Object Recognition by a Robotic Agent: the Purposive Approach

Ehud Rivlin, Yiannis Aloimonos and Azriel Rosenfeld
Computer Vision Laboratory
Center for Automation Research
University of Maryland
College Park, MD 20742-341

Abstract

We study the problem of object recognition by considering it in the context of an agent operating in an environment, where the agent's intentions translate into a set of behaviors. In this context, an object can fulfill a function; if the agent recognizes this, it has in effect recognized the object. To perform this type of recognition we need on one hand a definition of the desired function, and on the other the means of determining whether the object can fulfill that funaction. To illustrate this approach we describe the visual recognition abilities that might be needed by an autonomous cleaning robot.

1 Introduction

Before we design successful vision systems that can recognize objects, situations or any other patterns in their environment in real time and act appropriately in each situation, we must ask a set of basic questions. Our design methodology and most of the research that we will need to perform will depend directly on our answers to these questions. Among the questions are the following: What kind of information should a visual system derive from images? Should this information be described in some kind of language? If not, should the information be in a single general purpose form, leaving it to other modules to transform it to suit their needs, or can a visual system directly produce forms of information suited to specific modules? Are descriptions of 3-D location and structure the only descriptions that should be produced by a visual system? Are there sharply distinguished modules for vision, reasoning, memory, planning, etc., or are the boundaries blurred and different sub-systems closely integrated with one another?

When we work in the generally accepted paradigm of general recovery [5] we have already given an answer to the questions above. Indeed, if we consider vision as a recovery problem, i.e. if we maintain that the goal of visual perception is to create an accurate 3-D description of the scene (shape, location and other properties) which will then be given to other cognitive modules (such as planning or reasoning), then we are considering vision in isolation as a module of the whole system. Our answers to the previously mentioned questions are then obvious. Following this line of thought, most object recognition research can be classified in the "recover and match" paradigm, where first shape and properties of the object are recovered and then they are matched with stored versions of the objects that the system "knows" about. For this to happen, we need models that can describe the shape of a variety of objects using a few parameters; the literature is quite rich in such models [2, 3, 4]. Nevertheless, there continues to be a lack of vision systems that can recognize a variety of man-made or natural objects in real time.

With this in mind, and the realization that complete recovery of an object's shape and attributes is very hard [1], we propose an alternative approach to the problem of visual recognition. It is clear that a name or any symbol associated with an object means little to a vision system that needs to interact with that object. We need to ask the question: What are objects for? Objects can suit a purpose, fulfill a function. If the observer recognizes this, in effect he/she has recognized the object. To perform this type of recognition we need on one hand a definition of the desired function, and on the other the means of determining whether the object can fulfill that function.

[0] This work was funded in part by DARPA (ARPA Order No. 6989, through Contract DACA76-89-C-0019 with the U.S. Army Engineer Topographic Laboratories), NSF (under a Presidential Young Investigator Award, Grant IRI-90-57934), Alliant Techsystems, Inc. and Texas Instruments, Inc. The authors thank Barbara Burnett for her expert help in preparing this paper.

To find out if an object can fulfill a function we need to perform various *partial* recovery tasks.

Different vision systems working in different environments and in general performing different visual tasks do not necessarily recognize objects using similar algorithms. A vision system that needs to recognize only, say, ten types of objects from a very large set does not necessarily work in the same way as a vision system that needs to recognize two types or 500 types. A visual recognition system that serves a high velocity moving agent is not necessarily built in the same way as a vision system for a stationary agent. We propose that object recognition should be studied by also taking into account the agent that has to perform it. An agent is defined as a set of intentions and a set of capabilities for carrying them out given as parameters along dimensions such as size, mobility, sensory modality, etc. Since different agents, with different purposes, working in different environments do not recognize visually in the same manner, it would not make sense for us to seek a general (universal) theory for object recognition. Instead, we should concentrate on developing a methodology that, given an agent in an environment, will suggest how to perform a particular recognition task.

2 Purposive and Qualitative Object Recognition

We offer an understanding of recognition within the frame of the agent. An agent is defined as a set of intentions, $I_1, I_2, \ldots, I_n$. Each intention I_k is translated into a set of behaviors, $B_{k1}, B_{k2}, \ldots, B_{km}$. Each behavior B_{ki} calls for the completion of recognition tasks $T_{ki1}, T_{ki2}, \ldots, T_{kij}$. The agent acts in behavior B_{ki} under intention I_k. The behavior calls for the completion of recognition tasks $T_{ki1}, \ldots, T_{kin}$. The behavior sets parameters for the recognition tasks. Note that the same object can answer positively to several recognition tasks. Under some specific behavior a chair will answer yes to some recognition task that is asking for obstacles, under another behavior it will answer yes to a recognition task that is asking for a sitting place, and under another it will answer yes to a task that is asking for an assault weapon.

