Object Apprehension Using Vision and Touch

R. Bajcsy and S.A. Stansfield
University of Pennsylvania
Philadelphia, PA 19104

1. Abstract

We define object apprehension as the determination of the properties of an object and the relationships among these properties. We contrast this with recognition, which goes a step further to attach a label to the object as a whole. Apprehension is fundemental to manipulation. This is true whether the manipulation is being carried out by an autonomous robot or is the result of teleoperation involving sensory feedback. We present an apprehension paradigm using both vision and touch. In this model, we define a representation for object apprehension in terms of a set of primitives and features, along with their relationships. This representation is the mechanism by which the data from the two modalities is combined. It is also the mechanism which drives the apprehension process.

2. Introduction

It has been suggested by both psychologists and perceptual roboticists that objects are defined in terms of their parts and features. It has also been suggested that these features determine not only our recognition of objects, but also our interactions with them. It seems reasonable to say that these features (along with their relationships) are the outputs of the perceptual system. Our first task, then, in building a robotic perceptual system is to determine what this fixed set of features will be. This is also vital to our representation pardigm, since these features must form the building blocks of objects which are to be manipulated by the system. We have chosen a hierarchical representation for objects which consists, at the lowest level, of a set of modality dependent primitives. Examples of such primitives for touch include roughness, compliance, and several types of contact. For vision, primitives are region points and edges. These primitives are extracted by the sensors and then combined into successively more abstract features. During this process, the information from the two modalities is integrated and a symbolic representation is created. For example, a tactile edge and a visual edge are combined into a supermodal entity "edge" which may then be combined with other edges to form a contour. This contour may eventually be labelled as a rim and the rim determined to be part of a pan.

It is interesting to note that each modality measures only some apsect of reality. Take edges for example. The reality of an edge or a boundary of an object is what separates the object from other objects or from the background. Objects here can be either solid, gas or liquid. Sensors, however, only measure only some aspect of this reality. Hence the tactile sensor will measure and detect an edge as the difference between two substances, while the visual sensor will detect an edge as the difference between two colors or brightnesses. This object-background problem is similar to the figure-ground problem of psychophysics. In order to determine what an edge is, we must first determine what the difference is between the object and the background. This leads us to the question of calibration of the background. In our world we assume that the background is air and that the objects are solids. Hence a tactile edge is well defined as the difference between any reactive force from a solid and no force at all (the zero force). Needless to say this is a special, though frequent, case in our universe. A visual edge, on the other hand, does not always correspond to a physical edge. Shadows are an example. Hence the supermodal model must include some control strategies directing the individual modalities to calibrate. In particular, it must calibrate for the object-background relationship. The system might, for example, be required to differentiate between solids and gases in space applications, or between solids and liquids in underwater applications. Without this ability to differentiate between object and background, the concept of an edge is not well defined.

The features and primitives identified by the system are also combined into a hybrid model which is a combination of symbolic and spatial information. This representation is a loosely coupled collection of features and their relations. We call this representation a *spatial polyhedron* and it is, essentially, a user-centered guide to the features of an object and the relationships among these features in space. The identification of the features and relations of an object and their mapping to this spatial polyhedron then constitutes apprehension.

Interestingly, apprehension may be all that is required for grasping: visual apprehension, which gives us global shape and size information, would drive the initial stages of a grasp, such as hand shaping and bringing the manipulator into contact with the object. Tactile apprehension of such features as temperature and roughness would then aid in the fine adjustment stage of the grasp, when information such as weight and smoothness of an object is vital. Our spatial polyhedral representation provides both the visual cues, such as global size, and the tactile cues, such as roughness, which seem to be important to perceptually driven grasping. In addition, it provides

information about where the manipulator might expect to encounter each of the parts of an object, both in space (for the initial reach) and in relation to one another (for specific grasps and manipulation).

In the remainder of this paper, we present and discuss the issues involved in designing such a system.

3. System Configuration Issues

In this paper, we discuss the structure of a robotic perceptual system and the integration of information from different modalities. Let us begin, then, by presenting a system configuration to serve as a framework for this discussion. What are the issues? First, we consider the type of sensing desired. Sensors are often categorized as either contact or non-contact. Within these two broad classifications, we may place many different types of sensing devices, however, all devices within a classification have several characteristics in common. Contact sensors sense locally -- to gather a large amount of data requires sequential processing of the object. Contact sensors measure surface and material properties directly, for example temperature. And finally, a contact sensor may change its environment. Non-contact sensors tend more often to be global data gatherers, obtaining large amounts of information in parallel. They do not, as a rule, act on their environment (although they may -- for example, a vision system may have its own light source, changable at will.) Given the very different nature of these two types of sensors, it follows that we may make very different use of them. We have chosen to use one sensing mechanism from each of these categories. Our system makes use of a non-contact vison system and a contact sensor composed of a tactile array and a force/torque sensor. When we speak of the perceptual system later in this paper, we will refer to these sensing mechanisms as modalities. This is a term borrowed from the psychology literature. Our reasons for choosing these two sensing devices are twofold. First, they give us different, but complimentary information about the world. And second, they appear to be the two most important senses, in humans, for both recogniton and manipulation.

