Attention and Visual Search:
Active Robotic Vision Systems that Search

John K. Tsotsos and Ksenia Shubina

Dept. of Computer Science & Engineering, and
Centre for Vision Research,
York University, Toronto, Canada
{tsotsos, ksenia}@cse.yorku.ca

Abstract

Visual attention is a multi-faceted phenomenon, playing different roles in different situations and for different processing mechanisms. Regardless, attention is a mechanism that optimizes the search processes inherent in vision. This perspective leads to a sound theoretical foundation for studies of attention in both machine and in the brain. The development of this foundation and the many ways in which attentive processes manifest themselves will be overviewed. One particular example of a practical robotic vision system that employs some of these attentive processes will be described. A difficult problem for robotic vision systems is visual search for a given target in an arbitrary 3D space. A solution to this problem will be described that optimizes the probability of finding the target given a fixed cost limit in terms of total number of robotic actions the robot requires to find its visual target. A robotic realization will be shown.

1 Introduction

Attention is one of those visual phenomena that has been very easy to ignore in computer vision and robotics but seems to now be emerging as an important issue. Visual attention is a phenomenon that has been of interest to many disciplines for hundreds of years, with an enormous literature and thousands upon thousands of experiments investigating the vast range of its manifestations. Theoretical and computational models have been proposed since the 1950's in an attempt to explain how his phenomenon comes about and how it contributes to our perception of the real world (for a review see Rothenstein & Tsotsos 2006). The first formal proof for the necessity of attentive processes appeared in Tsotsos 1989 (see also Tsotsos 1987, 1990, 1992; Rensink 1989). There, the problem of visual matching - the task of determining whether or not an instance of a particular model exists in a given image without the use of any knowledge whatsoever - was shown to be NP-Complete. This means that is has exponential time complexity, in the size of the image, and further, the result is independent of implementation. In addition to other mechanisms, attention contributes to changing this problem into one with linear time complexity.

The breadth and variety of attentive phenomena as they relate to computer vision was first described in Tsotsos (1992). There, a spectrum of problems requiring attention was laid out including: selection of objects, events, tasks relevant for a domain; selection of world model; selection of visual field; selection of detailed sub-regions for analysis; selection of spatial and feature dimensions of interest; and the selection of operating parameters for low level operations. Most computer vision research makes assumptions that reduce the combinatorial problems inherent in the above tasks, or better yet, eliminate the need for attention, using strategies such as:

- fixed camera systems negate the need for selection of visual field or considerations of viewpoint invariant vision;
- pre-segmentation eliminates the need to select a region of interest;
- 'clean' backgrounds ameliorate the segmentation problem;
- assumptions about relevant features and their values reduce their search ranges;
- knowledge of task domain negates the need to search a stored set of all domains;
- knowledge of objects appearing in scenes eliminates search of a stored set of objects; and,
- knowledge of which events are of interest eliminates search of a stored set of events.

In this way, the extent of the search space is seriously reduced before the visual processing takes place, and often even before the algorithms for solution are designed. However, it is clear that in everyday vision, and certainly in order to understand vision, these assumptions cannot be made. Real vision, for humans as well as robots, is not so cooperative and attentive processes need to play a central role in all visual processes.

