the constraint  $TV \subseteq R$  can be written as follows. Every point  $\vec{p} \in V$  should be mapped by  $T$  to some point  $\vec{q} \in R$ , and so

$$A(t_{11}x + t_{12}y + t_{13}z + t_x) + B(t_{21}x + t_{22}y + t_{23}z + t_y) + C \geq 0. \quad (3)$$

This constraint is linear in the transformation parameters. Denote

$$\vec{w}^T = (t_{11}, t_{12}, t_{13}, t_x, t_{21}, t_{22}, t_{23}, t_y)$$

the vector of unknown transformation parameters, and

$$\vec{g}^T = (Ax, Ay, Az, A, Bx, By, Bz, B).$$

We can rewrite the forward constraints as

$$\vec{g}^T \vec{w} \geq -C. \quad (4)$$

We can similarly handle affine transformations followed by perspective projection. In that case

$$\begin{aligned} u &= \frac{f(t_{11}x + t_{12}y + t_{13}z + t_x)}{t_{31}x + t_{32}y + t_{33}z + t_z} \\ v &= \frac{f(t_{21}x + t_{22}y + t_{23}z + t_y)}{t_{31}x + t_{32}y + t_{33}z + t_z} \end{aligned}$$

where  $f$  is the focal length. The forward constraint  $Au + Bv + C \geq 0$  now contains the term  $t_{31}x + t_{32}y + t_{33}z + t_z$  in the denominator. This term must be positive since we require the object to appear in front of the camera. So, we can multiply both sides of the inequality by this term, again obtaining a linear constraint with the same general form as Eq. (4), with different definitions of  $\vec{w}$  and  $\vec{g}$ .

The set of forward constraints consists of all such constraints obtained for all pairs of bounding lines  $l \in L_R$  and model points  $\vec{p} \in V$ . This gives us one constraint for every point in the model volumes and for every tangent line to the image regions. For curved objects, therefore, the number of constraints is infinite, but we may sample them as accurately as we desire. The issue of sampling is addressed in [6]. For polyhedral volumes and polygonal regions the number of independent constraints is finite, and given by the vertices of the model volumes and the sides of the image regions. The rest of the constraints are redundant.

We therefore seek a set of transformation parameters that satisfy a set of linear constraints of the form:

$$\vec{g}_i^T \vec{w} \geq c_i, \quad i = 1, \dots, n. \quad (5)$$

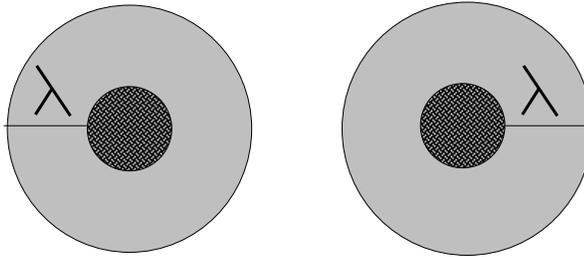


Figure 1: The dark circles are positioned by the similarity transformation that maximizes  $\lambda$  relative to the larger, shaded circles.

This can be written in matrix notation as:

$$\vec{G}\vec{w} \geq \vec{c}. \quad (6)$$

We may find a set of parameters that satisfy these linear constraints using linear programming. To do this, we must also specify a linear objective function. A common way of doing this is by introducing an additional unknown,  $\lambda$ , in the following way.

$$\begin{aligned} & \max \lambda \\ \text{s. t. } & G\vec{w} \geq \vec{c} + \lambda\vec{1} \end{aligned} \quad (7)$$

A solution to (6) exists if and only if a solution to (7) with  $\lambda \geq 0$  exists. (Note that other objective functions, e.g., the perceptron function, can be used for recovering  $\vec{w}$ , see e.g., [16] for a discussion of solutions to the linear discriminant functions problem.)

When  $\lambda \geq 0$  its value represents the minimal distance of a point to any line bounding the region (Figure 1). Maximizing  $\lambda$  amounts to attempting to contract the model volume inside the image region as much as possible. When  $\lambda < 0$  this attempt fails. In this case any model point that violates the constraints is mapped to a distance of no more than  $|\lambda|$  from its target regions. ( $|\lambda|$  in this case represents a maximum norm, and so it is related to the Hausdorff metric. For work on Hausdorff matching, see [24, 25]. Also, [4] specifically discusses the efficient Hausdorff matching of convex shapes undergoing translation and scaling.)

Solving the system (7) may result in over-contraction. Consider, for example, the case of matching a single volume  $V$  to a single region  $R$ . The forward constraints restrict the set of possible transformations to those that map every point  $\vec{p} \in V$  inside the region  $R$ . Assume  $T$  is a feasible transformation, that is  $TV \subseteq R$ , then applying any contracting factor  $0 \leq s \leq 1$  to  $V$  would also generate a valid solution; namely,  $T(sV) \subseteq R$ . (We assume here without the loss of generality that the origin of the model is set at the centroid of  $V$ .) Consequently, the case of matching one volume with one region necessarily introduces multiple solutions. The solution picked by Eq. 7 is the one with  $s = 0$ . This will contract  $V$  to a point, which is then translated to the point inside  $R$  furthest from any of its bounding tangent lines. This solution produces the largest value of  $\lambda$ . Clearly, the case of matching one volume to one region cannot be solved by the forward constraints alone. However, we will show that we can determine pose accurately when we match a larger number of volumes and regions.

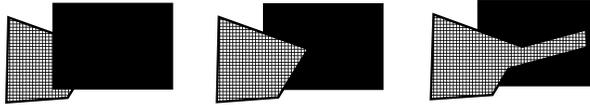


Figure 2: Suppose the cross-hatched, polygonal area is an image region known to be occluded by the dark rectangle (left). When the corresponding model volume is convex, we may apply the forward constraints, using the largest possible convex region consistent with this occlusion (middle). However, this may be incorrect if the occluded region is not convex (right), and we wish to use the convex hull of the region for the forward constraints.

## 2.2 The Forward Constraints with Known Occlusion

The forward constraints can also be validly applied in the presence of occlusion, provided that this occlusion can be identified in the image. Suppose we have identified a region in the image, but we know that some of the boundary of this region is due to another, occluding object and is not in fact the boundary of the region itself. By allowing for the region to be extended in the direction of such occlusions, we can construct a larger convex region which we know should contain the projection of the corresponding model volume.

This can be implemented with very little modification to the above algorithm. Suppose we approximate the border of a detected region with a set of line segments, some of which we know come from the boundary of the region, and some of which come from an occlusion. We apply the forward constraints, as described above, using only those line segments that originate due to the boundary of the object. Implicitly, this restricts the model volume to project within the largest possible convex region that is consistent with the known region boundary.

This method would be more complicated to apply, however, when we are using the convex hull of a concave object part as a volume. In that case, we cannot assume that by extending a region in the direction of an occlusion that we will cover all of the region that has been occluded. It is possible that the occluded portion of the region contains concavities, and extends outside of the largest possible convex region containing the detected region. This is illustrated in figure 2. We will not consider this problem in detail.

## 2.3 The Backward Constraints: Unknown Occlusions

We can allow for arbitrary, unidentified occlusions in the image with the *backward constraints*. The backward constraints are the inverse of the forward. Instead of requiring that each model volume project completely inside each corresponding image region, the backward constraints require that each image region lie completely inside the projection of each corresponding model volume. That is, the backward constraints are satisfied by the transformation,  $T$ , if and only if for every point,  $\vec{q}$ , in every image region,  $\vec{q} \in TV_i$  (that is,  $R_i \subseteq TV_i$ ).

It is preferable to enforce the backward constraints, rather than the forward constraints, when recognizing objects in the presence of occlusion. Typically, when a model volume is occluded by a different object or a different model volume, a region detector will locate image

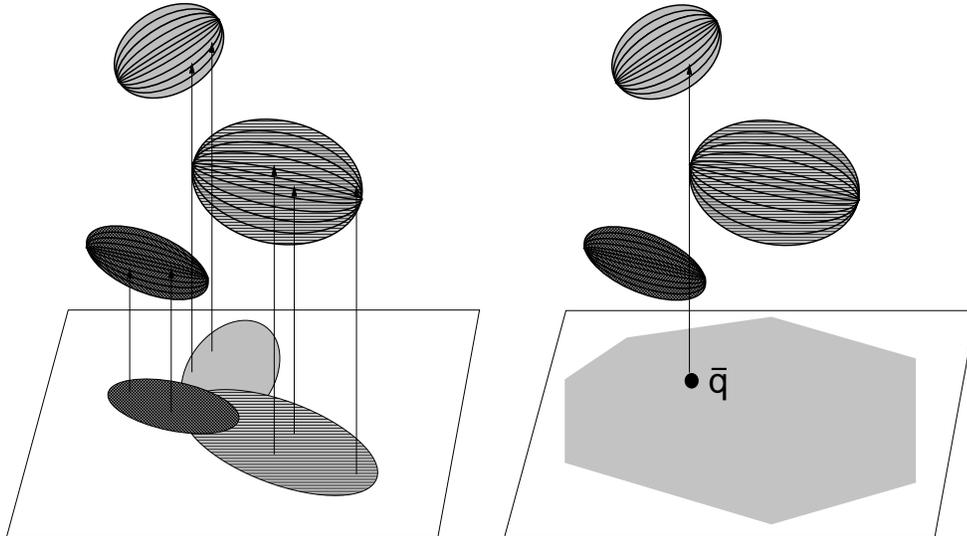


Figure 3: On the left, we show three model volumes projecting to three, overlapping image regions. On the right, we imagine that these regions are occluded, except for one point which they all share. In this case, the problem of solving the backward constraints is equivalent to that of finding a line traversal of the model volumes.

regions that are subsets of the true, unoccluded regions. In this case, the true transformation may map a model volume so that part of the volume accounts for the visible portion of the region, and part of the model volume is mapped outside this region. Therefore, the backward constraints express precisely our state of knowledge about the pose when we allow for arbitrary occlusion of the image regions.

As we show in [7], in the case of planar objects we may solve the backward constraints using linear programming. This is because the transformations that we use to relate a planar model to a planar image (similarity, affine or projective) are invertible. Therefore, requiring that the viewing transformation maps the model volumes so that they completely enclose the corresponding image regions is equivalent to requiring that the inverse transformation map the image regions completely inside the corresponding model regions. For planar models, the forward and backward constraints have an identical form, with the role of the image and model reversed, and so we may find poses that satisfy them using the same method.

