Spatial Attention and the Maintenance of Representations of a Robot's Environment

J. J. Clark
Centre for Intelligent Machines
McGill University
Montréal, Québec, Canada H3A 2A7

Abstract

Mobile robots are often faced with complex unstructured environments, which tax their ability to describe and represent. In this paper we describe an approach for handling this complexity by restricting the features of the environment that are represented to those that are relevant to a particular task. We propose an approach, based on the use of saliency driven spatial attention, to filter out irrelevant features, and to detect relevant changes in the robot's environment which can then be used to update the robot's representation of its environment.

1 Introduction

Robots, especially mobile robots, are often faced with extremely complex unstructured environments. This complexity leads to difficulties in achieving real-time peformance for robot vision systems. One way in which this complexity can be managed is to compute, and represent, only those apsects of the scene that are relevant to the activities of the robot (e.g. for obstacle avoidance, or for task performance). As Ballard points out [1] one can avoid the combinatorial explosion of absolute scene representations by using "indexical representations", wherein the system does not maintain an accurate representation of everything in the scene, but should instead only register objects and features that are relevant to the task being carried out.

In using the indexical representation approach the issues of defining relevancy and of updating and maintaining the representations must be considered. Considerations of these issues is the subject of this paper. We propose that these two issues are actually linked, with their common aspect being the role played by *spatial attention*. Spatial attention, in biological vision systems, is a process by which the sensitivity of visual feature detectors tuned to a certain area of the visual field are enhanced relative to other areas [9]. Such attentional processes are thought to reduce the influence of irrelevant visual information on image analysis operations. The spatial location of the attentional enhancement can shift about in the visual field. While the details of how this shifting is controlled in biological systems is as yet a subject of controversy, one commonly held viewpoint is that shifts are made to areas whose features are most *salient*, or most relevant to the visual task being performed [3].

We propose that the relevant scene features, which will be the only aspects of the scene that are stored in the scene representation, are those which are salient, that is, those which attract attention. The scene representation is updated only when a change is detected in previously stored aspects, or when attention is drawn to a new aspect of the scene. If there is no change detected in the scene, there is no reason to update the scene representation. A change in an aspect of a scene can only be reflected in a change in the scene representation if attention is directed to the location of the scene change.

In our approach, the following steps are neccessary to update or maintain the scene representation:

- A change in the scene is detected
- The detected scene change is localized
- The scene change is recognized
- The scene representation is updated

If any of the first three steps fails, the scene representation will not be updated, and the (robot) system will behave as if nothing in the scene has changed.

Humans appear to use this approach, as demonstrated by the experiments of Rensink *et al* [6]. Humans that view a scene ignore aspects of the scene that they do not pay attention to, and do not notice changes in these aspects as long as those changes do not draw their attention. The human perception of a rich external world is partly an illusion, created by

the ability of the eye to quickly move to, and acquire information about the scene if it is ever needed. In effect, one is using the world as an external memory buffer, which can be rapidly accessed by moving one's eye to the appropriate location [1, 5]. Mobile robots can conceivably benefit from this approach, as it would aid in managing the complexity of representing their environments. The cost of this, of course, is that the robots could be fooled on occasion, in the same way that humans are, as they will not notice every change in their environment. But they will, hopefully, notice changes that are important to them.

There is an increasingly large body of evidence that suggests that spatial attention in biological systems is closely linked to the planning and execution of movements (see the discussion in [2] for more details). This supposed link is the basis for a recent theory of attention, the *pre-motor theory of attention*, which holds that all spatial attention processes arise from, or are used in the generation of, movements or planned movements of all types [7].

The pre-motor theory is of particular interest to those involved in developing visual-motor systems for mobile robotics, as it provides the groundwork for integrating sensing and motor activities, through the reciprocal influence of motor commands and spatial attention. As we will see later in this paper, the approach that we take to the control of camera movements is based on using shifts in spatial attention to determine the targets of these motions. These motions will, in turn, be used to provide additional visual information to be used in updating the scene representation.

2 Maintenance of Scene Representations

To build and maintain a scene representation the robot visual system must detect changes in the scene, localize these changes to direct its visual processing resources, and then recognize the nature of the change. Based on this perception of change, the robot can either add a new entry to its representation (if the scene change is due to the appearance of a new object in the scene), delete an existing entry (if the scene change is due to the disappearance of a previously represented object), or modify an existing entry (if the scene change is due to some change in the characteristics of a previously represented object). These steps are discussed in the following sections.