The example robotic system described later in this paper looks at the viewpoint and selection of visual field issues in the context of search for a given, known object in an unknown 3D world. As such, it is an instance of the active vision approach (Bajcsy 1985). Bajcsy argued that rather than simply analyzing a set of prerecorded images, the observer should actively control its image acquisition process so that the acquired images are relevant and useful for the task at hand. In the case of region segmentation problems, the camera could be moved to a viewpoint in which, for example, a given object projects to a higher contrast region, or an objects edge projects to a higher gradient in the image. If a particular view of an object (or one of its parts) is ambiguous, the camera can be moved to disambiguate the object. For example, Wilkes and Tsotsos (1992) proposed a system that drives a camera to a standard viewpoint with respect to an unknown object. From such a viewpoint, the object recognition task is reduced to a two-dimensional pattern recognition problem. The authors choose to define a standard view as a position at which the lengths of two non-parallel object line segments are maximized, and the longer line has a specified length in the image. The standard view is achieved by moving the camera on the end of a robot arm. From a standard viewing position, the extracted line segments are used to index into the database to find a matching, stored object. In a different strategy for the same problem, Dickinson et al. (1997) combine an attention mechanism and a viewpoint control strategy to perform active object recognition. Their representation scheme is called the aspect prediction graph. Given an ambiguous view of an object this representation first tells if there is a more discriminating view of the object. If yes, the representation will indicate, in which direction the camera should be moved to capture that view. Finally, it specifies what visual events (appearance or disappearance of object features) one should encounter while moving the camera to the new viewpoint. In both cases, the image interpretation process is tightly coupled to the viewpoint selection and data acquisition process, as Bajcsy suggested. The success of these works lies in the fact that no assumptions about viewpoint were needed and attentive processes - selection processes - provided the reduction in the combinatorics of search that would cripple a brute-force, blind, search.

The remainder of the paper focuses on the object search problem, providing a brief description of the solution strategy and an example of the robot's performance.

2 A Robot that Searches

Suppose one wishes a robot to search for and locate a particular object in a 3D world. A direct search certainly suffices for the solution. Assuming that the target may lie with equal probability at any location, the viewpoint selection problem is resolved by moving a camera to take images of the previously not viewed portions of the full 3D space. This kind of exhaustive, brute force approach can suffice for a solution. However, this search is both computationally and mechanically prohibitive. As an alternative, Garvey (1976) proposed the idea of indirect search for a target: first a sensor is directed to search for an intermediate object that commonly participates in a spatial relationship with the target. For example, if one wants to find a telephone in an image of an office, it is easier to first locate flat surfaces, e.g. table tops, on which the phone is most likely to rest. Then the sensor is directed to examine the restricted region specified by the relationship, i.e. limit the search for the phone to the table tops. Indirect searches reduce the computationally expensive problem to a two-stage problem. In the first stage, one locates an intermediate object that typically participates in some spatial relationship with the target and which can be found with a lower resolution, i.e. with a wider field of view. In the second stage, the high-resolution search for the target is performed in the much smaller volume specified by the spatial relationship.

Wixson and Ballard, 1994, have elaborated the indirect search idea and have shown the efficiency gains both theoretically and empirically, Other demonstrations of the idea have also shown good performance (for example, Reece 1992). The problem with indirect search is that the spatial relation between the target and intermediate object may not always exist. In addition, the detection of the intermediate object may not be easier than the detection of the target. In fact, search for an arbitrary object in a 3D space is provably NP-hard (Ye & Tsotsos 2001).

Searching for an object in a cluttered environment is often complicated by the fact that portions of the area are hidden from view. A different viewpoint is necessary to observe the target. This requirement is also characteristic for the task of scene reconstruction where multiple viewpoints must be selected to acquire a model or a map of the environment. As a consequence, viewpoint selection for search tasks seems similar to viewpoint selection for data acquisition of an unknown scene: new viewpoints are determined by yet unseen areas, e.g. Connolly (1985), Maver & Bajcsy (1993). Wixson (1994) argues that the two tasks of visual search and scene modeling are not that similar. The viewpoint selection problem for search tasks involves not only a choice of position and direction of the sensor, but also a solid angle relative to the viewpoint. The author suggests that the difficult task of scene modeling that usually accompanies visual search is not necessary since the only requirement is that it brings otherwise hidden areas into view. Wixson proposes a model-free algorithm that first identifies an occluding edge and then rotates the sensor to a position where this edge becomes non-occluding. Yet it is unclear what is the criterion to abort the search and what is the strategy to decide on next sensor position.