In the case of a 3-D model and a 2-D image, the 3-D to 2-D viewing transformation is no longer invertible, however. It is therefore not obvious how one can apply the backward constraints in this case.

In fact, we can prove that it is not possible to directly extend our planar results on the backward constraints to 3-D models. To do this, we now show that there is a close connection between the problem of solving the backward constraints and the computational geometry problem of finding a *line traversal*, ie. finding a line that intersects a set of 3-D solids. Suppose first that we have an arbitrary set of model volumes, but that every image region consists

of only a single point, and that these points are identical, denoted by  $\bar{q}$  (see Fig. 3). The backward constraints are satisfied by a transformation if and only if the 3-D line that projects to  $\bar{q}$  intersects every model volume. Such a transformation exists if and only if there exists a single line that intersects each of the model volumes. This shows that the problem of solving the backward constraints is at least as hard as the problem of determining whether a line traversal exists for a set of 3-D volumes. Note that this reasoning applies equally to orthographic or perspective projection.

Unfortunately, Amenta[3] has shown that finding a line traversal to arbitrary 3-D volumes is not an LP type problem, because it can contain  $O(n)$  solutions that are disconnected in the space of all possible lines, for  $n$  3-D volumes. This implies that gradient descent methods of solving the backward constraints may produce locally optimal solutions that are not globally optimal. Moreover, existing algorithms for finding line traversals are not nearly as efficient as our linear programs, and do not appear to be extendible to solving the backward constraints without adding considerably greater complexity. For example, Pellegrini[38] gives an algorithm for finding a line traversal of general polyhedra with edges that have only a constant number of different possible directions. This algorithm requires  $O(n^2 \log(n))$  time for  $n$  polygons. Solving the forward constraints for such polygons would require only  $O(n)$  expected time, because each polygon would produce a constant number of constraints, yielding a total of  $O(n)$  constraints, and because fixed dimension linear programs can be solved in  $O(n)$  expected time (Seidel[49]). And solving the full backward constraints will be much harder than solving the line traversal problem, because the image regions will not, in general, all consist of a single point, and because a full 3-D to 2-D projection must be considered.

Our result is significant because it is not merely a comment on our algorithms; it reflects on the difficulty of a basic problem in recognition. We have shown that the general problem of determining 3-D model pose in a 2-D image, when one allows for arbitrary occlusions of unknown locus, is at least as hard as a difficult computational geometry problem. Moreover, [3]'s construction uses only volumes that are lines in 3-D. This implies that the problem is difficult even when we restrict ourselves to very simple shapes.

There is, however, a special case for which the line traversal problem can be solved efficiently. This is when the set of 3-D volumes are all axial rectanguloids, (ie. their sides are aligned with the  $x$ -,  $y$ -, and  $z$ -axes). [2] gives an efficient algorithm for this case, and subsequently [35] has refined this result, showing that this problem can be solved by linear programming.

Based on [35]'s method, we can solve the backward constraints using linear programming, for the case where the model volumes are axial rectanguloids and the projection model is affine. (Notice that since we are free to change the coordinate frame of the model, in fact we only require the models to be rectanguloids with parallel sides.) To do this, it is helpful to think of affine projection in the following way. First, the set of 3-D model points are extended in any one direction to a set of parallel lines; then a 3-D affine transformation is applied to this set of lines so that they are made normal to the image plane. Each 3-D point projects to a 2-D point where its corresponding line intersects the image plane. Now we may think of the inverse of this transformation as applying a 3-D affine transformation to the set of lines normal to the

image plane, that intersect each image point. Specifically, a set of image points,  $\{p_i\}$ , are the projection of a set of model points,  $\{q_i\}$ , if and only if there exists a 3-D affine transformation so that the line through each image point,  $p_i$ , that is normal to the image plane, is transformed to include  $q_i$ . The backward constraints, then, are satisfied by an affine transformation that maps each normal line through an image point so that it intersects the corresponding model volume.

We now show how to express the constraint that an image point  $p_i \in R_i$  must be the projection of some point in an axial rectanguloid,  $V_i$ , as a linear constraint on the inverse of the 3-D affine transformation that maps the model into the scene. Suppose  $V_i$  has a lower corner of  $(x_l, y_l, z_l)$  and an upper corner of  $(x_h, y_h, z_h)$ . Then the transformed line intersects this rectanguloid if and only if there exists a point on the line whose  $x, y$  and  $z$  coordinates lie within these ranges. Consider a line normal to the  $z = 0$  image plane passing through the point  $(u, v)$ . If we transform this line by  $T$ , and parameterize it by  $\tau$ , then a point on this transformed line has the coordinates:

$$(t_{11}u + t_{12}v + t_x + t_{13}\tau, t_{21}u + t_{22}v + t_y + t_{23}\tau, t_{33}\tau).$$

So the backward constraints are satisfied if and only if it is possible to satisfy the following inequalities:

$$\begin{aligned} x_l &\leq t_{11}u + t_{12}v + t_x + t_{13}\tau \leq x_h \\ y_l &\leq t_{21}u + t_{22}v + t_y + t_{23}\tau \leq y_h \\ z_l &\leq t_{33}\tau \leq z_h. \end{aligned}$$

Assume now that  $t_{13}, t_{23} > 0$  (in practice we must run four linear programs to account for the different possible signs of  $t_{13}$  and  $t_{23}$ ). Then we have:

$$\begin{aligned} \frac{x_l - t_{11}u - t_{12}v - t_x}{t_{13}} &\leq \tau \leq \frac{x_h - t_{11}u - t_{12}v - t_x}{t_{13}} \\ \frac{y_l - t_{21}u - t_{22}v - t_y}{t_{23}} &\leq \tau \leq \frac{y_h - t_{21}u - t_{22}v - t_y}{t_{23}} \\ \frac{z_l}{t_{33}} &\leq \tau \leq \frac{z_h}{t_{33}}. \end{aligned}$$

These inequalities are satisfied if and only if every quantity constrained to be less than  $\tau$  is smaller than every quantity constrained to be bigger than  $\tau$ . Ie. we can replace them with nine inequalities (three of which are trivial) that do not mention  $\tau$  at all. These inequalities appear non-linear, but we can linearize them with a change of variables. Let:

$$s_{11} = \frac{t_{11}}{t_{13}}, \quad s_{12} = \frac{t_{12}}{t_{13}}, \quad s_{13} = \frac{1}{t_{13}}, \quad w_x = \frac{t_x}{t_{13}}, \quad s_{21} = \frac{t_{21}}{t_{23}}, \quad s_{22} = \frac{t_{22}}{t_{23}}, \quad s_{23} = \frac{1}{t_{23}}, \quad w_y = \frac{t_y}{t_{23}},$$

and we get six non-trivial constraints such as:

$$x_l s_{13} - s_{11}u - s_{12}v - w_x \leq y_h s_{23} - s_{21}u - s_{22}v - w_y.$$

These inequalities are homogeneous, reflecting the fact that some components of the transformation are redundant. We can get around this by setting  $t_{33} = 1$ , creating difficulties only in a degenerate case. We can therefore express the backward constraints for rectanguloids with parallel sides as a set of linear inequalities, and solve them using linear programming, as we did the forward constraints. For the case of non-rectanguloid model volumes, we can approximate the backward constraints by replacing each model volume with the axial rectanguloid that bounds it.

With this approximation the backward constraints require that the bounding rectanguloid of each model volume project into the image so that it completely contains the corresponding image region. These provide correct constraints on the transformation, since if a volume projects into the image so that it contains a corresponding region, so must its bounding rectangle. The accuracy of the rectangular backward constraints will depend on the specific shape of the model volumes. However, we note that without loss of generality we may apply an arbitrary 3-D affine transformation to the set of model volumes before bounding them, to improve our approximation. In Section 3 we will describe the consequences of this approximation in more detail, showing that in some cases of interest we can find the correct model pose in spite of this approximation.

We are required to run four linear programs, which each correspond to a different visual aspect of the rectanguloids. It is not the case, however, that we can use more complex shapes to approximate the volumes, running a different linear program for each aspect, since the reduction to linear programming depends on decoupling the  $x$  and  $y$  coordinates, not on treating each aspect separately. This is clear also from [3]’s proof that the general problem is not an LP-type problem.

### 3 Uniqueness of Solutions

We have described two different types of constraints that we may use to efficiently determine model pose. The forward constraints are correct when there is only self-occlusion, or occlusion whose location is known. The backward constraints are correct even in the presence of unknown occlusion. However, these constraints are not complete: the forward constraints do not express the constraints that each model volume should explain every point in each matching image region, while the backward constraints only bound the true set of backward constraints, through the use of axial rectanguloids. We will now none-the-less show that, although incomplete, these constraints are sufficient in many realistic situations to correctly determine model pose.

We first consider the performance of the forward constraints, in the presence of only self-occlusion. We show that, in general, when we have matched three or four volumes, we may determine the correct model pose with linear programming, while, in general, two matches are not sufficient to determine pose uniquely. Next we consider the performance of the backward constraints when the model volumes are each planar (but not mutually coplanar). In this case, some subset of the exact constraints will be present, and we show that in many cases these can

determine the correct solution. We also point out that when we apply the rectangular back constraints to curved 3-D volumes they can only produce an approximate solution, and not an exactly correct one. Finally, we point out that the forward constraints with known occlusions can also produce the correct model pose. Throughout our discussion we will mention situations in which degenerate solutions may be found instead of the correct ones. By restricting the solution to represent a rigid transformation we can in many cases use a degenerate solution to determine the correct one. This will be shown in Section 4.

### 3.1 The Forward Constraints

In this section, we suppose that  $T$  is a 3-D to 2-D affine transformation that maps the model volumes to inside the corresponding image regions. We then ask under what circumstances  $T$  is the unique transformation that satisfies the forward constraints. When  $T$  is unique, this means that using a linear program to find a transformation satisfying the forward constraints must in fact produce  $T$ , the correct solution. We had previously ([7]) reported preliminary results on this problem for the special case of planar model volumes that are not mutually coplanar. We now extend those results to the case of fully three-dimensional model volumes.