2.1 Change Detection

The first step in our approach to the maintenance of scene representations is to detect that a change has occured in the scene.

This detection of change can occur in one of two manners. First, change can be detected by comparing the representation of a scene object currently being viewed with the set of stored scene objects. It may be the case that the current object is best interpreted as a change in one of the stored scene objects, in which case we could signal that a change has occured. For example, the location of the object in the scene may have shifted or it may have changed shape. Physical constraints such as object rigidity or velocity limits imply that these types of changes cannot be too extreme, otherwise the situation would be better interpreted as the stored object disappearing and a new object appearing.

The second way in which change can be detected is via the use of an image based "motion" feature. In this approach, a temporal derivative operator is applied to the image features and used as a measure of scene change. If this "motion" feature exceeds a given threshold then the system can signal that a scene change has been detected.

2.2 Change Localization

One a scene change has been detected, the robot's visual system must determine where in the scene the change is located.

If the change had been detected by comparing the currently attended object in the scene to the stored object list, then nothing further needs to be done, as the change location is at the current location of the attentional focus. If the change was detected due to an image motion cue, then the location of the change can be obtained by determining the location of the motion cue.

In many situations, however, multiple locations in the scene will be changing. Also, in some situations the change detection process is constantly being triggered. This is usually the case in an outdoor scene, where there can be leaves waving in the wind, cars going past, people walking, clouds scudding overhead, birds winging their way, and so forth. All of these will excite the change detection process. Thus an image transient based change detection scheme is of little use in a general setting. Our approach is that (potential) change is localized when spatial attention shifts to that location. If this shift in attention is caused by increased saliency at that location due to an image transient, then we could say that this transient has been localized. If there are many image transients at a given time, these all vie for the locus of attention, but only one will win out. The change at this winning location will be then localized. None of the other changes that are occuring at the same time will be

localized.

Sometimes the image may be globally disrupted, such as by an illumination change (e.g. by strobe light or a camera shutter), or by a rapid camera motion. In this case, there will be no clear location to draw the attention to, and there will be no shift in attention, and no direct change localization will take place. In such cases of de-localized change cues, localization of an actual object change must proceed via a serial search process. This search process relies on the fact that shifts in attention are possible even when no image transients are present, either by temporal inhibition of the currently attended to location (inhibition or return), by changing of saliency weighting factors, or by internally generated shifts from a gaze planning system. In this way, different locations in the scene can be examined and checks for scene changes at these locations can be performed. Thus if a change occurs in a part of the scene, but the resulting image transient does not attract attention to that location (perhaps due to de-localization of the transient, or to another location grabbing attention first) detection of the change at that location is only possible if attention is somehow allocated to that location at a future time, and only if the change persists until the time.

If attention is never directed to an image location, changes at that location will never be recognized. The fact that attention is never directed to that location, however, implies that that location is not of interest to the visual system and that it does not matter that changes at that location can not be detected. In this way, attention serves to reduce the distracting effect of irrelevant stimuli.

2.3 Change Recognition

The nature of the scene change is determined by a comparison of features of the attended object to those previously stored in the scene representation. The details of the change recognition process depend on the precise form of the structures used to represent scene objects, and on the visual features used to generate the object descriptions. We will give an illustrative example later in the paper.

3 Implementation in a Robotic Vision System

In this section we describe a robotic vision system which demonstrates the approach described earlier for the maintenance of scene representations.

We use a very simplistic world consisting of highly saturated coloured objects lying on a white background plane. These objects were characterized by their position in the world relative to the camera, by their hue, by their size, and by their shape.

3.1 Image Acquisition and Gaze Control

We use a fixed platform on which is mounted a Panasonic WVCP-410 solid-state colour video camera affixed to a Directed Perceptions pan-tilt unit. The pan-tilt unit was controlled via a serial link to a Pentium based PC. The PC contained a Matrox Meteor video digitizing card which was used to acquire the images from the colour video camera. The 640x480 pixel image acquired by the digitizer was compressed in software to a 5600 pixel foveal image, whose pixel structure is shown in figure 1. The use of a fixed, rather than mobile, platform for the camera motion control mechanism eliminates factors that are not immediately relevant to our demonstration. Our approach can be extended to handle moving camera platforms, however. In what follows, therefore, the body-centred and "head-centred" spatial maps are equivalent and we will just refer to head-centred maps.