The next issue which we must consider is how we are to use our devices. Both the visual and the touch systems may be used either actively or passively. Obviously, people use both actively (by which we mean that they are able to control the parameters of the devices at will.) It makes sense that a robot system should also be able to use both modalities actively. However, this is less vital for the vision system, which is able to gather large amounts of data in a single "view", than it is for touch. What is clear, however, is that each sensor is capable of only a partial view of its environment at any one time, and so it is imperative that at least one of the modalities be active.

The final issue is the coupling of the sensing devices. When two devices are coupled, changing the parameters of one will effect the parameters of the other. When they are uncoupled, then each may work independently of the other. One can imagine industrial applications in which the coupling of sensors which represent different modalities would serve the purpose at hand. In the case of general perception, however, it is not clear that the coupling of two modalities will provide any benifits since the information gathered by such devices is conceptually different. What does make sense, though, is to couple two sensors which provide different cues within the same modality -- force/torque and cutaneous, for example. Thus the feedback from one may be used to interpret the information from the other. One can think of various degrees of physical coupling, ranging from rigid coupling (for instance having several sensors on the same probe-finger) through a distributed system coupled via linkages (like the human arm, hand and body), to a physically decoupled system where each sensor system and/or modality functions independently.

The degree of coupling will have important consequences. We postulate that a necessary condition for integrating different sensory systems is that the world being sensed by those sensors which are to be integrated must remain invariant in space during the time interval in which the measurement is taking place or that the system contain some internal knowledge of the nature of the space-time change. We call this invariance spatial-temporal coherence. In the tightly coupled system, where several sensors are positioned on the same probe, the spatial-temporal coherence is guaranted by the physical setup. The disadvantage of this system is that there is no independent control of the data acquisition systems, although there is an independence in the processing of this data (i.e. of the logical sensors). The other extreme case is when the sensory systems are physically decoupled, hence there is independence in the control of the data acqusition process. In this case, in order to be able to integrate the data, and to guarantee the spatial-temporal coherence, one must introduce a supermodal space where the above conditions will be satisfied.

Thus the coupling of sensors will have an effect on both perception and control, particularly during the data acquisition process. This is manifested especially in the haptic modality where different primitives require different hand movement strategies [5]. However, the pairing of primitives and data acquisition strategies (movement of the probe) is universally true as soon as one accepts the concept of an agile (movable) sensor. Take the visual sensor for example, one positions the sensor so that it captures the optimal view and/or detail, depending upon the need. The open question, of course, is the identification of the parameters which will determine what the best view for a given time and context is.

For the remainder of the paper, we will assume that we are dealing with decoupled vision and agile touch sytems, and that the touch system provides us with tightly coupled force/torque and cutaneous sensors.

4. The Building Blocks of Perception

Primitives are the building blocks of perception. They are the lowest level input to the sensory system and require no inferencing capabilities. They are both modality and device dependent. By defining the primitives, we

define the features, and hence the objects, which our system will be able to handle. Our first step toward building a perceptual system must therefore be the determination of these primitives. Marr [8], for example, embraces this approach for vision when he presents the successively more complex stages of the visual system begining with zero crossings and ending with surfaces. The primal and two and a half dimenisional sketches embody the features of the system. In machine touch, less work has been done. However, studies with human subjects [5], [7] suggest that the haptic system computes information related to an object's form, substance and function. Form includes measurements of shape and size, while substance represents the properties of an object such as compliance and temperature. While it is not our intention to do cognitive modelling, we feel that the human system provides an excellent example of a working haptic system. Therefore, we propose the following seven primitives for touch: surface normals, contact (edge, point and area), roughness, compliance and elasticity [10]. Temperature, weight and size are also appropriate, but we are not able to detect them with our current device. For vision, the choice of primitives is richer still. For the time being, we will use simply two dimensional region points and three dimensional edge elements.

Once the primitives have been determined, the features of the system may be chosen. Important features for touch are contour, edge, global size and shape, and parts. For vision, surfaces, edges and regions are among the features which may be computed. These features, and their relations, form the output of the perceptual system.

5. Integration Techniques

Given a robot system with multiple sensors, we would like to somehow process and combine the information from each for further use by the system. We refer to this aggregation of disparate sensory data as integration, and it is currently an active topic of research. Several techniques for integration have been explored, each of them taking a very different approach to the problem. Three projects within the Grasp Lab of the University of Pennsylvania illustrate this diversity.