2.1 The Search Strategy

Ye & Tsotsos define object search as a problem of maximizing probability of detecting the target within a given cost constraint (Ye 1997; Ye & Tsotsos 1999). The formulation combines the influence of a search agent’s initial knowledge and the influence of the performance of available recognition algorithms. For a practical search strategy, the search region is characterized by a probability distribution of the presence of the target. The control of the sensing parameters depends on the current state of the search region and the detection abilities of recognition algorithms. In order to efficiently determine sensing actions over time, the huge space of possible actions is reduced to a finite set of actions that must be considered. The result of each sensing operation is used to update the status of the search space.

There are two characteristics of the approach. The first is decomposition of the huge number of camera’s parameters settings allowed by the hardware into a limited set of settings that must be considered. The second is the decomposition of the action space into a “where to look next” and a “where to move next” tasks. The spatial decomposition greatly reduces the complexity of the “where to look next” task. The second is the emphasis on guidance using a priori knowledge. The advantage of this is that when initial knowledge is relatively good, the most promising actions tend to be selected first. On the other hand, when the knowledge is too far from reality the performance of the algorithm will decrease greatly. At least three kinds of knowledge can be used: locations that are most likely to contain the target, locations of indirect objects that would be useful in finding the target, and, scene saliency tuned to target visual characteristics. The algorithm will operate in its default mode with no a priori knowledge. A brief description of the algorithm follows.

Objective Function

We wish to define an objective function that will direct the search for the target for an active robotic agent. Here, a qualitative description will be provided for this function; the mathematical details can be found in Ye & Tsotsos (1999). Suppose there is some maximum amount of time that is permitted for the search. The robot's task is to maximize the probability of finding the target within that time constraint. The environment is tessellated into cubes; only the boundaries of the 3D space need be known, not the structure of the interior. Each cube has associated with it the probability that the target is in that location. Initial probabilities can be set in a number of ways: no knowledge leads to a uniform distribution, 'hints' lead to higher probability values in specific regions, saliency representations computer with the robots sensing actions lead to an ordering of region to inspect, and so on. These probabilities are updated in Bayesian fashion as the robot searches. The probability of finding the target is maximized when a location's associated probability exceeds an acceptance threshold.

The actions that the robot can take include movement of the robot, movement of the sensor, changes in sensing parameters, and choice of a detection/recognition algorithm. Each of these have a cost in time, both processing time and mechanical time. A cost function for an operation gives the time required for its execution. It includes moving a sensor from one configuration to another, acquiring an image, running a recognition algorithm and updating the search agent's knowledge about the environment. The sum of these costs for each of the robots actions provide the total time to find the target.

Detection Algorithm

The detection algorithm a used in this work is an implementation of an appearance-based recognition method using gradient descent search in an image pyramid based on MacLean and Tsotsos (2000). Their method is an extension of a normalized grey-scale correlation (NGC) technique that has traditionally been slow due to the computational issues. The authors address this limitation by introducing a pyramid structure and a local estimate of the correlation surface gradient. Our search agent can find 2D objects (e.g. toys) in a 3D office environment. This algorithm is largely rotation (in the image plane as well as out of the image plane) and scale invariant.

Where to look next

Perhaps, the most obvious search strategy is to perform a 360 degree pan using only wide-angle settings. This strategy might succeed when looking for a soccer ball, but not for a small battery, because at a wide angle setting the image of the target will be too small. Therefore, the first task is to decide upon camera's angle sizes so that no matter where the target is in the environment, there is an angle that can make its image sufficient for detection. For any given angle size, the target can be detected only when it is within a certain range of distance from the camera called the "effective range". For each camera angle size, a group of viewing directions that examines a given effective range is then selected.