3.2 Spatial Maps

In our demonstration system, we need to worry about two types of spatial maps, those relative to an image-centred (or camera centred) coordinate system and those relative to a body-centred coordinate system.

Camera-centred maps are taken to be *foveal* in our system. The geometry of a foveal map is depicted in figure 1a. The foveal map has a high resolution in the centre of the map, and has a decreasing resolution as one moves radially outwards from the centre. There are a number of advantanges to using a foveal representation, the most important of which is perhaps the great reduction in the amount of data that needs to be processed [8]. This reduction is of special importance in mobile robotics applications, where real-time processing is desired.

We also adopt a specific geometry for the head-centred maps. We call this particular map a "Horizon" map, which is depicted in figure 1b. The horizon map has high resolution along the central horizontal axis of the head coordinate system and decreases in resolution as one goes away from this axis in the vertical direction. We are not claiming that this type of map is of any particular utility in practice, but are using it as an example of mapping between two different types of spatially non-uniform image representations. It can be thought of a spatial map which could be used in a navigation task in a mobile robot, where obstacles are likely to lie on a horizon plane, or in a manipulation task, where the objects to be manipulated are likely to lie on an elevated horizontal plane in front of the robot. This is an example of what Rizzolatti call a "pragmatic map", a spatiotopic map which is used for

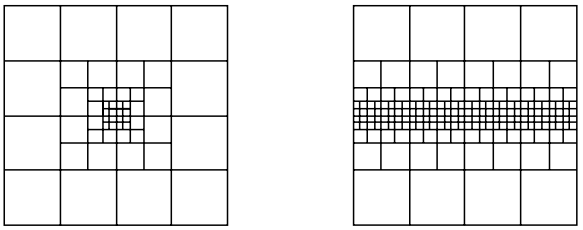

Figure 1: a) The foveal image map. Each square in the diagram represents an array of 10x10 map pixels. There are a total of 5200 pixels in the map. b) The horizon map. Each square in the diagram represents an array of 20x20 map pixels. There are a total of 73600 pixels in the map.

carrying out some action. In the pre-motor theory of attention, attention control signals are generated by, or are identical to, the activity in pragmatic maps.

To transform from a camera-centred map to a head-centred map, we do the following. For each element (pixel) in the horizon map, we compute all elements (pixels) in the foveal map whose receptive fields spatially overlap the horizon map element. The camera position needs to be known for this step, as this determines the offset of the camera-centred coordinate system relative to the head-centred coordinate system. The average of the (feature) values of these foveal elements is then assigned to the horizon map element. The mapping from the horizon map to the foveal map works in a similar fashion, that is, the value of a foveal map element is assigned the average of the values associated with all horizon map elements whose receptive fields spatially overlap the foveal map element.

3.3 Implementation of Spatial Attention

The attentional process that we use in our algorithm is based on the model of human spatial attention and saccade generation proposed by Clark [2], which is itself based on the Koch and Ullman Winner-Take-All model [3]. In this approach, a set of retinotopically mapped image features are computed from the image. These feature maps are weighted and summed to provide a camera centred (retinotopic) *saliency* map. The saliency map is then converted to a head-centred coordinate system, and is used to drive a winner-take-all dynamic neural network (see [2, 3] for details). This winner-take-all process selects a single region of the head-centred salience map. The centroid of this region is used to determine the target for a saccadic camera movement, if neccessary.

3.4 Saliency Computation

In our experiment, the saliency map which drives the attentional dynamics combines three different image based features. These three features are, the magnitude of the temporal derivative of the image intensity, the proximity of the object color to a target color, and the saturation of the color. The target, or most salient, color was shifted periodically, by 90 degrees in Red-Green/Red-Blue color opponent space, whenever more than 10 image frame times had passed without a camera movement. This causes a cycling of the focus of attention which produced "scan paths" similar to those investigated by Noton and Stark [4].

