COOPERATIVE INTEGRATION OF VISION AND TOUCH

Peter K. Allen

Department of Computer Science
Columbia University
New York, NY 10027

ABSTRACT

Vision and touch have proved to be powerful sensing modalities in humans. In order to build robots capable of complex behavior, analogues of human vision and taction need to be created. In addition, strategies for intelligent use of these sensors in tasks such as object recognition need to be developed. Two overriding principles that dictate a good strategy for cooperative use of these sensors are the following: 1) sensors should complement each other in the kind and quality of data they report, and 2) each sensor system be used in the most robust manner possible. We demonstrate this with a contour following algorithm that recovers the shape of surfaces of revolution from sparse tactile sensor data. The absolute location in depth of an object can be found more accurately through touch than vision; but the global properties of where to actively explore with the hand are better found through vision.

1. INTRODUCTION

Our research focuses on making dextrous robotic hands major components of complex robotic systems that can be used to grasp, manipulate, inspect and recognize 3-D objects. Our initial work on this problem [2, 3] has focused on developing a system of hand primitives that can be used to build higher level system modules for the tasks described above and creating a number of Exploratory Procedures (EP's) that are robotic analogs of methods humans use to perceive with touch [7]. The EP's can be thought of as a set of primitive haptic functions that can be used as the building blocks for an active, autonomous haptic recognition system. The requirements of such a system are more complex than a similar vision based recognition system, primarily due to the active control needed for the hand.

The three EP's we have built are fully described in [4]. They have been implemented on a system that consists of a Utah-MIT hand attached to a PUMA 560 robot and interlink tactile sensors attached to the distal links of each finger [2]. The first EP is grasping by containment. Grasping by containment is an attempt to understand an object's gross contour and volume by effectively molding the hand to the object. Using these primitives, we have been able to recover the global shape of objects such as cylinders, rectangular blocks, funnels, wedges and lightbulbs. The technique uses very sparse touch sensor data (30-100 points) and serves as an excellent initial shape estimator for further active sensor exploration. The second EP we have implemented with our robotic hand system is a Lateral Extent EP. This EP is used to explore a continuous, homogeneous surface such as a planar face, and to determine its extents. Research with human subjects suggests that this strategy is used until a discontinuity is found, which can cause a change in haptic sensing strategy. This EP is capable of determining the extents of a planar surface, and by using multiple fingers, different planar faces can be explored. The intersections of these planes form the edges and vertices of the object in question, and are easily computed. The third EP we have built is a Contour Follower EP which we have used to recover the shape of surfaces of revolution, which form a restricted class of generalized cylinders. This EP is the basis of our initial fusion of vision and touch and we now describe it in detail.

First, the PUMA is moved to a known location near one end of the explored object, and the thumb and index finger are opened enough to allow them to encompass the object without making contact with it. Then the thumb is slowly moved toward the object until the sensors detect contact between the thumb and the object. Next, the index finger follows the same movement. After detecting contact, the positions of the two contact locations are noted, and the fingers are backed off the object so that they are no longer in contact. The arm and hand are moved a small amount along the axis of the explored object, and the process is repeated. This exploratory procedure ends when one of the fingers moves toward the object and fails to make contact. (The location of the object and its axis are given to the system *a priori*

We have performed a series of experiments that try to recover the shape of a number of different surfaces of revolution including a wine bottle, a beer bottle, a coke bottle and an Orangina soft drink bottle (a flask like object). The procedure begins with exploring the object along an exploration axis that is assumed to be perpendicular to the support table. The points generated from these contour traces are then linked into a set of linear contour segments. Circular cross section curves are then fit perpendicular to the exploration axis that include trace points from each of the contours. The recovered shapes are shown in figure 1. The shapes are clearly distinguishable from this sparse data. An additional and important discriminating characteristic is actual 3-D size and volume which are calculable from these representations.