We first prove the following useful lemma:

**Lemma 1:** *Suppose that  $T'$  is a transformation that also satisfies the forward constraints. Then for each model volume  $V_i$  and image region  $R_i$ , there exists a point,  $\vec{p}_i \in V_i$  such that  $T\vec{p}_i = T'\vec{p}_i$ .*

**Proof:** Throughout this section, we will assume that the *contour generator* of  $V_i$  is a 1-D curve on  $V_i$  that projects to the 1-D boundary of  $R_i$ . This will be true for smooth, generic, convex volumes in general position. Our reasoning can be readily extended to other cases, however we will not consider those in this paper to simplify our arguments. Clearly we can construct a 2-D surface, call it  $S_i$ , such that  $S_i$  is bounded by the contour generator, such that  $S_i \subseteq V_i$  and such that  $TS_i = R_i$ . Therefore,  $T$  will define a continuous one-to-one mapping between  $S_i$  and  $R_i$ . So even though  $T$  is not in general invertible, we may invert  $T$  when we restrict its domain to  $S_i$ . Let  $\bar{T}$  denote the restriction of  $T$  to this domain, and let  $\bar{T}^{-1}$  denote its inverse, a mapping from  $R_i$  to  $S_i$ . Because  $T'$  satisfies the forward constraints,  $T'S_i \subseteq R_i$ . Therefore  $T'\bar{T}^{-1}$  defines a continuous mapping from  $R_i$  into  $R_i$ . Brouwer's fixed point theorem, a basic result in functional analysis (see, for example, Conway[15]), tells us that since  $R_i$  is convex, and hence topologically equivalent to a disc, such a mapping must have a fixed point. That is, there exists some point  $\vec{q}_i \in R_i$  such that  $\vec{q}_i = T'\bar{T}^{-1}\vec{q}_i$ . Therefore,  $T(\bar{T}^{-1}\vec{q}_i) = T'(\bar{T}^{-1}\vec{q}_i)$ , proving the lemma.  $\square$

Using this lemma, we may now show the following:

**Theorem 2:** *Suppose the transformation  $T$  maps the four model volumes,  $V_0, V_1, V_2, V_3$  to the four image regions,  $R_0, R_1, R_2, R_3$ , and there does not exist a plane that intersects all four volumes. Then,  $T$  is the only transformation that satisfies the forward constraints.*

**Proof:** Let  $T'$  map the four model volumes  $V_0, \dots, V_3$  to inside the corresponding image regions  $R_0, \dots, R_3$ . By lemma 1 for every volume  $V_i$ ,  $i = 0, \dots, 3$  there exists a point  $\vec{p}_i \in V_i$  such that  $T'(\vec{p}_i) = T(\vec{p}_i)$ . Since there exists no plane that intersects all four volumes the points  $\vec{p}_0, \dots, \vec{p}_3$  are not coplanar. Consequently, since correspondences of four non-coplanar points determine a 3-D to 2-D affine transformation uniquely then  $T = T'$ .  $\square$

We now show that typically, only three matches are required to uniquely determine the pose. In [6] we have proven the following lemma:

**Lemma 3:**  *$T$  is uniquely determined for the set of volumes  $V_i$  if and only if it is uniquely determined for the set of volumes  $QV_i$ , where  $Q$  is any 3-D affine transformation.*

This tells us that we may, without loss of generality, place the model volumes in any affine reference frame. This can significantly simplify our reasoning, since it allows us to assume without loss of generality that  $T$  is the identity transformation, plus orthographic projection. We now suppose that we have matched three model volumes to three image regions, and that the image regions are not all intersected by a single line. Further, we suppose that the image regions were produced by applying the transformation,  $T$ , to the model volumes, and that a different transformation,  $T' \neq T$ , also satisfies the forward constraints. Then there exist three non-collinear model points,  $\vec{p}_0, \vec{p}_1, \vec{p}_2$  such that  $T\vec{p}_0 = T'\vec{p}_0$ ,  $T\vec{p}_1 = T'\vec{p}_1$ , and  $T\vec{p}_2 = T'\vec{p}_2$ . Lemma 3 tells us that we may assume, without loss of generality, that these three points are fixed under the transformation  $T$ , and therefore also under  $T'$ . This means that the entire  $z = 0$  plane is fixed under these two transformations. Further, without loss of generality, we may assume that  $T(0, 0, 1) = (0, 0)$ . Therefore, we may write:

$$T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

and

$$T' = \begin{pmatrix} 1 & 0 & k_1 \\ 0 & 1 & k_2 \end{pmatrix},$$

where either  $k_1$  or  $k_2$  is non-zero.

By choosing this affine reference frame, we have constructed things so that each contour generator,  $s_i$ , projects orthographically to form the boundary of the corresponding region. Let  $\vec{p}$  be an arbitrary point on  $V_0$ 's contour generator, with coordinates  $(x, y, z)$ . Then  $T\vec{p} = (x, y)$ , and  $T'\vec{p} = (x, y) + z(k_1, k_2)$ . This tells us that  $T'$  maps the contour generator so that it is displaced from the region boundary in either the direction  $(k_1, k_2)$  or  $-(k_1, k_2)$ , depending on whether the point is above or below the image plane.

We may use this fact to place constraints on the direction of the contour generator's tangent. If one of the contour generators is either entirely above, or entirely below the image plane, then clearly  $T'$  will map some of these points outside the corresponding region, violating our assumptions.

Next, suppose that the image plane intersects the contour generator in at least two discrete points, but not in an entire subcurve of the contour generator. Let  $R_i \cap s_i$  be the points  $\vec{p}_i^1, \vec{p}_i^2$ . Let the tangent to  $R_i$  at  $\vec{p}_i^j$  be  $\vec{w} = (w_x, w_y, 0)$ . Let the tangent to  $s_i$  at the point  $\vec{p}_i^j$  have the direction  $\vec{v}$ . Then the directions of  $\vec{w}, T\vec{v}$  and  $T'\vec{v}$  must all be the same. That is true because if a point in the model projects to the boundary of the image region, and the tangents of the region point and the projected model point differ, then the projected model volume will not be contained in the image region. So, since  $\vec{w} = T\vec{v}$ , we must have  $\vec{v} = (w_x, w_y, v_z)$  for some  $v_z$ . The points  $\vec{p}_i^j$  are also fixed under  $T'$ , since they lie in the  $z = 0$  plane which is fixed under  $T'$ . Therefore, the tangent to  $T'V_i$  at  $\vec{p}_i^j$  is  $(w_x + k_1v_z, w_y + k_2v_z)$ . Again, the condition that  $T'V_i \subseteq R_i$  implies that  $T'\vec{v}$  must have the same direction as  $\vec{w}$ . Since either  $k_1 \neq 0$  or  $k_2 \neq 0$  it follows that the directions of  $(w_x, w_y)$  and  $(k_1, k_2)$  must be parallel. Therefore, the tangents to each region  $R_i$  at a point  $\vec{p}_i^j$  must all be parallel to  $(k_1, k_2)$ , and so they must all be parallel to each other.

If the image plane intersects the contour generator in an entire curve, we may apply the same reasoning to the two end points of the curve. In the special case that the entire contour generator lies in the image plane, then that volume will not constrain  $k_1$  and  $k_2$ , for  $T'$  close to  $T$ .

In sum, we can see that  $T$  is non-unique, and an appropriate  $T' \neq T$  exists if and only if we can find a plane such that for all contour generators, either this plane completely contains the contour generator or it intersects the contour generator in two points, such that the projection of the tangents to these contour generators orthographically onto the plane are all parallel. Simple variable counting now tells us that for general shapes, in general position, this will not be possible. Given a set of model volumes and a particular viewpoint, the contour generators will be fixed. First, for general objects the contour generator will not be planar. Next, we have three degrees of freedom in choosing a plane to intersect the contour generators. Each plane will determine the direction of six tangent vectors projected into that plane. For all these tangents to be parallel provides five degrees of constraint; however we have only three degrees of freedom available to satisfy these constraints. Therefore, in general, given a set of model volumes and a view of them, there will be a unique transformation that satisfies the forward constraints, provided that the image regions cannot be intersected by a single line.

We now consider the case in which only two model volumes are matched to two image regions. We do not provide necessary and sufficient conditions for these to determine a unique transformation. However, we do note that any two model volumes may be viewed so that their corresponding image regions intersect. When this occurs, the forward constraints will be satisfied by any transformation that shrinks the model volumes to a very small size, and projects them inside this region of intersection. Therefore, any two model volumes will lead to a non-unique solution when viewed from a significant range of viewpoints.

### 3.2 The Rectangular Backward Constraints

We now address the problem of determining when the rectangular backward constraints are sufficient to correctly determine the pose of a model. Recall that the backward constraints bound each model volume with an axial rectanguloid, and then require a valid transformation to map this rectanguloid into the image so that it completely contains the corresponding image region.

Let  $C_i$  denote the axial rectanguloid that bounds the volume  $V_i$ . Note that the contour generator for  $C_i$  will consist of six line segments, or four line segments for those special viewpoints in which only one face of the rectanguloid is visible. This means that the projection of  $C_i$  will be a six (or four) sided convex polygon. The rectangular backward constraints require that  $R_i \subseteq TC_i$ . Since  $V_i \subseteq C_i$ , the correct transformation will satisfy these constraints.

In general,  $TV_i$  may lie completely inside  $TC_i$ . In fact,  $TV_i$  and  $TC_i$  will touch only in the case where  $V_i$  touches one of the line segments that form the boundary of  $C_i$  (or for special viewpoints in which a side of  $C_i$  projects to a line segment in the image). For a general smooth 3-D volume, no bounding rectanguloid (or polygon of any sort) will touch the volume in one of its edges, since this would require a discontinuity in the volume. On the other hand, convex planar volumes can always be oriented so that a bounding rectanguloid is also planar, and touches them in at least four points. Therefore, cases of interest exist in which either  $TV_i$  is always completely inside  $TC_i$  or in which  $TV_i \cap TC_i$  contains at least four points.

Suppose first that  $R_i = TV_i$  lies entirely inside  $TC_i$ , for all volumes  $V_i$ . In this case, it is obvious that any small perturbation to the transformation  $T$  will not violate the rectangular backward constraints. Therefore, in such a situation the backward constraints will not uniquely determine the correct pose, although they may still produce a good approximation to this pose.

This is not surprising; after all the rectangular backward constraints involve an approximation, and so one expects them to produce poses with some error. Surprisingly, though, we now show that for an interesting class of volumes the rectangular backward constraints will produce exactly the correct answer. Suppose that the volumes are all planar, and that they each lie in either of two different planes. This is a common occurrence if the volumes are either surface markings on two different faces of a polyhedron, or on the walls or ceiling of a room. We show that such volumes can lead to a unique solution to the backward constraints.

