Scene Description Using Range Data

Prasanna G. Mulgaonkar, Cregg K. Cowan, and Jeff DeCurtins
SRI International, Menlo Park, California 94025

ABSTRACT

A technique for describing scenes consisting of piles of simple but unknown objects using dense range images is discussed. The technique uses concepts such as symmetry, stability, viewpoint independence, and object solidity to hypothesize the unknown shapes and sizes of the objects. These hypotheses are analyzed using the known geometry of the range sensor to rule out the inconsistent configurations. The final result of the analysis is one or more descriptions of the 3-D scene, each of which is consistent with the sensed data, and with the constraints imposed by the physics of objects in contact.

INTRODUCTION

Analyzing the contents of a pile of 3-D objects using noncontact sensing is an area of great research interest, and one that is encountered in a large variety of industrial tasks of the bin-picking type. Traditional approaches to bin-picking systems use strong object models. The pile is constrained to consist of objects whose geometry is known. This is reasonable in a large number of industrial vision tasks, but such approaches do not apply to an equally large set of unstructured material handling applications such as mailpiece singulation, random or mixed part feeding, scavenging, fire fighting and other similar tasks. In such applications, the key vision problem is determining the size and shape of the objects and how the objects relate to each other and to other invisible objects that may be underneath; in other words, describing the scene in terms that would be useful for controlling subsequent manipulation operations.

In this paper, we show how heuristics, such as object symmetry, and assumptions, such as general viewpoint, can be used to generate “initial” descriptions of the partially visible objects in a pile. We then examine these hypotheses to determine their possible extension using criteria such as coplanarity of disconnected surfaces and intersection of swept volumes to produce “complete” hypotheses. Finally, we analyze the configuration of the hypothesized objects to determine their contacts, the contact forces, and support relationships, and from these, we estimate what the invisible interior of the pile looks like.

DATA-DRIVEN INTERPRETATION

Piles consisting of one to a dozen objects such as the ones shown in Figure 1 constitute the scenes that are the input to our interpretation system. We obtain dense range data from the scene using a triangulation range sensor shown schematically in Figure 2. The pile moves on a conveyor belt at a constant velocity through a plane of light obtained by passing a laser beam through a cylindrical lens. The scene is viewed by a 2-D CCD array camera that images cross sections through the scene at each position of the belt. The resulting 254 x 254 range images have a nominal spatial and depth resolution of 0.09 inch, and take about a minute to acquire.

It is intuitively clear from this description that we obtain only partial information about the scene. This restriction is common to all forms of noncontact sensing that use...
there are three techniques for overcoming this limitation:

- Use of multiple sensors or a single sensor that can be moved to different viewpoints.
- Use of CAD or other geometric models that can be matched against the visible scene components. Once a match is found, the invisible portions of the scene can be filled in using the data in the CAD models.
- Use of data-driven reasoning by which complete scene descriptions are extrapolated from the visible data using simple generative rules.

In general, multiple viewpoints have limited applicability because sensors cannot penetrate piles of objects to view the interior, and sensor mobility may not be an option in many applications such as scavenging or fire fighting. Detailed CAD models are available in standard manufacturing domains. Consequently, object extents can be inferred using model-based scene interpretation techniques such as those of Horand and Bolles [1], and Faugeras and Hebert [2]. In such techniques, a matching procedure is used to associate aspects of the visible surfaces with aspects of the known 3-D models. These associations uniquely define the geometric transformation that relates model features to the scene. Since the models are complete 3-D descriptions, the transformation provides a way for filling in the portions of the scene that were occluded from the camera. Such techniques cannot be applied in situations where dimensioned CAD models are unavailable.

Because we are interested in unstructured domains, we have developed techniques that are data driven, and extrapolate the sensed data to form hypotheses about the scene. For example, in piles consisting of single objects, we can detect the presence of global regularities (such as rotational, translational, or reflectional symmetry) in the shape of the visible surfaces and, under the assumption that all of the object shares the same regularity, the invisible portions of the scene can be estimated.

All hypotheses are complete descriptions of 3-D objects in the scene. To be considered valid, all hypotheses must satisfy two types of constraints. First, because the hypotheses refer to physical objects in a physical world, valid hypotheses must satisfy physical constraints on the position, orientation, and stability of rigid objects in the presence of gravity and friction. Second, the hypotheses must also satisfy imaging constraints imposed by the range-image-acquisition system used to gather the data.