2. USING MULTIPLE SENSING MODALITIES FOR AUTONOMY

The EP's reported above work well in recovering shape information about rigid 3-D objects. However, the exploration of these objects by an active hand requires a level of control that is not present in typical vision based robotic systems. We believe that by integrating both vision and touch sensing we can increase the autonomy in these systems. Our overall approach to the problem of robotic object recognition lies in a multi-sensor approach; we believe no single sensing modality is currently powerful enough to robustly perceive and recognize its environment. Just as humans exploit a multitude of sensor systems, robotic systems need to use multiple sensors for perception as outlined in Allen [1] and Kak and Chen [6]. A central idea in using multi-sensor data is that over-reliance on one sensor can cause error. It has been empirically observed that trying to extract too much information from a single sensing modality results in a degradation of results; however, using only the most reliable and highest confidence sensor data allows one to proceed along a path that is known to be correct. We call this principle *less is more,* in that reduced amounts of reliable data from a single sensor are more useful than large amounts of data which may be spurious. By combining the data that is most reliable from each of a number of sensors, more accurate results may be computed. An interesting analogy to our work is in the ability of blind people to perceive the world. While blind people can and do function in complicated environments, small amounts of other sensory data (such as verbal visual guidance cues from sighted people) can extend their ability to perceive.

3. EXAMPLE: FINDING AN EXPLORATION AXIS FOR CONTOUR FOLLOWING

We now demonstrate these ideas by describing a vision module we have implemented to be used with the Contour Follower. Determining the exploration axis is a key part of the Contour Follower EP. Knowing in which direction to trace the object is important to higher level recovery procedures which need to use this information in the recognition process. Once the hand makes contact with the object, it explores the contour along a known axis which we calculate a priori. We have implemented a vision based technique to determine this axis. Our method of visual recovery of the exploration axis exploits the recent work of Wolff [9] in stereo line matching. Point-based stereo techniques tend to be unreliable in that multiple correspondences between images can cause mismatches and error. More stable matching