The recognition process is viewed along the axis intention, behavior, recognition task. In order for us to have a full and complete theory of purposive object recognition we should be able to make the two basic transformations: first from the desired intention to the set of behaviors that achieve it, second from a specific behavior to some needed recognition task(s). In what follows we will present some of the problems concerning these transformations.

It is possible to give a lower bound on the general intention-to-behaviors transition. It can be showed that the intention-to-behaviors problem with a finite number of behaviors is undecidable by reducing it to the halting problem. If we add constraints to our definition of the problem we can move from undecidability to intractability. For example, by constraining ourselves to a constant set of objects we can show a PSPACE-hard lower bound.

For the translation from behavior to a recognition task we should have the following transformations. One transformation that we have is from that general function needed to a collection of sub-functions that the object has to provide. This transformation is not trivial. In addition, we need another transformation into the needed perceptual data. Objects have limited observable characteristic information from which we can infer which functional category the object belongs to. For example, does the object appear to be immobile or mobile? (It can be momentarily stationary.) Is the object graspable? Does the object appear to be organic or inorganic (animal, vegetable, mineral)? These are functional relationships (here functional is been used in the utilitarian sense), which can be translated, for example, into surface characteristics and geometric properties in a crude qualitative way. It should be emphasized, however, that not all functional relationships can be determined by vision alone. Vision can only extract a limited set of object properties such as shape, color, motion, etc. There exists a plethora of object properties, such as electric, magnetic, chemical, thermal and mechanical, that can be detected by specialized sensors other than visual. For example, it is impossible to recognize that an unfamiliar object[1] is hard or rigid (especially when it is not moving) using vision, while it is an easy task if we use taction.

To summarize: under our framework an agent acts in behavior B_{ki} under intention I_k. The behavior calls for the completion of recognition tasks $T_{ki1}, \ldots, T_{kin}$. The behavior sets parameters for the recognition tasks. Each recognition task activates a different collection of basic perceptual modules. Each module qualitatively finds a generic object property which is a result of one or a combination of direct low-level computations on some sensory data (possi-

[1] Interaction with an object, experience, and learning attach many properties to the object that were not acquired by vision alone.

bly done by other modules). The result of a module's operation is given as a qualitative value. Each module has its own neighboring open intervals which are parameter-specific. The i^{th} module can take one of $q_{i1}, \ldots, q_{in}$ qualitative values.

The state of our recognition system, denoted by Q_i, is a tuple of all the qualitative values of our modules $(q_1, \ldots, q_m)$ under recognition task T_{kij}. Each recognition task T_{kij} defines a system state that will constitute a positive answer to that recognition task. Recognition is done when we complete our task, which means a stable answer from our modules. At this point we want to remark that a common recognition task can be defined as a new module.

We have described the recognition process along the axis intention, behavior, recognition task (i.e. top-down). This process is well suited to a purposive agent performing active vision. From this point on we will restrict ourselves to a single recognition task T_{kij} under behavior B_{ki} and intention I_k. We will assume that some parameter setting is done by the intention and the behavior. These parameters fix the setting for the current recognition task, which includes the required system state (some of the modules might be in the don't care position) and possibly some additional "common knowledge" parameters, such as environmental parameters (outdoor, indoor), predator, size, etc. From this point of view the recognition process is using high-level information.

3 An autonomous cleaning robot - the needed visual abilities

Following the above methodology we describe the design of an agent equipped with visual sensors whose purpose in life is to clean corridors. Here we are only interested in the recognition tasks and not in the control aspects of this problem. We consider the environment to be a single corridor, thus concentrating on the visual problems involved and avoiding problems related to planning a path through the building. It is worth noting that some of the visual abilities of this cleaning robot constitute a basic skeleton that is common to any autonomously moving agent in an indoor environment.