Physical constraints manifest themselves by limiting the possible sizes of the partially visible objects. Objects may not intersect each other (the solid-object constraint), objects must contact either the support surface or another object in a set of discrete points, lines, or planes (the support constraint), and the sum of the forces and torques due to gravity at the various contact points for an object must add to zero (the stability constraint). The computational aspects of such physical rules are currently under study by a large number of researchers in the field of automatic design (Murthy and Addanki [3], Joskowicz [4]), qualitative physics (deKleer and Brown [5], Hayes [6], Forbus [7]), and physics-based graphical modeling (Wilhelms [8], Barzel and Barr [9], Terzopoulos and Fleischer [10], Hahn [11]; for a collection of articles, see Barr et al. [12]). Although we use only a few rules in the current level of our system, we plan on expanding the rule base with time.

Sensing constraints imposed by the image acquisition system are implicitly present in the sensed data. In the past, computer vision researchers have viewed the data obtained by range sensors as defining a thin sheet lying on the visible surfaces of all objects in the scene. Assume a simplified model of the sensor geometry for the purpose of this discussion: Let the sensor be a single point above the scene that produces a perspective image of the objects in front of it as shown schematically in Figure 3a. Under such a model, the range data obtained from the scene lie on the infinitely thin sheet shown in Figure 3a. In reality, due to quantization both spatially and in depth, this sheet is neither continuous nor is it vanishingly thin. However, we can ignore reality for a moment.

Viewing the relationship between the sensed data points and the scene as a thin sheet is, in our view, restrictive because it hides one important fact about the sensing: To be able to obtain any given data element, the sensor has to have a clear line of sight from a sensing element to that data element. Thus, the range data define more than a set of points in 3-D, they define lines of empty space in the scene. In other words, the sensed data do not define surfaces, but volumes in space. These volumes of free space (shown schematically in Figure 3b) define constraints that must not be violated: None of the objects can "penetrate" into these free spaces, because if they did, they would have been visible in the original data.

FIGURE 2. A TRIANGULATION RANGE SENSOR

a single viewpoint. We do not know the true extent of the objects because we cannot see inside the pile. In general, there are three techniques for overcoming this limitation:

- Use of multiple sensors or a single sensor that can be moved to different viewpoints.
- Use of CAD or other geometric models that can be matched against the visible scene components. Once a match is found, the invisible portions of the scene can be filled in using the data in the CAD models.
- Use of data-driven reasoning by which complete scene descriptions are extrapolated from the visible data using simple generative rules.

In general, multiple viewpoints have limited applicability because sensors cannot penetrate piles of objects to view the interior, and sensor mobility may not be an option in many applications such as scavenging or fire fighting. Detailed CAD models are available in standard manufacturing domains. Consequently, object extents can be inferred using model-based scene interpretation techniques such as those of Horand and Bolles [1], and Faugeras and Hebert [2]. In such techniques, a matching procedure is used to associate aspects of the visible surfaces with aspects of the known 3-D models. These associations uniquely define the geometric transformation that relates model features to the scene. Since the models are complete 3-D descriptions, the transformation provides a way for filling in the portions of the scene that were occluded from the camera. Such techniques cannot be applied in situations where dimensioned CAD models are unavailable.

Because we are interested in unstructured domains, we have developed techniques that are data driven, and extrapolate the sensed data to form hypotheses about the scene. For example, in piles consisting of single objects, we can detect the presence of global regularities (such as rotational, translational, or reflectional symmetry) in the shape of the visible surfaces and, under the assumption that all of the object shares the same regularity, the invisible portions of the scene can be estimated.
In this section, we describe techniques for extending the visible surfaces of objects in a pile to form solid objects. We begin with geometric surface descriptions extracted from the range data, i.e., each surface is classified as planar or cylindrical \cite{13}. We then use heuristics to generate "complete" descriptions of the partially visible objects.

We use visible surface patches larger than a fixed size threshold as seed hypotheses for objects. Each of these seed hypotheses is then made into one of two solid-model primitives: a prism or a cylinder.