Durrant-Whyte [3] takes a purely mathematical view of the problem. In his research, all sensors are considered as independent agents. The system contains a world model, and the goal is to maintatin the consistency of this model. Objects are modelled as geometric positions using homogeneous transforms, and uncertainty in these positions is modelled as a contaminated gaussian. Integration is achieved mathematically using a baysian statistical model of the sensory data. Resulting changes in the position of the object being sensed are propagated throughout the world model to maintain consistency. There are two aspects of Durrant-Whyte's work as it currently stands which do not make it adequate for perception. The first is that it requires a world model. A perceptual system should make no intial assumptions about the world. The second is that it represents all objects geometrically as homogeneous transforms. Such a representation in not adequate for apprehension or recognition.

Allen [1] applies well-known modelling techniques to the integration problem. The goal is object recognition and objects are modelled geometrically using a Coon's patch representation. Vision and touch are used in a complimentary fashion: Passive vision is first used to define the regions to be explored and to make an initial fit of the data. Touch is then used to explore the regions and to build successively better approximations of the surfaces. In this way, the information from the two modalities is integrated at the level of the geometric model. This instantiation is then matched against a data base of objects created using a CAD/CAM system. Allen's system suffers from the limitations imposed by the use of geometric modelling techniques. It can only recognize precisely modelled objects, although some variation may be allowed by the use of bounds on particular parameters of the object. The recognition of generic objects is not possible.

In our work [9], the goal is object apprehension of generic objects. By apprehension we mean the determination of the properties of an object and the relations among these properties. As we stated earlier, passive vision and one-fingered active touch are used. Objects are modelled symbolically using a hierarchy of frames: frames at the lower levels represent the primitives and features specific to each modality. Intermediate levels represent super-modal features and parts, and at the highest level is a representation of the object as a whole. As the system explores an object, it extracts and identifies the modality dependent primitives and features. Other modules in the system then combine this information into the supermodal entities described above. (As we said in the introduction, this supermodal model must contain the basic physical assumptions about substances (solid, gas and liquid) and the laws that apply to them. These laws, and the subsequent properties which they imply, will then be translated to the individual modalities in terms of expectations (or hypotheses). An important result will be the establishment, for the given world, of the object-background relationship from which the calibration procedure will follow.) Integration within this system occurs at the symbolic level, as modules gather primitives and features (which are themselves already symbolic) and combine them into more abstract entities.

There are several reasons why we have chosen this structure. First, by defining our primitives based upon the sensory systems available, and not upon the objects to be considered, we hope to build a more generalized perceptual system. Because the goal is to apprehend, and eventually recognize, generic objects, it does not make sense to require specific models of each individual object. For the same reasons, we will need to be able to reason about our object categories, both for recognition and for exploration. Reasoning falls into the domain of Artificial Intelligence, and it is from this field that we have chosen to take our representational paradigm. There is also a psychological basis to our design. It has been suggested [11] that humans reason about objects based upon parts and features. Therefore, it seems reasonable to have our perceptual system compute such features and parts.

6. Using Vision to Guide Touch

In manipulation, of which we may consider tactile exploration a subset, there appear to be two stages. First, there is the reach -- a gross motion and orientation mechanism using the arm; then there is the fine adjustment and manipulation stage using only the wrist and fingers. The former is likely feed forward, while the latter uses feedback. It seems reasonable to suggest that the initial reach and hand-shaping is visually-guided, while the fine manipulation (or exploration) is primarily tactile in nature. As a matter of fact, the very properties which each system is most adept at extracting are imperative to the stage at which we suggest it is used. Thus: vision is excellent at determining position in space, rough size and shape, and part segmentation. These are the very parameters required for the initial reach and hand-shaping. Once contact with the object has been made, however, vision may no longer be useful. Often the object is occluded by the grasping hand, or contacting finger. In addition, the parameters important to successful manipulation (again, we use the term to encompass exploration) may not be easily computed by the visual system. One example is the use of roughness and temperature to access the possibility of slip during a grasp. Another is the use of kinesthetic feedback for positioning of the finger during exploration*.

It therefore makes sense to take a "look before you touch" approach, and that is what we have done. The vision system operates first, obtaining initial position, segmentation, and orientation information. This information is used to drive the initial reach and positioning of the finger on the object. The haptic system then takes over to do the tactile exploration. In our case, since we have only a single finger, we choose to approach the object several times and from several different directions in order to fully do the exploration. Because we have no a priori knowledge of the object, and only partial information from our visual system, we need a general exploration method. We use a representation called the spatial polyhedron to accomplish this. The spatial polyhedron is a collection of approach planes. Mapped onto the face of each of these planes is the set of features of the object which one might expect to encounter while exploring the object from that direction. Thus the robot aproaches and contacts an object from each of a set of predetermined orientations. It then invokes the haptic system to explore the features encountered. The end result is a set of extracted features and their relations as defined both implicitly by the relations among the faces of the polyhedron and explicitly by the relations of each feature on a given face. This is in fact apprehension as we have defined it.