A "non-planning" strategy is to examine all discretized camera configurations one by one. However, when the number of settings increases or image analysis takes too much time, this strategy becomes inefficient. By applying the most promising settings first, the probability of detecting the target at an early stage increases and the time and effort spent on the search decreases. What information should be considered when selecting the best camera's settings? Almost every search process is guided by our knowledge: we always search those regions that we believe are the most likely to contain the target. Here, the knowledge about the potential target locations is encoded as a target probability distribution. By combining the target distribution and detection function, the probability of detecting the target by operation can be calculated. We use this information to select imaging geometries that yield useful results. The details can be found in Ye & Tsotsos (1999) and Shubina (2007).

Where to move next

The previous subsection describes the search strategy when the robot is still. Once the utility of detecting the target at robot's current position becomes small, the search agent should move to a different location. The best next position must satisfy two requirements: it must be reachable and have a high probability of detecting the target. As we have mentioned before, the search space is tessellated into a 3D grid. This tessellation divides the horizontal plane of the search region into a 2D grid. Since the robot moves only horizontally and the height of its camera does not change, only the vertices of the 2D grid are considered as the possible robot positions. For each possible position, the local probability distribution of the local grid elements is examined. The sum of those probabilities in a local region determines the overall strength of a position. The robot moves to the accessible position with highest overall strength.

2.2 An Example

Our search agent is implemented on a Pioneer 3 robot, a mobile four-wheel differentially steered drive ActiveMedia Robotics' platform. The platform is equipped with a Point Grey Research Bumblebee camera mounted on a Directed Perception pan-tilt unit. It is a two lens stereo vision camera that is used for both target detection and environment data acquisition. To obtain depth information, the Bumblebee camera uses Triclops Stereo Vision Software Development Kit that provides stereo processing capabilities. This library does stereo processing on the images obtained from the cameras. It establishes correspondence using the Sum of Absolute Differences correlation method. In addition, it offers a number of validation techniques to reject incorrect correspondences.

Several examples of the search process, including investigations of the performance of the algorithm under differing lighting conditions, different poses of the object, etc. can be found in Shubina (2007). Here, one example is shown. The robot is asked to find a simple colored block (see yellow arrow in Fig. 1) in the environment shown in Fig. 1. The starting position of the robot, shown in Fig. 2, is such that the target is not visible at all.