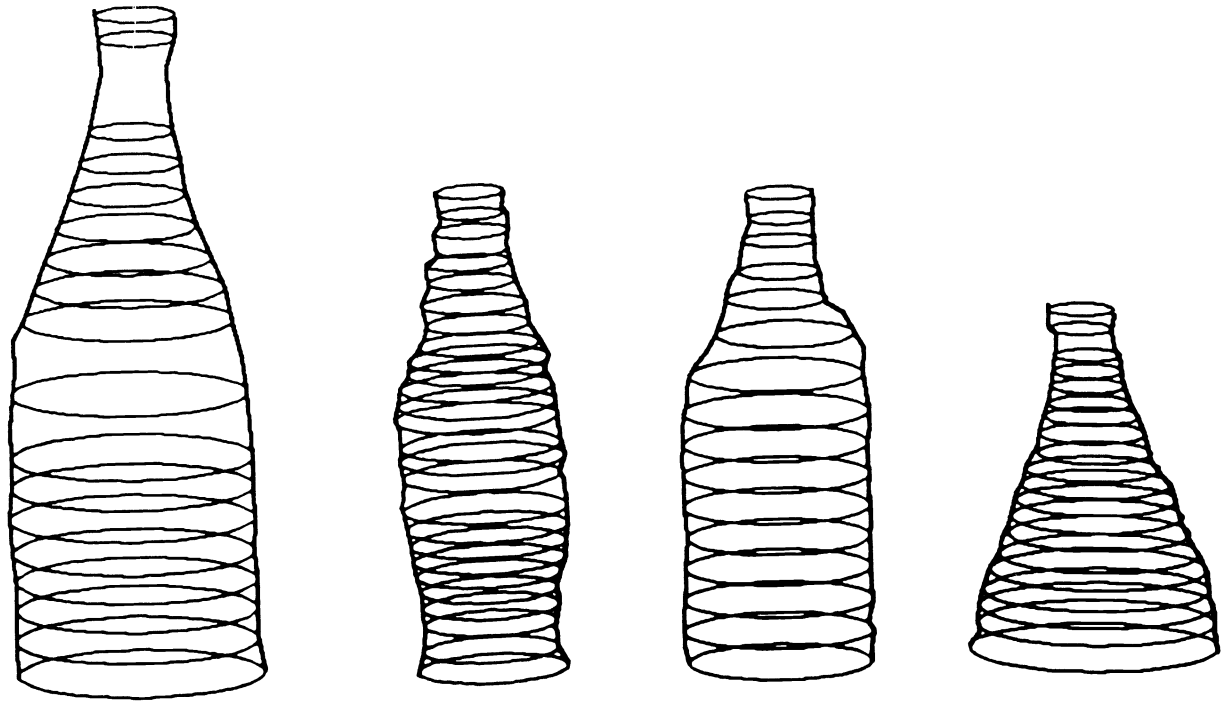

Figure 1: Linked contour points and recovered surfaces of revolution from Contour Explorer EP (left to right wine bottle, coke bottle, beer bottle, Orangina bottle).

can occur using larger primitives such as lines [5]. Even using line-based matching, problems can still occur. Matching the endpoints of lines can be prone to errors in the output of the line finder which may break a single line into multiple segments due to differing edge strengths along the line. The problem here is that 3-D depth is being computed, which requires an absolute correspondence of points (whether from point-based or line-based methods).

Our method alleviates this dependence on absolute matching of unstable primitives to generate 3-D depth. All we require of the algorithm is an *orientation* vector in 3-D. We do not need to have its absolute depth, but need to generate a match between a family of parallel lines sharing the same orientation. This orientation can then be used by the active hand as the exploration axis. The 3-D depth has already been determined from the contact of the hand with the object. Given this 3-D depth from tactile contact, we can follow the 3-D axis determined by the line based stereo matcher to continue our exploration.

It is important to note that this method is less sensitive to matching errors and baseline measurement, another common cause of stereo error. In addition, it is also less prone to the effects of physical point mismatches as the baseline increases, since we are still matching a larger entity, the line itself. Intuitively, the method creates a 3-D plane in space from the camera center and any two points on the

line. This plane and a similar plane from the other camera are all that are needed to create a 3-D intersection line which we can use as the exploration axis.

3.1. Vision Module

The vision module consists of a pair of cameras that image the object in the scene. To extract linear features from the images, an algorithm developed by Singh and Shneier is used [8]. This is a two stage process that uses the real-time edge detection capability of the PIPE image processing system to apply smoothing, gradient, threshold and thinning operators to the image in real-time. The PIPE allows images to have local operators work on multiple images in time as well as pipelining images from processing stage to processing stage for sequential processing of a multi-step image algorithm. Each processing step takes one field time (1/60 sec). The initial operator is a first derivative of a Gaussian smoothed image, which is used as an edge strength measure. This image is then thresholded using a local averaging technique over a 5x5 window to isolate edges. This procedure, while able to be realized in real-time, has the unwanted side effect of isolated "spot" edges. These can be eliminated by sending the edge image to another processing stage in the PIPE where a morphological thinning operator can be applied, again in one image field time. The thinning is accomplished by two operators, one for removing spots and another restoring template to merge multiple parallel edges.

To create full linear features, a non-local edge following algorithm is needed. This is done by shipping the edge marked binary image to a host over a high speed interface. The edges are linked into linear features by using a raster scan connected component analysis. The connected edge pixels are broken into linear pieces using a recursive subdivision algorithm. These linear-connected edge pixels are then fit to a least-square fit line that minimizes the sum of the pixels distances from the line. Finally, the lines endpoints are calculated by projecting the ends of the connected pixel chains onto the fitted line, and attributes of each line (orientation, extent, length).

The vision module performs the algorithm for each of the two images. The cameras are calibrated but no attempt is made at scan line registration. The intent is that the linear features are much more stable as candidates for stereo matching than pixel level data.