We begin our analysis by supposing that at least two of the model volumes,  $V_0, V_1$  are planar and lie in the same plane. Without loss of generality, we assume that this is the  $z = 0$  plane. Note that we are free to preprocess the model volumes with an affine transformation to achieve this before approximating them with rectangles. The volumes will then be approximated by 2-D rectangles in this plane, which we call  $C_0$  and  $C_1$ . Each side of each rectangle will touch the corresponding volume in at least one point. These will be the points with the highest and lowest values in their  $x$  and  $y$  coordinates. We will call these points  $\vec{x}_{0,l}, \vec{x}_{0,h}, \vec{y}_{0,l}, \vec{y}_{0,h}, \vec{x}_{1,l}, \vec{x}_{1,h}, \vec{y}_{1,l}, \vec{y}_{1,h}$ , where, for example,  $\vec{x}_{0,h}$  is the point in  $V_0$  with the highest  $x$  coordinate (see Figure 4). Note that if there is a line segment on the boundary of one of the volumes with constant  $x$  or  $y$  values we can pick any of these points, or we can easily extend

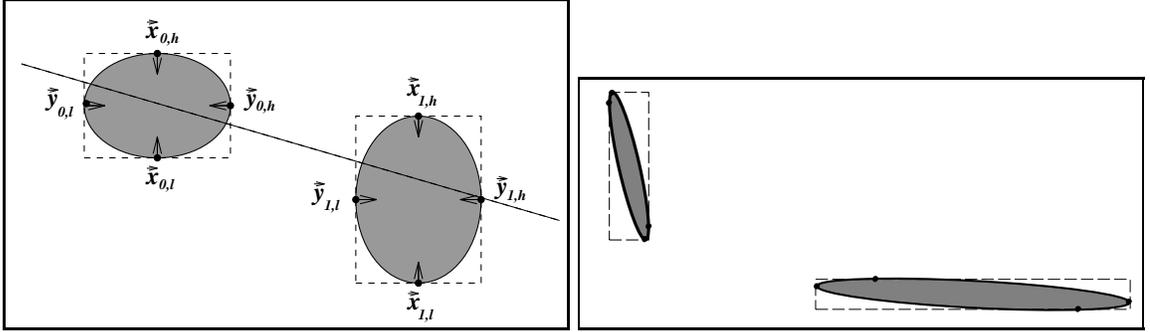


Figure 4: Two coplanar volumes  $V_0$  and  $V_1$  bounded with flat rectanguloids  $R_0$  and  $R_1$ . The points  $\vec{x}_{0,l}, \vec{x}_{0,h}, \vec{y}_{0,l}, \vec{y}_{0,h}, \vec{x}_{1,l}, \vec{x}_{1,h}, \vec{y}_{1,l}, \vec{y}_{1,h}$  are the contact points of the volumes and the rectanguloids. In the case on the left, the two pairs of points  $\vec{x}_{0,l}, \vec{x}_{1,l}$  and  $\vec{x}_{0,h}, \vec{x}_{1,h}$  are linearly separable, as is indicated by the dashed line. On the right, the volumes lead to a unique solution in the  $z = 0$  plane.

the reasoning given below to include this case. Also, one point may be extremal in both the  $x$  and  $y$  direction; this does not effect the argument given below. At these eight points, the backward constraints will require that the corresponding region point found in the image will lie on the appropriate side of the projection of a horizontal or vertical line passing through that point in the model. Therefore, we will have eight constraints that are not approximate, but that are subsets of the constraints we would have if we could apply the backward constraints exactly. A transformation infinitesimally different from the correct one will not violate any of the other backward constraints, but it may violate these ones.

Now we consider when these eight constraints will suffice to uniquely determine that portion of the projection that effects the  $z = 0$  plane. From Lemma 3 we may assume, without loss of generality, that  $T$  leaves the  $z = 0$  plane fixed<sup>1</sup>; we must then determine the circumstance under which there exists a different transformation,  $T'$ , which has a different effect on the  $z = 0$  plane while respecting the approximate backward constraints. Note that the  $x$  coordinates of points in the  $z = 0$  plane are effected only by  $t'_{11}, t'_{12}$  and  $t'_x$ , while the  $y$  coordinates are changed by  $t'_{21}, t'_{22}$  and  $t'_y$ . Since the  $\vec{x}_{0,h}, \vec{x}_{0,l}, \vec{x}_{1,h}, \vec{x}_{1,l}$  points are constrained only in the  $x$  direction, and the corresponding  $y$  points are constrained only in the  $y$  direction, we may consider these two sets of points separately. A non-unique solution exists if and only if either there exist values of  $t'_{11}, t'_{12}$  and  $t'_x$  that map the points  $\vec{x}_{0,h}, \vec{x}_{0,l}, \vec{x}_{1,h}, \vec{x}_{1,l}$  within the minimum and maximum  $x$  values of their bounding rectanguloids, or if there similarly exist values of  $t'_{21}, t'_{22}$  and  $t'_y$  that map the  $y$  points in the appropriate  $y$  directions.

The transformation  $T'$  will change the  $x$  coordinate of the point  $(x, y)$  by the amount:

$$T'_x(x, y, 0) - x = (t'_{11} - 1)x + t'_{12}y + t'_x.$$

Therefore, the line

$$(t'_{11} - 1)x + t'_{12}y + t'_x = 0$$

<sup>1</sup>This point is somewhat subtle. We may assume this because the bounding rectanguloids can be axial to any affine reference frame, which need not be the natural Euclidean reference frame in which we store the model.

will have its  $x$  coordinates fixed (note that in the case of pure translation we may think of the equation  $t'_x = 0$  as describing a vertical line at  $x = \infty$ ). On one side of this line, the value of  $(t'_{11} - 1)x + t'_{12}y + t'_x$  is positive, on the opposite side this value is negative. This means that either all points will shift their  $x$  coordinate towards this fixed line, or all points will shift away from it. Therefore, the only way such a transformation can satisfy the backward constraints is if we can draw a line that separates the points  $\vec{x}_{0,l}, \vec{x}_{1,l}$  from the points  $\vec{x}_{0,h}, \vec{x}_{1,h}$ . In that case any such separating line can be the fixed line, and  $T'$  can map all points towards that line, in the  $x$  direction. It may or may not be possible to find such a separating line, depending on the configuration of the model.

Exactly the same reasoning holds for the points  $\vec{y}_{0,l}, \vec{y}_{0,h}, \vec{y}_{1,l}, \vec{y}_{1,h}$ . Therefore, there exist models for which the approximate backward constraints uniquely determine the transformation of the  $z = 0$  plane, as illustrated in Figure 4. While it is someone difficult to find such a configuration in two model volumes, similar reasoning holds when more than two volumes are coplanar, except that now a larger set of points must be linearly separable for a transformation that satisfies the backward constraints to be non-unique.

Suppose now that the model contains a set of volumes for which the backward constraints uniquely determine the transformation of the  $z = 0$  plane. This tells us that any transformation,  $T' \neq T$  satisfying the backward constraints must have the form:

$$T'\vec{p} = \begin{pmatrix} 1 & 0 & t'_{13} \\ 0 & 1 & t'_{23} \end{pmatrix} \vec{p},$$

with either  $t'_{13} \neq 0$  or  $t'_{23} \neq 0$ . We now ask when an additional planar model volume, not in the  $z = 0$  plane, will suffice to completely determine the model-to-image transformation. Without loss of generality we may assume that this volume lies in the  $y$ - $z$  plane, so that it is bounded by a rectangle in this plane, and has four extremal points in the  $y$  and  $z$  directions. Furthermore, we may assume that the transformation  $T$  has the form:

$$T\vec{p} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \vec{p}.$$

That is,  $T$  is the identity transformation when applied to the  $x$ - $y$  plane, while it maps the  $y$ - $z$  plane to the  $x$ - $y$  image plane by mapping  $(0, y, z)$  to  $(z, y)$ . Therefore,  $T$  acts also like an identity transformation on the  $y$ - $z$  plane, while converting it into the  $x$ - $y$  plane. Now consider the effect that  $T'$  has as it maps the  $y$ - $z$  plane into the  $x$ - $y$  plane. It leaves the  $z = 0$  line fixed, so  $T'$  can consist only of a contraction of the  $y$ - $z$  plane towards this line, along with an interchange of coordinates. If both extremal points of the model volume in the  $z$  direction are on the same side of the line  $z = 0$ , then any contraction or expansion towards or away from that line will violate one of the constraints imposed by the bounding rectanguloid. That is, once the backward constraints imply a unique solution within one plane, they fail to produce a unique solution only when this plane intersects the remaining volumes. This will never happen if the volumes are surface markings on a convex polyhedra, as for example, when they lie on the walls or ceiling of a room.

We have now shown that, though approximate, the rectangular backward constraints may still uniquely determine the correct model to image transformation in situations of real interest. At the same time, unlike our previous 2-D formulation of the backward constraints, we can now integrate information from planar volumes that are not coplanar, or from non-planar volumes.

### 3.3 The Forward Constraints, with Known Occlusion

The forward constraints are suitable when there is only self-occlusion; the backward constraints hold even in the presence of occlusion. We now consider one other possibility, that some model volumes may be partially occluded in the image, but that the extent of this occlusion is known. That is, we assume that some portion of boundary of the image regions may be caused by an occluding object in front of the volume, rather than the boundary of the volume itself, but that we have identified those portions of the region boundaries. This is a situation of considerable interest; for example, if we identify several different parts of an object we may be able to determine that some of these parts lie in front of and occlude others (e.g., by identifying concave sections in the boundaries of a region corresponding to a convex volume). Also, it is well-known in the psychology literature that the knowledge that certain boundaries are due to occlusions can greatly assist human perception (Rock[46]).

We now point out that our previous discussion of the backward constraints also provides an example of a situation in which the forward constraints, with known occlusion, can lead to a unique transformation. When we use the forward constraints with known occlusions we are making use of a subset of the constraints that would be available in the absence of occlusion. As we have described, the approximate backward constraints can include a small subset of the complete backward constraints. For planar volumes, those points that are extremal in the axial directions lead to tight constraints on the model pose. Suppose now that the same extremal points that we make use of in section 3.2 are visible in the image. In this case, exactly the same reasoning applies to show that a unique transformation will satisfy the forward constraints. Consequently, if any but these extremal points are occluded, we may still apply the forward constraints to find the correct transformation. This provides an example that shows that even with a large amount of occlusion, the forward constraints can produce the correct transformation, provided that this occlusion is identified.