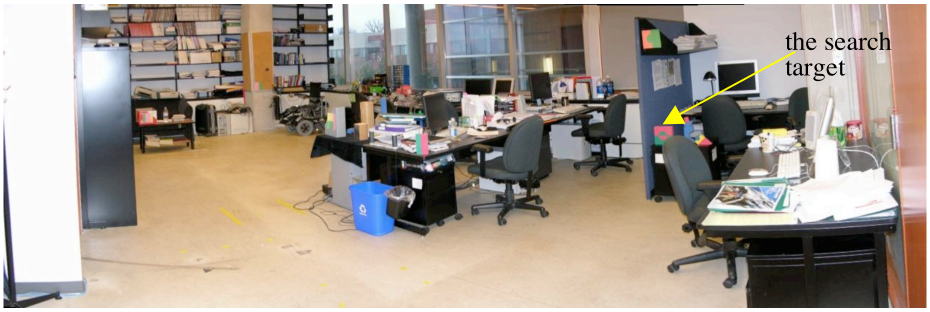


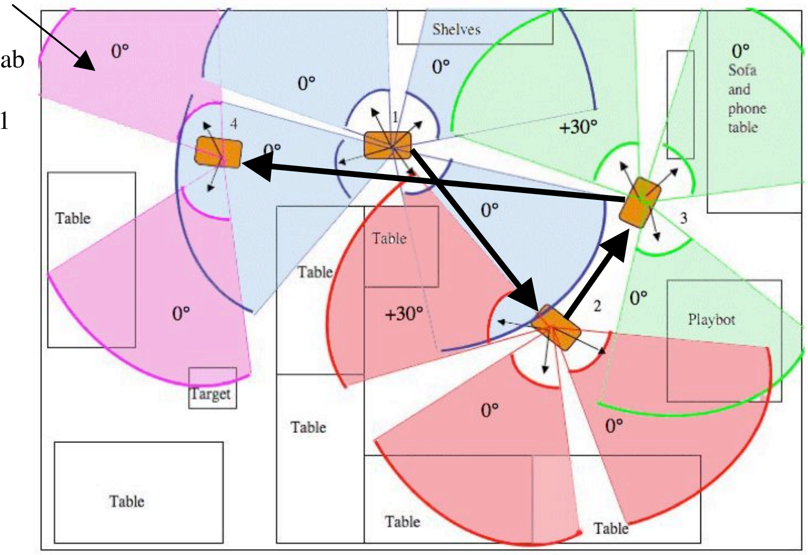
Fig. 1. The Lab Environment

In total, the robot examines its world from 4 positions, and examines several visual angles from each, for a total of 12, as shown in Fig. 2, before finding the target. Each of these 12 fixations entails the acquisition of an image, search across that image using the MacLean & Tsotsos method, and an updating of the probability distribution throughout the search environment using the results of the search. Each fixation is specific for a particular depth of field determined by the visual angle subtended by the search target. This depth of field is termed the effective range of detection - the range at which the target can be successfully recognized, that is, the depths from the camera for which the object will be imaged with a sufficient number of pixels in the image for recognition. For example using our camera system, if a target is 25 cm in diameter, its effective range of detection is 3 m; if a target is 12 cm in diameter, its effective range is 1.5 m. The impact of this on the search process is that the size of the movements required for the search depend on the size of the target: smaller targets lead to smaller movements and thus more of them to cover the search space. The target object that we used in the example here measured 25 cm in diameter.

3 Conclusions

The search for an object in a 3D space benefits greatly from attentive mechanisms that limit the search space in a principled manner. We presented a solution to this problem with an effective implementation using a robotic agent. The solution can be contrasted with previous attempts for related problems.

Bourgault et al. (2003) employed a Bayesian approach where the target probability density function is used as prior information. The target PDF is updated using the model of the sensor and expected target motion. The optimal search trajectory is defined as the one that maximizes the cumulative detection probability over a limited time period. The key assumptions in their strategy are that the PDF of the target locations is smooth and the search space is free from obstacles constraining the searcher's motion, e.g. for rescue air vehicles. Cowan and Kovesi (1988) explicitly specify configurations of the camera's state parameters in order to perform a certain task. In these methods the effectiveness of the system can largely be determined by the locations, types and configurations of the sensor used. Some redefine the search task as given a set of locations that completely cover the environment the shortest path that visits all of these points. The robot of Fukazawa et al. does just this, as it moves along the path, it looks for objects. Kim et al.



camera angle
for picture of lab
environment
shown in Fig. 1

Fig. 2. The sequence of viewpoints examined during the search. The brown rectangle represents the search robot. Its path is shown with bold arrows. The room layout includes tables, shelves, and other items shown in light gray outline. At each position, small arrows show the viewpoints inspected. The visual field for each viewpoint is shown as a colored region, with the depth of field for which recognition was possible delimiting the extent of the region. Within each region the degree of elevation from the horizontal is given, where 0° is horizontal. The target is marked (middle of lower left quadrant of room).

(1985) has studied the problem of determining camera's viewpoints for successive views looking for distinguishing features of an object. The distance of the camera to the object is determined by the size of the object and the size of the feature. Within this distance, the shape of the feature and presence of occluding objects determine the direction. An aspect graph with nodes being assigned values representing the goodness of the view is suggested to guide the motion of the camera on the sphere.

Our solution performs search for a target object in an unknown 3D environment. No assumptions are made about the configuration of the environment nor the position of the target object. Since our search agent generates its path based on its current knowledge of the target's location encoded by the probability distribution computed during the search process, it minimizes the expected time to find the object, and does not simply determine a path that covers the entire environment. The state of our solution in terms of its experimental verification under different lighting conditions and differently cluttered environments can be seen in Shubina (2007); 'find' rate is 85% assuming that the target is visible from the free floor space available to the robot. Future enhancements include extending the recognition function to 3D objects and adding a full path planning capability so that arbitrary movements can be used.