To form a prism, the polygonal outline of a planar surface patch is projected some distance along the plane normal away from the sensor. A box, for example, is a special case of a prism in which the polygonal outline is a rectangle. To form a cylinder, the center, radius, and length are taken from results of the surface segmentation algorithm.

In the rest of this section and in the next, we will use the pile shown in Figure 4 as an example. It consists of a rectangular box butted against the end of a cylinder. Figure 5 shows the surface descriptions for the example image.

To generate a hypothesis about the structure of a scene, the following steps are necessary:

- Estimate the maximal extent of each potential object in the scene.
- Determine and combine objects that appear fragmented because of occlusion by other objects in the scene.
- Verify resulting hypotheses for stability.

The scene descriptions (hypotheses) that are generated depend on the order in which object fragments are examined and processed. Typical scenes contain between 2 to 20 visible fragments. Clearly, the combinatorics of the situation do not permit a generation of all permutations in which n fragments can be ordered. To simplify this situation, we process surface fragments in two fixed orders: One hypothesis is generated in decreasing order of fragment size. A second alternative is generated in increasing order of height above the support surface. The first order essentially allows large and presumably better defined objects to be formed first. The second order allows supporting objects to be completed before objects that they support resulting in smaller extents for most objects.

To estimate the extent of a partially viewed object, it is necessary to know in which directions the boundary of such an object hypothesis may be extended. To determine this information, we analyze the detected boundary of each surface patch in the image.

A boundary (edge) pixel may arise from one of four sources in a triangulation (camera plus laser light plane) range image: (1) a convex or concave slope discontinuity, (2) an occluding edge of a solid object, (3) a shadow cast by an occluding edge along either the camera's line of sight or the laser's line of sight, and (4) the boundary of the picture itself, if the object extends outside the field of view of the sensor.

Note that the last two, the shadow and picture boundary, are artifacts of the imaging system and do not represent physical object boundaries. A hypothesized object, if extended along one of these artifact boundaries, may be made considerably larger than the visible boundary indicates without contradicting the observed data. Thus, we label all surface patch edge pixels as extensible or nonextensible to restrict the range of our hypotheses.

Labeling a picture boundary edge is simple. Any edge pixel on the border of the image pixel array is marked as extensible. To label the remaining edge pixels, we note that each occluding edge pixel casts a shadow and thus should have a corresponding shadow edge pixel in the image. If we can pair up occluding and shadow pixels, the shadow pixels can then be marked as extensible. All remaining unmarked edges are by default nonextensible.
The pairing of occluding and shadow pixels relies on knowledge of the sensor geometry. In particular, a shadow in a range image obtained by triangulation may be the result of occlusion of either the camera’s view or the laser’s light. Thus, if we take an arbitrary edge pixel, A, and form three-dimensional vectors from this pixel’s coordinates to all other edge pixels, one of these vectors (to pixel B, say) should lie along either the laser’s ray or the camera’s view to pixel A. If pixel A is closer to the camera than pixel B, then pixel A belongs to the occluding edge; otherwise, it is on the shadowed edge.

Examination of all possible pairs of edge pixels for alignment with the camera or laser vectors requires on the order of $n^2$ camera, laser, and pixel vector operations, where $n$ is the number of edge pixels in the image. In a typical multiobject pile, we find thousands of edge pixels; thus, some method is necessary to speed the computation.

As an alternative to the exacting but exhaustive method, we approximate the camera and laser vectors to each edge pixel by a single camera and laser vector pair: The vectors to the center of the image. This is a good approximation if the distance from the scene to the sensor is large compared with the dimensions of the scene itself. In the discussion that follows, we describe the procedure for using the central camera vector. The procedure for using the central laser vector is identical.

Using the central camera vector, we form a coordinate frame $(u, v, w)$ in which the $w$ axis is the camera vector and the $(u, v)$ axes form a plane perpendicular to $w$. The $uv$ plane is then discretized into a grid where the size of the grid squares is a function of the sensor resolution. The reason for this is explained as follows.