7. Some Thoughts About Grasping

We believe that the structure of our perceptual system, and its attendant representations, will extend painlessly to multi-fingered grasping. We have tried to keep the primitives and features dependent only on the modality. Hence they should be as easily computable by several fingers as they are by one. Since the integration within the system occurs at the symbolic level, any number of sensors may input information. New information available only to a multi-fingered hand, such as weight and gross size, can be easily incorporated. Finally, the method by which the reach and object contact are made is designed specifically to be generalizable to a hand, and the spatial polyhedron will allow the simultaneous extraction and aggregation of features from several positions on the object.

Finally, there is evidence that, in humans, grasping and manipulation are perceptually driven, and that the mechanisms for manipulation, such as hand shaping, may actually be part of the stored representation of an object [4]. Thus the development of a haptic perception system, the integration of visual and tactual cues, and the mechanism for visually-guided touch would all appear to be vital to the development of such a perceptually driven manipulation system.

8. Conclusion

In this paper we have presented the framework of a bimodal (contact and non-contact) robotic perceptual system. The concrete study of this general problem is done by investigating vision and touch. Within this framework we have discussed such issues as the system configuration, the choice of perceptual primitives, the integration technique and how vision is used to guide tactile information acquisition. We have further analyzed the consequences of the degree of physical coupling of different sensory systems. We introduce the concept of spatial-temporal coherence and postulate that a necessary condition for integrating different sensory systems is that the world which is being sensed by those sensors which are to be integrated must remain invariant in space during the time interval for which the measurements are taking place or that the system contain some internal knowledge of the nature of the space-time change. Furthermore, the supermodal model must contain facts about the physical world that are true independent of the individual sensors, but that describe the particular world in which the robot must function. This in turn will determine the parameters for calibration of the object-background relationship in the supermodal world, which will then will be translated for the individual modalities.

9. Acknowledgements

The work described herein was performed in the Grasp Lab of the University of Pennsylvania and was supported in part through the following grants: ARO DAA6-29-84-k-0061; AfOSR 82-NM-299; NSF MCS-8219196-CER; NSF MCS 82-07294; AVRO DAABO7-84-K-FO77; NIH 1-RO1-HL-29985-01 DARPA/ONR N0014-85-K-0807.

*A further example of the different ways in which visual and tacile information is processed by the human perceptual system involves the perception of texture [6]. While visual texture is primarly used for grouping and segmentation purposes, the tactile texture determines the properties of a surface, such as roughness. This difference also shows up the data acquisition process: The visual texture detector must be applied over the entire scene, while the tactile texture detector need be applied only locally.

10. References

1. Allen, P. *Object Recognition Using Vision and Touch.* Ph.D. Th., University of Pennsylvania, December 1985.

2. Bajcsy, R., D. Brown, J. Wolfeld, and D. Peters. What Can We Learn From One Finger Experiments? A Report on a Joint US-France NSF-CNRS Workshop on Advanced Automation and Robotics, June, 1982, pp. 119-158.

3. Durrant-Whyte, H. *Integration, Coordination and Control of Multisensor Robot Systems.* Ph.D. Th., University of Pennsylvania, August 1986.

4. Klatzky, R., B. McCloskey, S. Doherty, J. Pellegrino, and T. Smith. Hand Shaping and Object Processing. Cognitive Science - 8605, University of California at Santa Barbara, 1986.

5. Klatzky, R. and S. Lederman. Hand movements: A Window into Haptic Object Recognition. Paper presented at the 26th Annual meeting of the Psychonomic Society, Boston, Nov. 1985.

6. Lederman, S., G. Thorne, and B. Jones. "The Perception of Texture By Vision and Touch: Multidimensionality and Intersensory Integration". *Journal of Experimental Psychology: Human Perception 12*, 2 (1986), 169-180.

7. Lederman, S. and R. Klatzky . Knowledge-based Control of Human Hand Movements. Paper presented at the conference on Biomechanics and Neural Control of Movement, Henniker, NH, July 1985.

8. Marr, D.. *Vision.* W. H. Freeman and Co., 1982.

9. Stansfield, S. Representation and Control within an Intelligent, Active, Multisensor System for Object Recognition. Proceedings of the IEEE Conference on Systems, Man, and Cybernetics, November, 1985.

10. Stansfield, S. Primitives, Features, and Exploratory Features: Building a Robot Tactile Perception System. Proceedings of the IEEE Conference on Robotics and Automation, April, 1986.

11. Tversky, B. and K. Hemenway. "Objects, Parts, and Categories". *Journal of Experimental Psychology 113*, 2 (June 1984), 169-193.