References

Bajcsy, R., Active perception vs. passive perception. In Proc. IEEE Workshop on Computer Vision: Representation and Control, pages 55–62, 1985.

Bourgault, F., Furukawa, T., Durrant-Whyte, H., Coordinated decentralized search for a lost target in a bayesian world, IEEE/RSJ, International Conference on Intelligent Robots and Systems, pages 48–53, 2003.

Connolly, C. I., The determination of next best views, IEEE International Conference on Robotics and Automation, page 432-435, 1985.

Cowan, G.K., Kovesi, P., Automatic sensor placement from vision task requirements, IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 10, 1988.

Dickinson, S.J., Christensen, H.I., Tsotsos, J.K., Olofsson, G., Active object recognition integrating attention and viewpoint control. Computer Vision and Image Understanding, 67(3):239–260, 1997.

Fukazawa, Y., Trevai, C., Ota, J., Yuasa, H., Arai, T., Asama, H., Controlling a mobile robot that searches for and rearranges objects with unknown locations and shapes, IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 1721– 1726, 2003.

Garvey, T. D., Perceptual strategies for purposive vision. Technical report, SRI International, 1976.

Kim, H.S., Jain, R.C., Volz, R.A., Object recognition using multiple views, IEEE International Conference on Robotics and Automation, page 28-33, 1985.

MacLean, W., Tsotsos, J.K., Fast pattern recognition using gradient-descent search in an image pyramid, International Conference on Pattern Recognition, pages 877–881, 2000.

Maver, J., Bajcsy, R., Occlusions as a guide for planning the next view. In IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 15, pages 417–433, May 1993.

Reece, D. A., Selective perception for robot driving. Technical report, Carnegie Mellon Computer Science, 1992.

Rensink, R.A., A New Proof of the NP-Completeness of Visual Match. Technical Report 89-22 (September 1989), Computer Science Department, University of British Columbia.

Rothenstein, A.L., Tsotsos,J.K., Attention links sensing to recognition, Image and Vision Computing, Available online 29 March, 2006.

Shubina, K., Sensor Planning for 3D Object Search, MSc Thesis, Dept. of Computer Science & Engineering, York University, Toronto, Canada, January 2007.

Tsotsos, J.K., Culhane, S., Wai, W., Lai, Y., Davis, N., Nuflo, F., Modeling visual attention via selective tuning, Artifical Intelligence 78(1-2),p 507 - 547, 1995.

Tsotsos, J.K., On the Relative Complexity of Passive vs Active Visual Search, International Journal of Computer Vision 7-2, p 127 - 141, 1992.

Tsotsos, J.K., Analyzing Vision at the Complexity Level, Behavioral and Brain Sciences 13-3, p423 - 445, 1990.

Tsotsos, J., The Complexity of Perceptual Search Tasks, Proc. International Joint Conference on Artificial Intelligence, Detroit, August, 1989, pp1571 - 1577.

Tsotsos, J.K., A `Complexity Level' Analysis of Vision, Proceedings of International Conference on Computer Vision: Human and Machine Vision Workshop, London, England, June 1987.

Wilkes, D., Tsotsos, J.K., Active object recognition. CVPR’92, pages 136–141, 1992.

Wixson, L., Ballard, D., Using intermediate object to improve efficiency of visual search. International Journal of Computer Vision, 18(3):209–230, 1994.

Wixson, L., Viewpoint selection for visual search, Computer Vision and Pattern Recognition, pages 800–805, 1994.

Ye, Y., Tsotsos, J.K., A Complexity Level Analysis of the Sensor Planning Task for Object Search, Computational Intelligence Vol. 17, No. 4, p. 605 – 620, Nov. 2001.

Ye, Y., Tsotsos, J.K., Sensor Planning for Object Search, Computer Vision and Image Understanding 73-2, p145 - 168, 1999.