3.5 Structure of the Scene Representations

In our demonstration system, the scene representation consists of a linked list of object description structures. These structures contain the following pieces of information about the object: The object area (in head coordinate pixels), the object centroid (in head coordinates), and the object color (expressed as an angle in Red-Green/Red-Blue opponent space, and taken from the point of maximum saliency).

Obviously, in a more practical application the structures used to represent different objects would be much more complex. The above structure is sufficient to demonstrate the approach that we are proposing, however.

3.6 Change Detection, Localization and Recognition

In our experiment change detection was carried out in two ways. The first involved comparison of a currently attended-to object with objects stored in the scene representation. This comparison only takes place if the centroid of the currently attended-to object is close (within a radius of 10 pixels) to one of the stored objects. Otherwise it is considered to be a new object. If the currently attended-to object is close in spatial location to a previously stored object then the change recognition process is carried out.

The second way in which scene changes are detected is via a visual motion cue. In our test system the motion cue is obtained by simply taking the magnitude of the first temporal derivative of the image intensities. As we weight the contribution of the motion saliency much more strongly than the color saliency, this high motion salience will, in most cases, cause a shift in the attention to that location. This shift in attention will signal that a change has been detected, and the change is automatically localized by the shift in attention.

After every shift in attention, the change recognition process is carried out. This involves the object comparison process described earlier. If there is no previously stored object near the newly attended-to location, then the presence of a new object is signalled and its representation is added to the scene representation. If there is a previously stored object near the object under consideration then the features of the two objects are compared. If they are the same (within a tolerance) then it is concluded that no scene change has occured and no updating of the scene representation is done. If there is a small difference in one or more of the object features then a change in the object is signalled and the representation is updated to reflect that change. If there is a large difference in one or more of the object features then a large change is signalled and the previous object entry in the scene representation is deleted and a new one is added corresponding to the object currently being attended to.

The shift in attention may, or may not, be followed by a camera movement to the new locus of attention. To determine whether a camera motion is neccessary, a coarse comparison process is first carried out. If this coarse comparison suggests that a change has occured then the camera motion corresponding to the new attentional focus is executed. Otherwise no motion is made, as the system has decided that nothing has changed in the scene. Note that, because of our application of the pre-motor theory of attention, the spatial map used for targetting and control of the camera motion is the same as that used in the generation of the attentional modulation signals. Whenever the system generated a camera motion, the motion feature computation was suppressed for 3 frame times, to prevent the image change that results from the camera movement from creating a spurious change signal. This is similar to the suppression in visual sensitivity observed in mammalian visual systems when saccadic eye movements are performed.

When an attention shift occurs, we change the target color to be the average color of the newly attended to region. This is required to handle shifts in attention caused by motion cues. In general, the object causing the motion cue will have a different color than the current target color. If the target color is not changed to match this object's color then attention would shift away from the object after the motion cue died away.

3.7 Experiment

A typical run of our system is depicted in figures 2 through 4. Shown are the sequences of images obtained as the system is acquiring it's initial scene representation. In the left part of each figure is shown the scene that the camera is viewing. Note that the intensity images have been converted to the foveal camera-centred representation of figure 1a). In the scene the triangular object to the left has a green colour, the square object on the top has a red colour, the triangular object to the right has a blue colour, and the rectangular object near the bottom has a yellow colour. The right hand image of each figure represents the corresponding head centred saliency map.

The following is a transcript of the messages produced by the system:

```
New blue (80,67) A=2080.
New red (-57,4) A=4421.
New green (-104,87) A=1072.
Red at (-57,4) shift to (-1,30).
```

The last message is in response to a motion of the red object to the right and downwards. These are the only messages generated, and thus correspond to the only modifications made to the scene representation. Note that the yellow object was not found, as yellow was never a target colour. Moving the yellow object would cause it to be detected, however, as the motion cue would attract attention to it.

4 Summary and Conclusions

In this paper we have proposed an approach to the construction and maintenance of representations of a mobile robot's environment. It is based on the principle that only those objects in the environment that are relevant to the task at head need be represented. Relevance is determined by an attentional mechanism which is driven by the saliency of object features. Maintenance of the representation is performed by detecting changes in the relevant aspects of the scene. If no such change is detected then the representation is not altered. Relevant scene changes are only detected if attention is directed to the change location, as it is this shift in attention which defines relevancy. Unlike other approaches to change detection, image transients serve only to direct attention to locations in the scene, and are not used directly to detect scene changes. Once attention has been allocated to a scene location, change recognition processes are carried out on the attentionally modulated image at that location. Scene changes which result in image transients that do not attract attention will not be detected, localized, or recognized, unless attention is directed to the scene location by some other non-transient means in the future.