3.2. Matching

The linear feature extraction process serves as input to the matcher. The matching process is simplified by filtering the candidate lines on a number of criteria. The initial filtering criteria is edge length. All edges are rejected below a minimum length, which are taken from the statistics of the line feature extraction process. The edges remaining in each image are the longest and most stable lines in the image. The preference for line orientation is again taken from the distribution of the filtered lines, with the maximum of the orientation histogram chosen (an appropriate bucket size of 30 ° is used). The lines are then matched according to orientation and extent, with a horizontal disparity window imposed to prevent gross mismatches. The procedure works well since it is matching a very small subset of lines, which, because of their length, are quite stable. Figure 2 shows the left and right stereo images of the beer bottle with the extracted linear features overlaid on the bottle, and the matched filtered lines.

As mentioned previously, once the lines are matched, it is still not possible to fully recover depth of the object points, since the matches are not on a pixel by pixel basis. There is still a degree of freedom (translation along the stereo lines matched) inherent in the recovered line match. However, given our problem of finding a 3-D exploration axis on the object, this degree of freedom is acceptable. In fact, it is exactly the axis of exploration, along the matched line. The hand system can locate any point along the line where the object is imaged, move until it contacts this point, and begin exploring along the axis.

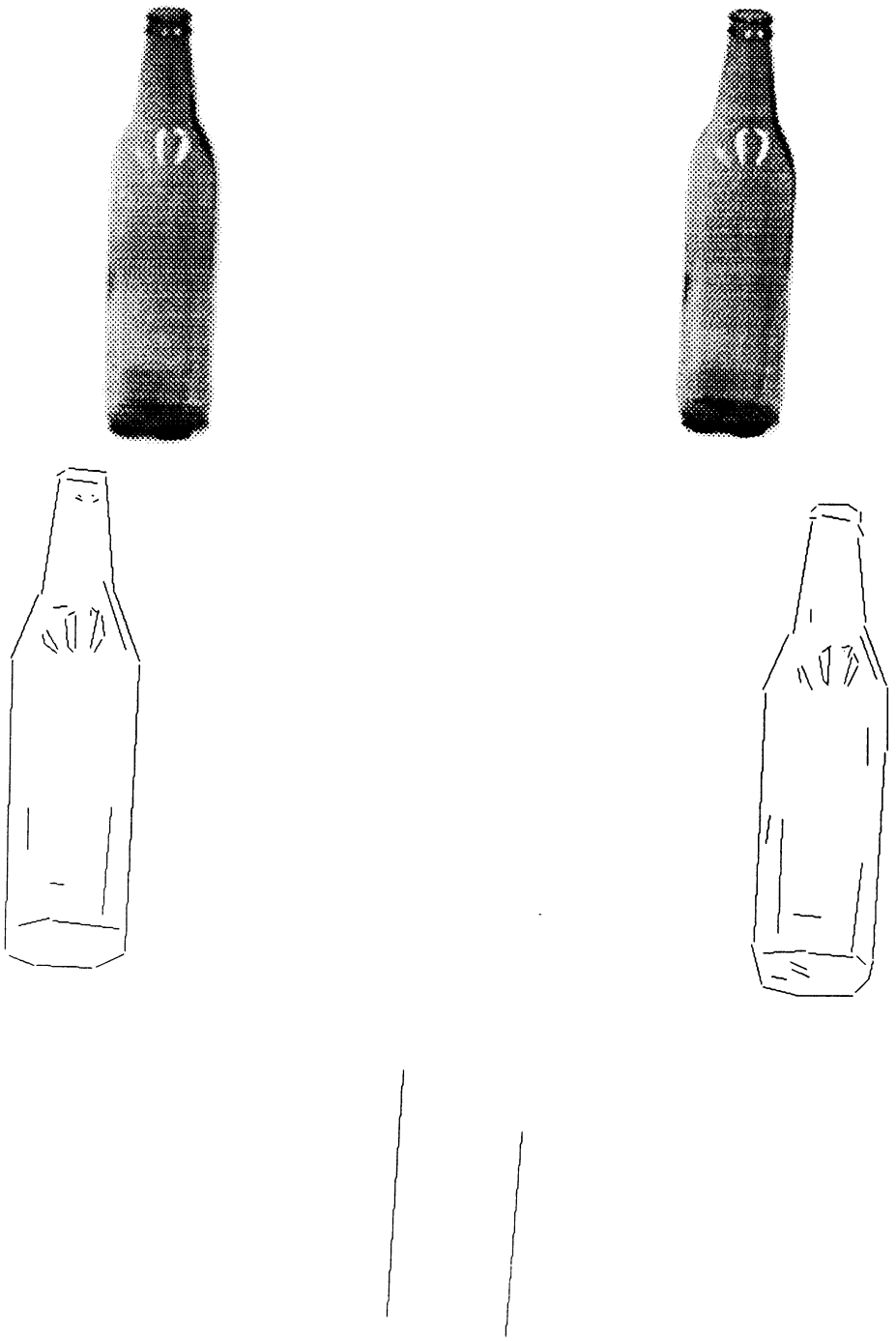

Figure 2: a) Image of beer bottle b) linear features c) matched features

4. SUMMARY

We have described a method whereby the difficult problem of actively controlling a robotic hand can be helped by supplying small amounts of stable and reliable visual data. This visual data, given in the form of the 3-D orientation vector, is then used by the active hand system to explore an object. Our next step is to use all three EP's together to recover the exact shape of a 3-D object.