In general the forward constraints with known occlusions will be much more effective in determining a unique solution than will the approximate backward constraints, since every unoccluded point on the boundary of the image regions will provide a tight constraint on the transformation. It is, however, beyond the scope of this paper to fully characterize when these constraints lead to a correct solution.

## 4 Recovering from Degeneracies

We now consider how we may handle certain situations in which the model volumes lead to a non-unique solution. The most common cause of this occurs when the model itself is in a

sense degenerate, and so a 3-D affine transformation that satisfies analogs to the forward and backward constraints may be found. However, we will show that in this case we can often reconstruct the correct transformation, undoing the effects of this degeneracy.

Suppose there exists a 3-D to 3-D affine transformation,  $A \neq I$ , such that  $AV_i \subseteq V_i$  for all  $i$ . In this case, the forward constraints can almost never lead to a unique solution. If the correct transformation is  $T$ , then the transformation  $TA$  will clearly also satisfy the forward constraints. Moreover,  $TA \neq T$  except for special cases when projection completely removes the effects of  $A$ .

If the model contains three distinct volumes, any such transformation must contain three fixed points. If the volumes are not traversed by a single line,  $A$  must consist of a fixed plane, with a contraction in some direction toward that plane.

Similarly, a degeneracy will almost always occur in the backward constraints when there is an affine transformation  $A$  such that  $V_i \subseteq AC_i$ . This typically occurs when a plane exists that intersects the  $C_i$  rectanguloids on sides that can be divided into two parallel sets. In this case  $A$  can be the affine transformation that leaves this plane fixed, and expands the model in the direction shared by the rectanguloid sides intersected by the plane. For example, suppose all the rectanguloids have minimum  $x$  coordinates less than zero, and maximum  $x$  coordinates greater than zero. In this case, the  $x = 0$  plane will intersect each  $C_i$  in two sides parallel to the  $y = 0$  plane and in two sides parallel to the  $z = 0$  plane. In this case, we may expand the model rectanguloids in the  $x$  direction (which is the intersection of the  $y = 0$  plane and the  $z = 0$  plane) so as to leave the  $x = 0$  plane fixed. This transformation has the form:

$$A\vec{p} = \begin{pmatrix} 1 + \epsilon & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \vec{p}$$

for any  $\epsilon > 0$ . In this example,  $V_i \subseteq C_i \subseteq AC_i$ . In such a situation, whenever  $T$  satisfies the backward constraints, so will  $TA$ .

However, since the model is accessible to us in advance, we may detect when such degeneracies occur, and undo their effect. To determine the possibility of such a degeneracy, we may apply the forward or backward constraints for the case of 3-D affine transformations. For the forward constraints we compare the model to itself. For the backward constraints we compare the model volumes to their bounding rectanguloids. Since 3-D affine transformations are linear, these constraints can be solved for using linear programming, and will reveal the presence of a transformation such as  $A$ . However, in the experiments described below, we have simply detected this possibility by hand.

If a model can be contracted or expanded perpendicularly in a single direction, we first preprocess the model so that this direction is aligned with one of the axes. Without loss of generality, assume this is the  $x$ -axis. Let the contraction/expansion transformation applied to the model be given by

$$A\vec{p} + \vec{b} = \begin{pmatrix} 1 + \epsilon & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \vec{p} + \begin{pmatrix} a_x \\ 0 \\ 0 \end{pmatrix}.$$

Notice that we must allow for an unknown translation of  $a_x$  in the  $x$  direction, as a part of the contracting/expanding transformation. This is because, although we assume that contraction or expansion is in the  $x$  direction, and that there is a fixed plane perpendicular to the direction, we do not assume that this is the  $x = 0$  plane. In some cases there will not be a unique plane that might be the fixed plane, there might be a family of parallel planes any one of which might have been fixed in the contraction/expansion of the model. Now, suppose the image is obtained by applying a rigid transformation to the model followed by a scaled orthographic projection. The imaging process can be written as

$$S\vec{p}' + \vec{t} = \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \end{pmatrix} \vec{p}' + \begin{pmatrix} t_x \\ t_y \end{pmatrix},$$

where the entries  $s_{ij}$  are the first two rows of a scaled rotation matrix. Our solution methods, then, will produce a transformation  $T$  that is a composition of the two transformations, as follows:

$$T\vec{p} = S(A\vec{p} + \vec{b}) + \vec{t} = \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \end{pmatrix} \left( \begin{pmatrix} 1 + \epsilon & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \vec{p} + \begin{pmatrix} a_x \\ 0 \end{pmatrix} \right) + \begin{pmatrix} t_x \\ t_y \end{pmatrix}.$$

Given the transformation,  $T$ , we wish to recover the matrix  $S$  that indicates the true scaled orthographic projection of the model. This is easily done, since

$$\begin{aligned} s_{12} &= t_{12} & s_{22} &= t_{22} \\ s_{13} &= t_{13} & s_{23} &= t_{23}. \end{aligned}$$

We may then readily determine the values of  $s_{11}, s_{21}$  which will satisfy the rigidity constraints of the matrix, so that:

$$\begin{aligned} s_{11}s_{21} + s_{12}s_{22} + s_{13}s_{23} &= 0 \\ s_{11}^2 + s_{12}^2 + s_{13}^2 &= s_{21}^2 + s_{22}^2 + s_{23}^2 \end{aligned}$$

Thus we can recover the scaled rotation matrix and also the  $y$  translation that produced the image regions. We cannot directly recover the translation  $a_x$ . However, once we have determined the rest of the transformation, it is easy to determine the appropriate translation. For example, we could run linear programming again, allowing for only an  $x$  translation.

## 5 Experiments

We now present some experiments that demonstrate the feasibility of our approach to recognition. These experiments will provide useful information about the accuracy of the poses that we can recover using both the forward and the approximate backward constraints.

In these experiments a model was first constructed by hand, using images of the object and knowledge of its structure. Then, the Canny edge detector ([13]) was run on a new image of

the object. We automatically extracted sets of edges that formed salient convex groups, using the grouping system described in [28]. The localization of groups in the image will therefore contain errors due to running a real edge detector and grouping system on real images. A subset of these groups were then extracted and matched by hand to groups in the model. See Nayar and Bolle[37] for one suggestion about how to use intensity information to match such regions automatically. Our system then used these matches to determine the model poses shown here.

In the first set of experiments we use a black box with different shaped regions painted on the side. Although this object is somewhat artificial, it allows us to experiment with a variety of different conditions. In figures 5 to 11 we show the poses derived by the forward and backward constraints using different combinations of regions. This shows that both sets of constraints are able to produce accurate model poses when at least three regions correspondences are used. It also illustrates how the poses degrade as we use fewer regions, and as the amount of occlusion increases. Of course the forward constraints are especially vulnerable to unidentified occlusions.

In figures 12, 13 and 14 we demonstrate the method discussed in section 4 for determining the correct scaled orthographic projection from a degenerate transformation. The model regions used in these examples are coplanar, and thus lead to a degeneracy. Note, however, that degeneracies can occur even when 3-D model volumes are used.

Figure 15 shows the forward and backward constraints being used to recognize two other objects. These experiments show that even with noise due to the edge detection and grouping process, we can use region matches to accurately determine pose without explicit correspondences between local features such as points or line segments.

Finally, figures 16 through 19 show the results of our algorithms for a synthetic model composed of 3-D volumes. The object is composed of seven volumes: the four legs, the body, the neck and the head. That is each volume is a part; the faces of volumes are not used as separate parts. We segment and match these parts perfectly in the images. Being synthetic, this object is simple to model accurately. But moreover, since the only source of error is digitization in the projection, any inaccuracies that we find are due to limitations of the methods proposed. Figure 17 shows that the backward constraints can produce very accurate poses, in spite of the approximations made by taking bounding rectanguloids. Note that none of the model parts are at all rectangular; the legs and neck, for example are nine and seven sided polygons, respectively. Figure 18 shows that the forward constraints do indeed produce accurate results when only three volumes are matched. Finally, figure 19 shows the performance of the methods when some model volumes are partially occluded, in this case by other parts of the object.

## 6 Conclusion

Recognition of 3-D objects in 2-D images has been hampered by the difficulty of finding representations that can faithfully model complex 3-D objects and still be used to determine pose based on their 2-D images. In this work we make use of a simple representation, which divides objects into parts and then represents each part as a volume of points. This representation

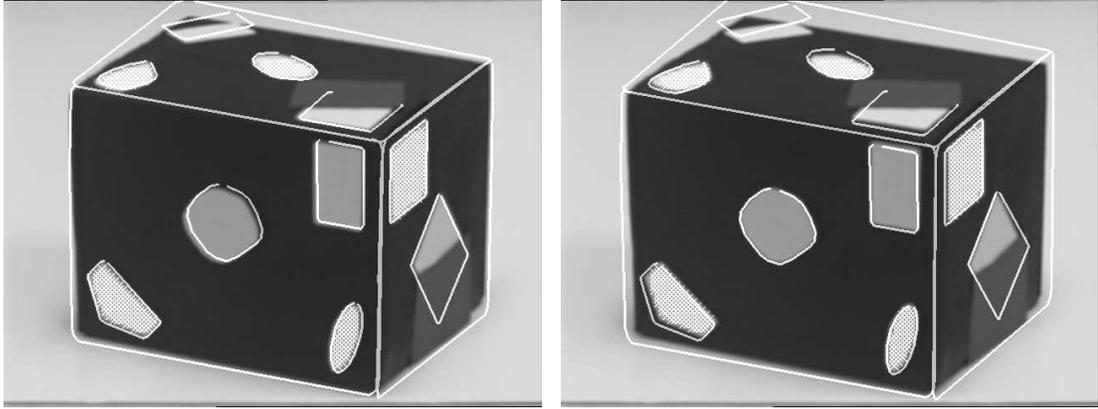


Figure 5: This shows poses derived using the forward and backward constraints. The image is shown, with the regions used to derive the pose marked with a white hatching. Superimposed over the image are the white outlines of all model groups, to indicate the pose that has been derived using the forward constraints (left) and backward constraints (right). This figure shows the poses derived using five regions.