By mapping each pixel’s 3-D coordinates $(x, y, z)$ into the $uvw$ frame, we associate the pixel with a particular $uv$ grid square. The data structure for this square keeps track of the minimum $w$ value of all pixels mapped to this square; i.e., the pixel closest to the camera. Note that in any occluding and shadow pair relationship, the pixel closest to the camera will be the occluding pixel. The first pass over the edge pixels in the image establishes the minimum $w$ value for each grid square. During the second pass, a pixel’s $w$ coordinate is compared with the minimum $w$ for the associated grid square. If this $w$ is within some tolerance (roughly twice the sensor resolution) for the minimum for the associated square, the pixel is an occluding pixel; otherwise, it is a shadow pixel (and thus extensible). The size of the grid square is important here. If the square is too small, occluding pixels and their corresponding shadow pixels may be mapped to different squares and never compared against each other. If the square is too large, the spread in the $w$ coordinates of occluding pixels mapped to the same square will be too large for edges that do not parallel the $uv$ plane. In practice, we have used a grid square dimension approximately four times the sensor resolution.

The grid approach is linear in the number of edge pixels. The process is repeated for the central laser ray. When complete, all extensible edge pixels have been labeled (see Figure 6).

In addition to knowing whether an edge of a surface patch is extensible, we also need to know how far the edge may be extended without violating one of the physical constraints discussed. To determine the limitations on the extensibility of object edges, we explicitly model all volume known to be free of objects. Essentially, this volume consists of the space between the camera or laser and any visible point (either on a surface patch or on a portion of the underlying support surface). No unseen portion of a hypothesized object may enter this freespace. If it did, it would have come between the sensor and a surface seen by the sensor and thus would have been visible itself.

To model this freespace above the visible surfaces, we must first note that because our sensor sweeps the volume of the scene through an imaged plane of light, the apparent camera and laser position is different for each stripe of the image. Thus, we represent freespace above any visible surface patch by two truncated, tesselated pyramids, one with respect to the camera, the other with respect to the laser. The base of each of these is the polygonal boundary of the surface patch. To form the sides of the pyramid, we compute a vector from each vertex of the base polygon to the camera (or laser) and find the intersection of this vector with an arbitrary horizontal plane located well above the imaged volume. For each pair of adjacent base vertices, $B_i$ and $B_{i+1}$, with their associated horizontal plane intersections, $I_i$ and $I_{i+1}$, we form two triangles,
and \((B_{i+1}, L_i, L_{i+1})\), which cover the possibly nonplanar area enclosed by \((B_i, L_i, L_{i+1}, B_{i+1})\). The tessellated pyramid defined by a pair of surfaces is shown in Figure 7.

As we have discussed, the goal of the construction is to hypothesize a set of prisms or cylinders or both that incorporate all of the visible surface patches and which, taken together, form a plausible explanation of the pile's configuration. The collection of construction rules we have described is applied repeatedly to the next surface patch in the list generated either by decreasing size or by increasing height above the support. When a patch becomes explained as part of a hypothesized object, it is removed from the candidate list.

As a first step, the largest remaining unexplained patch is chosen as the seed for the next hypothesized object. Note that the scene description that results depends on the order in which patches are selected for expansion. In a future paper, we will describe our current efforts in generating all possible hypotheses. If the patch is planar, we form a prism using the patch's plane normal to the prism axis and the patch's polygonal boundary as the endface of the prism. If the patch boundary has been additionally classified as rectangular by the surface segmentation process, a rectangular boundary is substituted for the observed polygonal boundary. The polygonal or rectangular boundary is then swept along its axis until some portion of it intersects known free space, a previously hypothesized object, or the known support surface. If the patch is cylindrical, a cylinder is hypothesized based on the axis, length, and radius determined by the surface segmentation preprocessing. If the ends of the cylinder judged to be extensible by the process described above, the body of the cylinder is swept along its axis in the same manner as the polygonal boundaries of prisms.

After this object hypothesis has been formed, we test all remaining unexplained patches to see if any appear to lie on the surface of our newly formed object. Because there are typically thousands of pixels on each surface patch, we test only a sample of these to see if they lie on the endfaces or the trunk of the hypothesized prism or cylinder. If any patch passes this test, we mark it as "explained" and associate it with the newly hypothesized object.