Working top down from intentions to behavior we got the following seven behaviors:

- Cleaning the corridor: Move and clean (M&C)
- Without hurting human beings: Stop and wait (when a human is within some distance) (S&W)
- Detecting obstacles: Obstacle (OBS)
- Picking up small objects: Pick up small object (PICK)
- Pushing aside medium-sized objects: Push (PUSH)
- Bypassing large obstacles: Bypass (BYPASS)
- Completing the task: (STOP)

The following recognition tasks needed to implement the various behaviors [2]:

- Recognize the main cleaning area (i.e. find free space, the main axis of the corridor along which the robot needs to move as well as the position of the floor): MCA
- Recognize independently moving objects: IMO
- Recognize static obstacles and their distances from the floor: OBS
- Recognize the size category of the obstacle: SIZE (big, small, med)
- Recognize dead-end (no free space): DEAD-END

The following visual modules are sufficient to implement the above mentioned recognition tasks [3]:

- Computing normal flow series: VM1
- Computing time-to-collision hazard maps: VM2
- Detecting anomalies (and finding the floor): VM3
- Detecting size: VM4
- Steering: VM5

The recognition tasks will be accomplished using in parallel the above visual modules [4] Clearly, there exists a vast amount of redundancy in the above modules, which, however, contributes to the robustness of the main behavior. The density of features in the image can be used for assigning various degrees of confidence to the results of the visual processing described.

In figure 1 we see the *behavior-transition* diagram where each node represents a distinct behavior. To

[2] A systematic way to do the translation is presented in the full paper

[3] A full description of the visual modules needed can be found in [6]

[4] For example, to recognize the main cleaning area we use the time to collision,detecting anomalies and the steering visual modules.

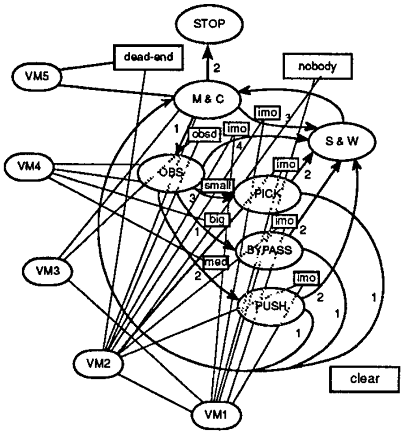

Figure 1: The behavior-transition diagram. The arcs represent those recognition tasks that trigger a different behavior. The numbers on the artcs indicates priority (large number indicates high priority). The different visual modules that are needed for each behavior are connected appropriatly with lines the the specific behavior.

accomplish a behavior we need to carry out several recognition tasks which are not shown in the diagram. The arcs represent those recognition tasks that trigger a different behavior [5]. Since various events may happen simultaneously a priority schema is needed to arbitrate (large number indicates high priority). The different visual modules that are needed for each behavior are connected appropriatly with lines the the specific behavior.

4 Conclusions

We have presented an alternative approach to the problem of object recognition. Instead of the common bottom-up process of recovery, followed by fitting to a model and matching in a database, we have formulated the problem as a top-down process. Recognition is studied in terms of the agent performing it, under its intentions and the behaviors triggered by them. From this point recognition translates partially to a verification process that checks for the existence of physical properties that provide needed functionality. In this way partial recovery of the scene is sufficient and can be performed in a robust manner using qualitative techniques. The above methodology was demonstrated for an autonomous cleaning robot.

We see the main contribution of this paper as being the methodology proposed for building machines with vision (intentions, behaviors, recognition tasks under some environment). Our approach should not be confused with unexpected object recognition [7], which is a different problem from the one studied here and is not directly relevant to the construction of robust machine vision systems. To better understand our approach, one should think of recognition as utilization, i.e. we recognize an object when we know enough about it so that we can utilize it. This problem of utilization is easier than the traditional problem of recognition which amounts to assigning symbols to perceptual data.

References

[1] Aloimonos, J. Y. (1990). "Purposive and qualitative active vision," *Proc. of the DARPA Image Understanding Workshop*, pp. 816–828.

[2] Binford, T.O. (1982). "Survey of model-based image analysis systems," *International J. of Robotics Research*, Vol. 1, pp. 18–64.

[3] Besl, P.J., & Jain, R.C. (1985). "Three dimensional object recognition," *Computing Surveys*, Vol. 17, pp. 75–145.

[4] Chin, R.T., & Dyer, C.R. (1986). "Model based recognition in robot vision," *Computing Surveys*, Vol. 18, pp. 67–108.

[5] Marr, D. (1982). "VISION: Computational Investigation into the Human Representation and Processing of Visual Information", Freeman, San Francisco.

[6] Rivlin, E. & Aloimonos, J. Y. (1991). "Purposive recognition: an active and qualitative approach " *SPIE Proc.*, Vol 1611, Boston, MA.

[7] Rosenfeld, A. (1987). "Recognizing unexpected objects:A proposed approach," *International J. of Pattern Recognition and Artificial Intelligence*, Vol. 1, pp. 71–84.

[5] these recognition tasks (outgoing edges) are running in parallel to whatever activity is taking place during any behavior.