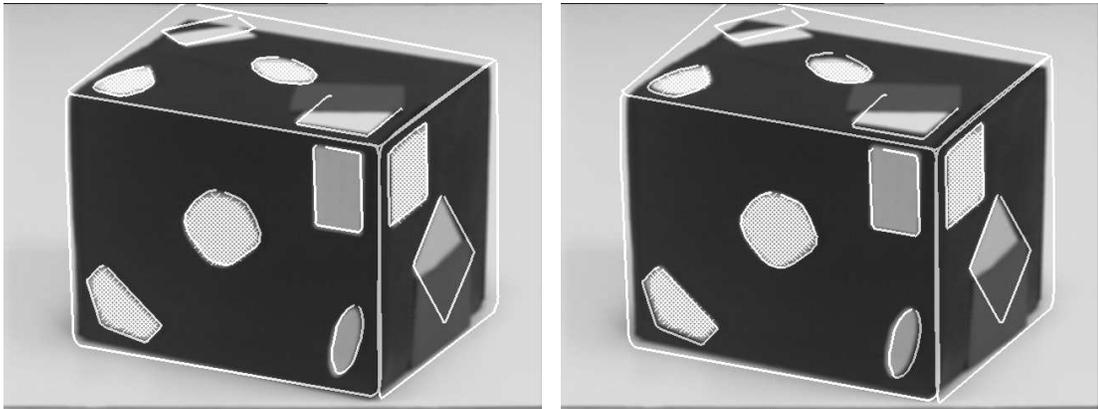


Figure 6: Another example using five regions. The pose derived using the forward constraints is on the left, and the pose from the backward constraints is on the right.

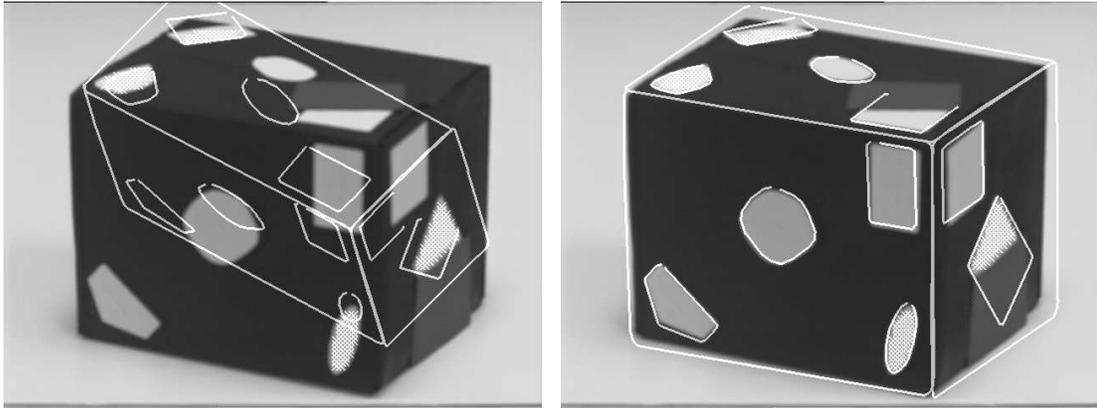


Figure 7: An example using four regions. Note that one of the regions is partially occluded, leading to a poor solution with the forward constraints (left).

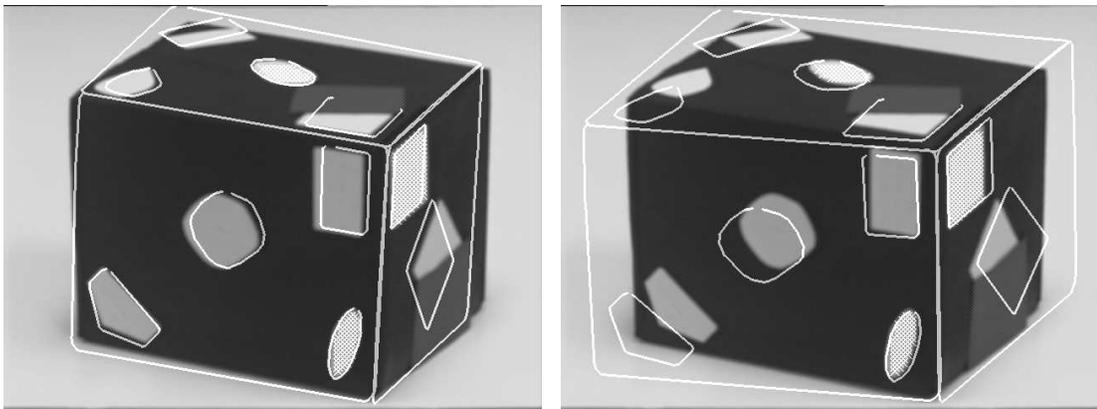


Figure 8: An example using three regions. The backward constraints (right), which are approximate, lead to a much noisier solution than the forward (left).

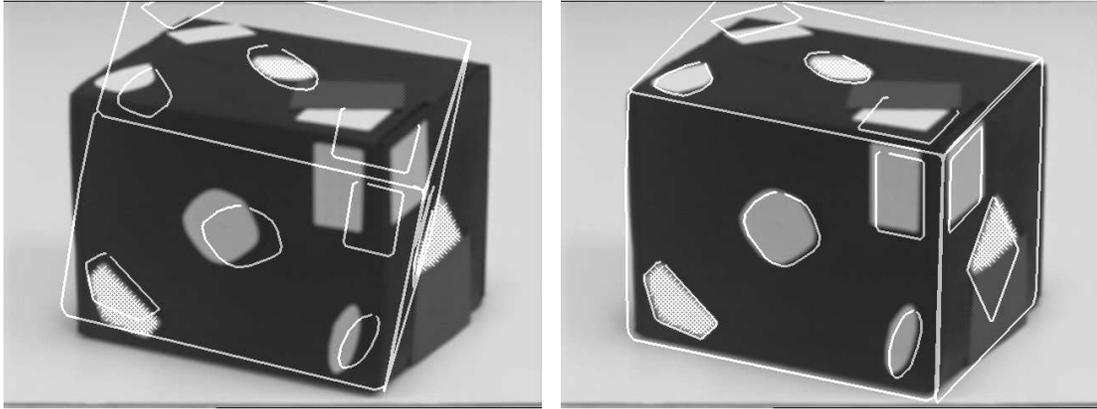


Figure 9: An example using three regions, one of them occluded. This leads to a poor solution for the forward constraints (left), which do not allow for occlusion. The backward constraints (right), produce a much more accurate pose. Note that in spite of the occlusion, the backward constraints produce a more accurate solution than they did in Figure 8 because a more stable triple of image regions are matched.

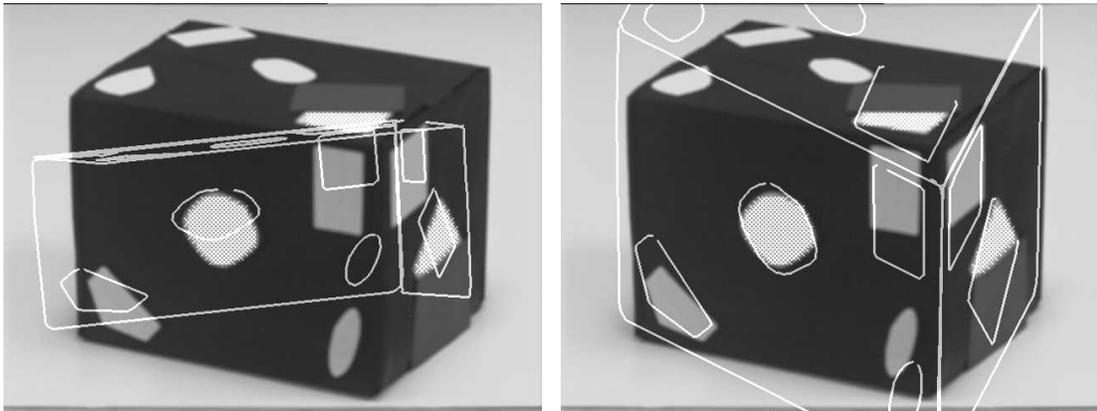


Figure 10: An example using three regions, with two of them occluded. This leads to a very poor solution for the forward constraints (left), while the solution for the backward constraints also degrades. Occlusion may cause the correct solution to become non-unique, and indeed we can see that the pose found satisfies the backward constraints very well.

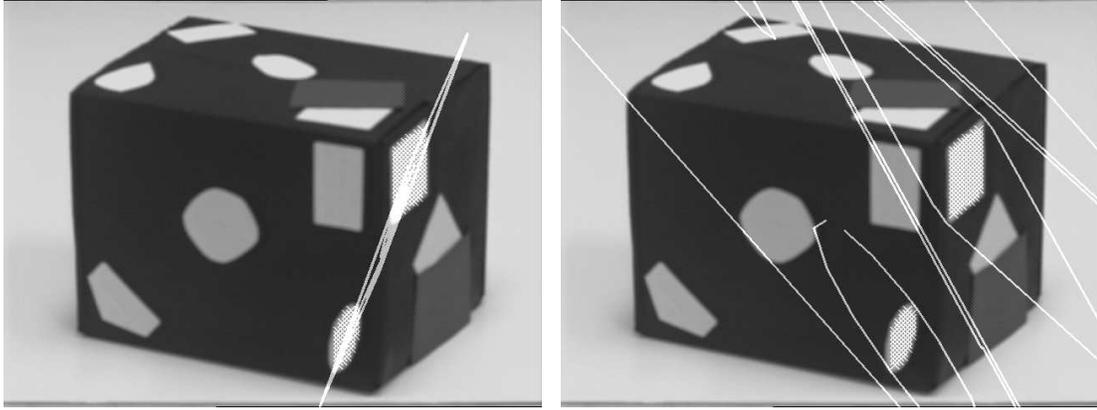


Figure 11: This example shows poses derived from two regions. We can see that these are not sufficient to determine the correct pose.