5. ACKNOWLEDGEMENTS

This work was supported in part by DARPA contract N00039-84-C-0165, NSF grants DMC-86-05065, DCI-86-08845, CCR-86-12709, IRI-86-57151, North American Philips Laboratories, Siemens Corporation and Rockwell Inc. Thanks to Ken Roberts and Paul Michelman for helping to implement the EP's on our hand system.

References

1. Allen, Peter, "Integrating vision and touch for object recognition tasks," *International Journal of Robotics Research*, vol. 7, no. 6, pp. 15-32, 1988.
2. Allen, Peter, Paul Michelman, and Kenneth S. Roberts, "An integrated system for dextrous manipulation," *IEEE Conference on Robotics and Automation*, pp. 612-617, Scottsdale, AZ, May 15-19, 1989.
3. Allen, Peter and Kenneth S. Roberts, "Haptic object recognition using a multi-fingered dextrous hand," *IEEE Conference on Robotics and Automation*, pp. 342-347, Scottsdale, AZ, May 15-19, 1989.
4. Allen, Peter K. and Paul Michelman, "Acquisition and interpretation of 3-D sensor data from touch," *IEEE Workshop on Interpretation of 3-D scenes*, Austin, TX, November 27-29, 1989.
5. Henriksen, Knud, "Line-based stereo matching," MS-CIS-87-52, Grasp Lab 109, Department of Computer and Information Science, University of Pennsylvania, Philadelphia.
6. Kak, A. and S. Chen editors, *Proceedings of the AAAI Workshop on Spatial-Reasoning and Multisensor Integration*, Morgan Kauffman, Los Altos, CA, Oct. 1987.
7. Lederman, Susan and Roberta Klatzky, "Hand movements: A window into haptic object recognition," *Cognitive Psychology*, vol. 19, pp. 342-368, 1987.
8. Singh, Ajit and M. O. Shneier, E. W. Kent, M. O. Shneier, and R. Lumia, "PIPE: Pipelind image processing engine," *Journal of Parallel and Distributed Computing*, no. 2, pp. 50-78, North American Philips Laboratories, Briarcliff Manor, NY, 1985.
9. Wolff, Lawrence B., "Measuring the orientation of lines and surfaces using translation invariant stereo," *SPIE Conference on Sensor Fusion*, vol. 1003, Cambridge, MA, November 1988.