Once all unexplained patches have been checked for inclusion in the new hypothesis, we examine all existing prism hypotheses to see if any may be unioned into a single prism. This test looks for sets of prisms in the hypothesis list that have coplanar endfaces and the same length along their axis. To form the union, we assume the prisms are simply portions of the same partially occluded object. In the common plane of their edges, we form a two-dimensional convex hull around the collection of their edge pixels. This convex hull is the edge boundary of a new prism hypothesis. We then test this new hypothesized object to ensure that no visible surface patch is inside it. If an interior patch is found, this implies the sensor has seen through the hypothetical union and thus the union is implausible. If no such contradicting evidence is found, the union prism replaces its component prisms in the hypothesis list.

Similar to the operation linking prisms, cylinders may also be unioned to form a single hypothesized cylinder. For cylindrical unions, we test for coaxiality, equal radius, and no visible patch enclosure. If the union passes these tests, it replaces the component cylinders in the hypothesis list. In the case of the two objects shown in Figure 4, the visible area of the box is larger than that of the cylinder. Consequently, in the hypothesis generated by decreasing size, the box gets expanded first. The resulting hypothesis is shown in Figure 8a. The box is taller than the cylinder, and consequently, in the hypothesis generated by increasing height above the support surface, the cylinder gets expanded first into a description shown in Figure 8b. This second hypothesis is clearly unstable, and is ruled out by subsequent stability analysis.

Results thus far indicate that the above rules, when applied to an image until no surface patch remains unexplained, generate a set of hypotheses at least one of which reflects the actual pile configuration quite well, even for fairly complex scenes.
ESTIMATING STABILITY

The stability of objects in a pile arises from support forces generated at the regions of contact between objects. These contact regions may take the form of points (for example, a cylinder leaning against the edge of a box), lines (a box on edge or a cylinder lying on a planar surface) or planes (a box lying flat). If, for a given hypothesis, we can find the regions of contact between an object and its surroundings, we can compute the forces and torques on the object. Since the sum of forces and torques must add to zero for a static scene, the closeness of our calculated sum to zero provides a measure of the validity of the hypothesis.

To compute the contact regions, we form a 3-D voxel array whose volume encloses the entire scene. We then sample the surfaces of our hypothesized objects and, for each sample point, mark the voxel that contains it. The mark indicates from which object the point was derived. When this process is complete, any voxel with marks from more than one object is a contact point. If we collect all voxels marked by two particular objects, say A and B, we get the region of contact between object A and object B.

We may characterize the region as a plane, line, or point by computing the moments of the voxels in the region. A planar region will, for example, have two relatively large moments and one small amount. Using the set of these contacts, we plan to estimate the stabilizing force each could supply and from this, generate free-body diagrams for each hypothesized object. This leads directly to evaluation of the stability of the configuration and hence, to an additional test of the validity of the hypothesis. The contact points between the two objects shown in Figure 4 are shown in Figure 9.

We currently use a simplified test for object stability. We project all the contact points at which any object touches its neighbors onto an artificial horizontal surface below the object. We then construct the 2-D convex hull of the projected contact points called the support footprint. If the centroid of the hypothesized volume of the object is directly above the support footprint, we declare the object stable. Figure 10 shows the support footprint and the centroid for the box in the unstable configuration.

The current test is only approximate because it cannot analyze cantilevered objects. In these situations, some interobject contacts push down on the object of interest. Figure 11a shows one hypothetical case where this footprint test breaks down, incorrectly calling an unstable configuration stable, and Figure 11b shows a stable configuration that will be marked unstable. A true test for stability involves the simultaneous solution of a set of inequalities that define the contact forces and friction at each contact point on each of the objects in the scene. Computationally, this is difficult because of the variety of ways in which contact forces may distribute over line or area contacts. We are currently investigating iterative ways in which such computation can be done following the work of Fahlman [14].
CONCLUSIONS

We have described a system that is capable of analyzing piles of 3-D objects to estimate the dimensions and configuration of objects within. The system uses physical and geometric constraints to control the generation and evaluation of these estimates. In its present form, the system uses only a few simple rules such as object solidity, friction, stability, support, and so on. Work is underway to make the system more robust and capable of analyzing a wider variety of objects, and capable of analyzing pile stability.

ACKNOWLEDGMENTS

Research work leading to this paper was funded by the Office of Advanced Technology, United States Postal Service, as Task Order 104230-87-M-0194, under the Basic Ordering Agreement 104230-84-D-0963 for Electro-Optics and Character Recognition.

REFERENCES