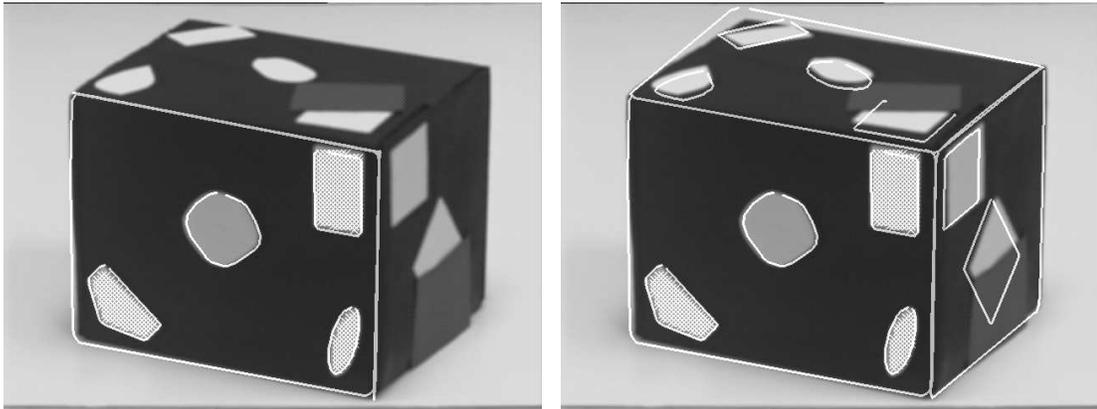


Figure 12: Here we show a set of regions that produce a degenerate solution. The forward constraints are applied. However, since the model regions matched are really coplanar, a solution is found that contracts the 3-D model into a single plane (left). Since this potential degeneracy can be detected ahead of time, we may postprocess the pose to “uncontract” it, producing a scaled orthographic transformation (right) that matches the model and image better.

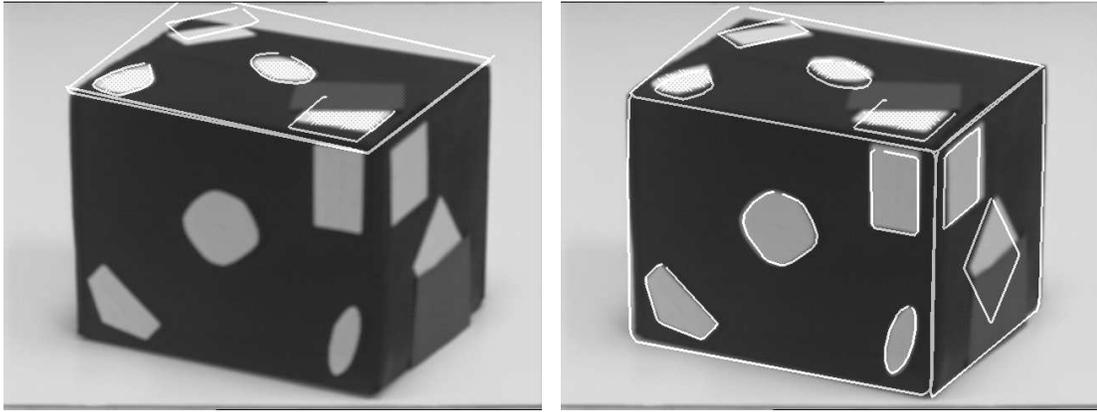


Figure 13: Similarly, we show the backward constraints applied to produce a degenerate solution (left), along with the “uncontracted” pose.



Figure 14: Similar results for a different object. The pose of the soda can found by the forward constraints (left) is significantly contracted. After “uncontracting” the pose (right) into a scaled orthographic projection, we obtain a better fit.



Figure 15: These figures show the performance of the system on more realistic objects. On the left, the system uses the forward constraints to accurately determine the pose of the soda can. Four regions are used in this case. The regions are surface markings on the cylinder of the can, and a circular region from the top of the can. On the right, the backward constraints are used to locate the pen box. Note that the soda can is occluding some of the regions used.

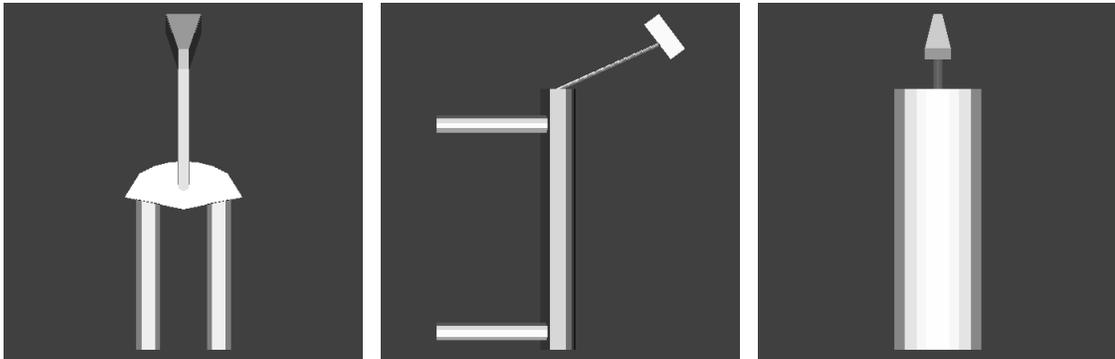


Figure 16: A simple synthetic animal, seen from along the  $x$ -axis (left)  $y$ -axis (center) and  $z$ -axis (right). Bounding boxes for the rectangular backward constraints were built with the model in this reference frame.

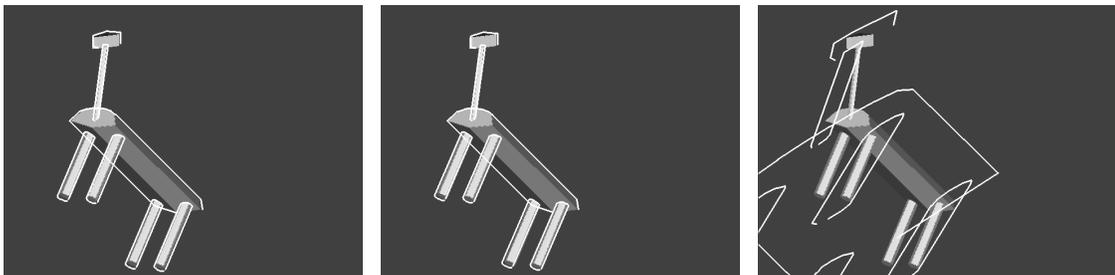


Figure 17: These figures show the rectangular backward constraints applied to a synthetic, 3-D object. We show the image used with some of the boundary of the volumes of the projected model superimposed in white. On the left, the pose found using all seven model volumes. In the center, we use four volumes: the head, the two left legs, and the front right leg. In both cases, we find accurate poses. On the right, we use only the head and the two left legs, and a poorer pose is found.

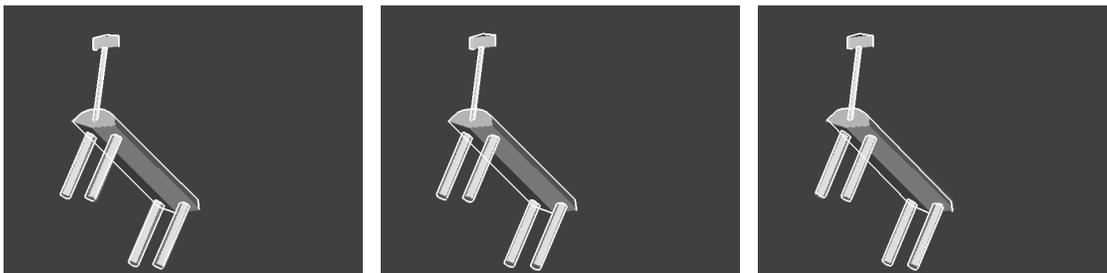


Figure 18: The results using the forward constraints using the same image and volume/region matches as in the previous figure. Since there is little occlusion of objects parts by other parts, the forward constraints produce accurate poses. When only three volumes (right) are used, they continue to produce an accurate pose. Note that the backward constraints lead to a much less accurate pose in this case, as shown in the previous figure.

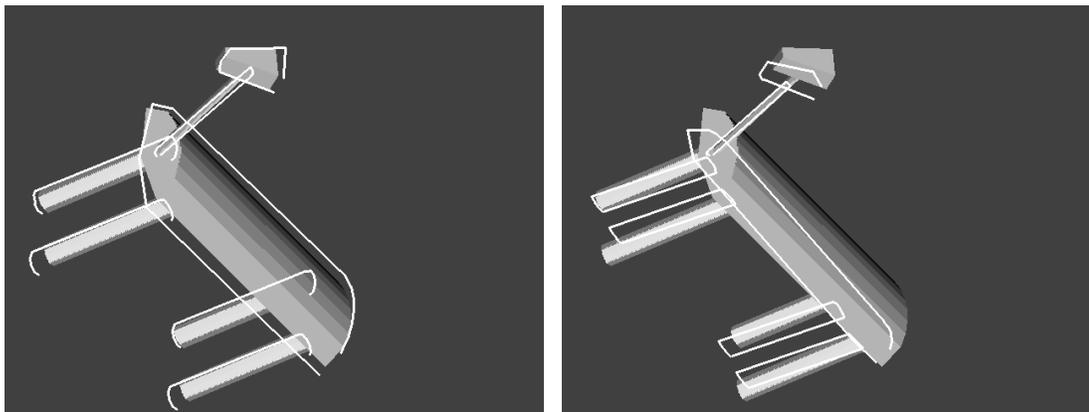


Figure 19: Here we show a view of the object in which the leg volumes are significantly occluded by the body. We match all seven volumes to regions. The rectangular backward constraints (left) produce a fairly accurate pose. The forward constraints (right), which do not allow for such occlusion, produce a much noisier pose.

can clearly be applied to a large class of objects. Our contribution is to show that it can also be used to accurately determine the pose of these objects. We show that even without specific correspondences between local geometric features, we may use region matches to determine the correct model pose. At the same time our method allows us to incorporate correspondences between points, lines or line segments, should they be available.

Specifically, we present new results that show that the forward constraints, which allow for self-occlusion, may correctly determine the pose of a 3-D object, typically when a correspondence has been found between three 3-D parts of the object and three matching 2-D image regions. We have also shown that it is a more difficult problem to find poses that satisfy the backward constraints, which allow for unidentified occlusions. This problem can have multiple disconnected solutions. These results apply not just to our algorithm but to any parts-based recognition system which determines pose while allowing for arbitrary image occlusions of unknown locus. However, we have also devised a novel algorithm, based on work in computational geometry, that finds solutions that satisfy an approximation to the backward constraints, using linear programming. And we show that in some cases of interest, this approximate solution leads to our finding exactly the correct model pose. These results demonstrate that we can recognize 3-D objects using a very simple, and novel representation of their structure.