Our implementation of spatial attention is based on the pre-motor theory of attention. As such, the spatial maps representing attentional activity are the

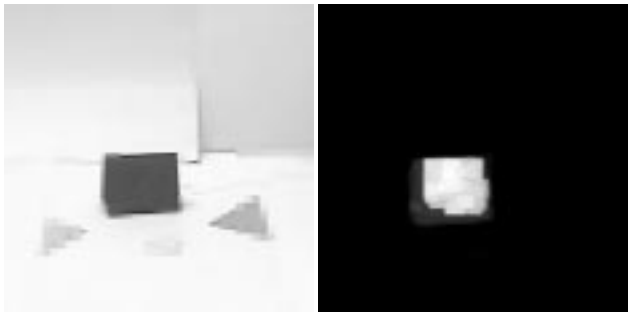

Figure 2: Left: The foveal image of the scene, just after a camera movement based on a target colour of red. Right: The corresponding head centred saliency map.

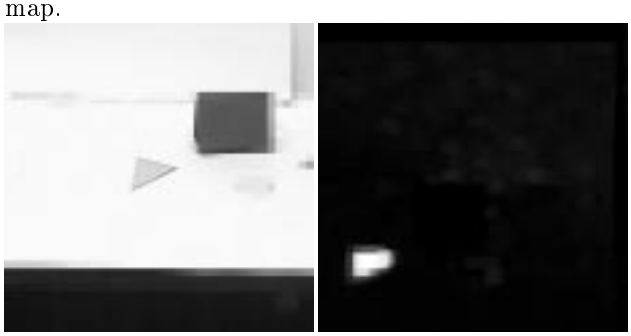

Figure 3: Left: The foveal image of the scene, just after a camera movement based on a target colour of green. Right: The corresponding head centred saliency map.

Figure 4: Left: The foveal image of the scene, just after a camera movement based on a target colour of blue. Right: The corresponding head centred saliency map.

same as those used in planning and executing motor actions. This means that once a change has been detected, motor activities appropriate to that change can be quickly carried out. In our systems this motor activity was a shift in the camera gaze to view the change location, but in more complex systems these motor activities could be reaching for an object with a manipulator, or avoiding an obstacle.

Acknowledgements

This research was supported by NSERC grant number 229-66.

References

[1] Ballard, D.H., (1990) "Animate Vision", University of Rochester, Department of Computer Science, Technical Report No. 329

[2] Clark, J.J. (1997), "Spatial attention and latencies of saccadic eye movements", accepted for publication in *Vision Research*

[3] Koch, C. and Ullman, S. (1985), "Shifts in selective visual attention: Towards the underlying neural circuitry", *Human Neurobiology*, Vol. 4, pp 219-227

[4] Noton, D. and Stark, L. (1971), "Scanpaths in saccadic eye movements while veiwing and recognizing patterns", *Vision Research*, Vol. 11, p 929.

[5] O'Regan, J.K. (1992), "Solving the 'real' mysteries of visual perception: The world as an outside memory", *Canadian Journal of Psychology*, Vol. 46, pp 461-488

[6] Rensink, R.A., O'Regan, J.K., and Clark, J.J. (1997), "To see or not to see: The need for attention to perceive changes in scenes", *Psychological Science*, Vol. 8, No. 5, pp 368-373

[7] Rizzolati, G., Riggio, L., and Sheliga, B.M., (1994), "Space and selective attention", in **Attention and Performance XV**, Umilta, C. and Moscovitch, M., eds., MIT Press, Cambridge, MA, pp 231-265

[8] Swain, M. and Stricker, M. (eds.) (1991), "Promising Directions in Active Vision", University of Chicago Technical Report, CS 91-27.

[9] Posner, M.I., Cohen, Y., and Rafal, R.D. (1982), "Neural systems control of spatial orienting", *Philosophical Transactions of the Royal Society, London B*, Vol. 298, pp 187-198