## Acknowledgment

The authors wish to thank Baskaran Vijayakumar for his assistance in building the experimental system and in running the experiments.

## References

- [1] Alter, T. D. and D. Jacobs, 1994, "Error Propagation in Full 3D-from-2D Object Recognition", *IEEE Conf. on Computer Vision and Pattern Recognition*: 892–898.
- [2] Amenta, N., 1992, "Finding a Line Traversal of Axial Objects in Three Dimensions," *Proc. of the Third ACM-SIAM Symp. on Discrete Alg.*:66-71.
- [3] Amenta, N., personal communication.
- [4] Amenta, N., 1994, "Bounded boxes, Hausdorff distance, and a new proof of an interesting Helly-type theorem", *Proceedings of the 10th Annual ACM Symposium on Computational Geometry*: 340–347.
- [5] Basri, R., 1994. "Paraperspective  $\equiv$  affine", *International Journal of Computer Vision*, forthcoming. Also *The Weizmann Institute of Science*, T.R. CS94-19.
- [6] Basri, R. and D. W. Jacobs, 1995. "Recognition using region correspondences", Forthcoming Technical Report, the Weizmann Institute.

- [7] R. Basri and D. Jacobs, “Recognition Using Region Correspondences,” *Int. Conf. on Comp. Vis.* 1995.
- [8] R. Basri and D. Jacobs, “Matching Convex Polygons and Polyhedra, Allowing for Occlusion”, R. Basri, D. Jacobs. First ACM Workshop on Applied Computational Geometry. pp. 57–66, 1996.
- [9] R. Bergevin and M. Levine, “Generic Object Recognition: Building and Matching Coarse Descriptions from Line Drawings,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **15**(1):19-36, 1993.
- [10] Biederman, I., 1985, “Human Image Understanding: Recent Research and a Theory,” *Computer Graphics, Vision, and Image Processing*, (32):29-73.
- [11] Binford, T., 1971, “Visual Perception by Computer,” *IEEE Conf. on Systems and Control*.
- [12] R. Brooks, “Symbolic Reasoning Among 3-D Models and 2-D Images,” *Artificial Intelligence*, **17**:285-348, 1981.
- [13] J. Canny, “A Computational Approach to Edge Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**(6):679-698, 1986.
- [14] D. Clemens, *Region-Based Feature Interpretation for Recognizing 3D Models in 2D Images*, MIT AI TR-1307, 1991.
- [15] Conway, J.B., 1990. *A Course in Functional Analysis*. Springer-Verlag.
- [16] Duda, R.O. and Hart, P.E., 1973. *Pattern classification and scene analysis*. Wiley-Interscience Publication, John Wiley and Sons, Inc.
- [17] Dudani S.A., Breeding K.J., and McGhee R.B., 1977. “Aircraft identification by moments invariants”. *IEEE Transactions on Computations*, bf C-26(1): 39–46.
- [18] Fischler, M.A. and Bolles, R.C., 1981. “Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography”. *Com. of the A.C.M.* **24**(6): 381–395.
- [19] Forsyth, D., Mundy, J., Zisserman, A., and Rothwell, C., 1992. “Recognising Rotationally Symmetric Surfaces from their Outlines,” *European Conf. on Comp. Vis.*:639–647.
- [20] Forsyth, D., 1993. “Recognizing Algebraic Surfaces from their Outlines,” *Fourth Int. Conf. on Comp. Vis.*:476–480.
- [21] Gross, A. and Boulton, T., 1990, “Recovery of Generalized Cylinders from a Single Intensity Image,” *Proc. of the DARPA IU Workshop*:557–564.
- [22] Horaud, R., 1987, “New Methods for Matching 3-D Objects with Single Perspective Views,” *IEEE Trans. Pattern Anal. Machine Intell.*, **9**(3): 401–412.

- [23] Hu M.K., 1962. "Visual pattern recognition by moment invariants". *IRE Transactions on Information Theory*, **IT-8**: 169–187.
- [24] Huttenlocher, D., G. Klanderman, and W. Rucklidge, 1993, "Comparing Images Using the Hausdorff Distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15(9)**:850-863.
- [25] Huttenlocher, D., J. Noh, and W. Rucklidge, 1993, "Tracking Non-Rigid Objects in Complex Scenes," *4th Int. Conf. on Computer Vision*: 93–101.
- [26] Huttenlocher, D.P., and Ullman, S., 1990 "Recognizing Solid Objects by Alignment with an Image", *Int. J. Computer Vision*, **5(2)**: 195–212.
- [27] Jacobs, D., 1992. "Space efficient 3D model indexing", *IEEE Conference on Computer Vision and Pattern Recognition*: 439–444.
- [28] Jacobs, D., 1992, "Recognizing 3-D Objects Using 2-D Images". *M.I.T. A.I. Memo 1416*.
- [29] Jacobs, D. and R. Basri, 1997, "3-D to 2-D Recognition with Regions", *IEEE Conference on Computer Vision and Pattern Recognition*.
- [30] Koenderink, J. and van Doorn, A., 1991. "Affine structure from motion", *Journal of the Optical Society of America*, **8(2)**: 377–385.
- [31] Kriegman, D. and J. Ponce, 1990, "On Recognizing and Positioning Curved 3-D Objects from Image Contours," *IEEE Trans. Pattern Anal. Machine Intell.*, **12(12)**: 1127–1137.
- [32] Lamdan, Y. & H.J. Wolfson, 1988, "Geometric Hashing: A General and Efficient Model-Based Recognition Scheme," *Second International Conference Computer Vision*: 238–249.
- [33] Lowe, D., 1985, *Perceptual Organization and Visual Recognition*, The Netherlands: Kluwer Academic Publishers.
- [34] Marr, D. and Nishihara, H., 1978, "Representation and Recognition of the Spatial Organization of Three Dimensional Structure," *Proceedings of the Royal Society of London B*, **200**:269-294.
- [35] Megiddo, N., forthcoming.
- [36] Nagao, K. and Grimson, W., 1994, "Object Recognition by Alignment using Invariant Projections of Planar Surfaces," *12th Int. Conf. on Pattern Rec.*:861–864.
- [37] Nayar, S. and Bolle, R., forthcoming, "Reflectance Based Object Recognition," *Int. J. of Comp. Vis.*
- [38] Pellegrini, M., "Stabbing and Ray Shooting in Three Dimensional Space," *Proc. of the 6th Annual Symp. on Comp. Geometry*:177–186, 1990.

- [39] A. Pentland, "Recognition by Parts." *Proceedings of the First International Conference on Computer Vision*:612-620, 1987.
- [40] Persoon E. and Fu K.S., 1977. "Shape descimination using Fourier descriptors". *IEEE Transactions on Systems, Man and Cybernetics* **7**: 534-541.
- [41] Poelman, C.J. and Kanade, T., 1994. "A paraperspective factorization method for shape and motion recovery". *Proc. of European Conf. on Computer Vision*.
- [42] Ponce, J., Chelberg, D., and Mann, W., 1989, "Invariant Properties of Straight Homogeneous Generalized Cylinders and Their Contours," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(9): 951-966.
- [43] Reeves A.P., Prokop R.J., Andrews S.E., and Kuhl F.P., 1984. "Three-dimensional shape analysis using moments and Fourier descriptors". *Proc. of Int. Conf. on Pattern Recognition*: 447-450.
- [44] Richard C.W. and Hemami H., 1974. "Identification of three dimensional objects using Fourier descriptors of the boundry curve". *IEEE Transactions on Systems, Man and Cybernetics*, **4**(4): 371-378.
- [45] Rivlin, E., Dickinson, S., and Rosenfeld, A., 1994, "Recognition by Functional Parts," *Proc. of Int. Conf. on Pattern Recognition*: 267-274.
- [46] Rock, I., 1983. *The Logic of Perception*, MIT Press, Cambridge, MA.
- [47] Rothwell, C., Forsyth, D., Zisserman, A. and Mundy, J., 1993, "Extracting Projective Structure from Single Perspective Views of 3D Point Sets," *Third Int. Conf. on Comp. Vis.*:573-582.
- [48] Sadjadi F.A. and Hall E.L., 1980. "Three-dimensional moment invariants". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**(2): 127-136.
- [49] Seidel, R., 1990, "Linear Programming and Convex Hulls Made Easy," *Proc. of the Sixth Annual Symp. on Comp. Geometry*, pp. 211-215.
- [50] Shafer, S. and Kanade, T., 1983, "The theory of stright homogeneous generalized cylinders," Tech. Report CS-083-105, Carnegie Mellon Univ.
- [51] Solina, F. and Bajcsy, R., 1990, "Recovery of Parametric Models from Range Images: The Case for Superquadrics with Global Deformations," *IEEE Trans. on PAMI*, **12**(2): 131-146.
- [52] Sugimoto, A. and Murota,, K., 1993. "3D object recognition by combination of perspective images". *Proc. of SPIE, Vol. 1904*: 183-195.
- [53] Terzopoulos, D. and D. Metaxas, 1991, "Dynamic 3D Models with Local and Global Deformations: Deformable Superquadrics," *IEEE Trans. on PAMI* **13**(7):703-714.

- [54] Thompson, D. and Mundy, J. 1987. Three-Dimensional Model Matching From an Unconstrained Viewpoint. In *Proceedings IEEE Conference Rob. Aut.*, pp. 208-220.
- [55] Tomasi, C. and T. Kanade, 1992, "Shape and Motion from Image Streams under Orthography: a Factorization Method," *International Journal of Computer Vision*, **9**(2):137–154.
- [56] Ulipinar, F. and Nevatia, R., 1995, "Shape from Contour: Straight Generalized Cylinders and Constant Cross Section Generalized Cylinders," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**(2): 120–135.
- [57] Ullman, S. and Basri, R., 1991. "Recognition by linear combinations of models". *IEEE Trans. on PAMI*, **13**(10): 992–1006.
- [58] Vijayakumar, B., Kriegman, D., and Ponce, J., 1995. "Invariant-Based Recognition of Complex Curved 3D Objects from Image Contours," *Fifth Int. Conf. on Comp. Vis.*:508–514.
- [59] M. Zerroug and R. Nevatia, "Using Invariance and Quasi-Invariance for the Segmentation and Recovery of Curved Objects," *Applications of Invariance in Computer Vision*, edited by J. Mundy and Z. Zisserman, Springer-Verlag, Berlin, Heidelberg, 1994.