Overt visual attention inside JDE control architecture *

José M. Cañas[1], Marta Martínez de la Casa[1], Pablo Bustos[2] and Pilar Bachiller[2]

[1] Universidad Rey Juan Carlos, 28933 Móstoles, Spain
[2] Universidad de Extremadura, 10071 Cáceres, Spain

Abstract. In this paper a visual overt attention mechanism is presented. It builds and keeps updated a scene representation of all the relevant objects around a robot, specially if they are far away of each other and do not lie in the same camera image. The algorithm chooses the next fixation point for a monocular camera, which is mounted over a pantilt unit. Our approach is based on two related dynamics: liveliness and saliency. The liveliness of each relevant object diminishes in time but increases with new observations of such object. The position of each valid object is a possible fixation point for the camera. The saliency of each fixation point increases in time but is reset after the camera visit such location. Real experiments with a pioneer robot endowed with a firewire camera on a pantilt unit are displayed.

1 Introduction

The use of cameras in robots is continuously growing. Cameras have become a cheap sensor in last years and they can potentially provide the robot with much information about its environment. But processing the huge amount of data from the cameras is not easy. Attention mechanisms of human vision system has been source of inspiration for machine visual systems, in order to sample data nonuniformly and to utilize computational resources efficiently [13].

On one hand, it offers a solution for the processing bottleneck generated by the huge amount of data carried by video streams. The *covert attention mechanism* [5, 8, 12] focuses in the areas of the images relevant for the task at hand, leaving out the rest. Biological models are moving in last years to the real-time arena and offer an impressive flexibility to deal simultaneously with generic stimulus and with task specific constraints [7, 10].

On the other hand, visual representation of interesting objects in robot's surroundings may improve the quality of robot behavior as its control decisions may take more information into account. This poses a problem when such objects do not lie into the cameras field of view. Some works use omnidirectional vision

* This work has been funded by Spanish Ministerio de Ciencia y Tecnología, under the projects DPI2004-07993-C03-01 and DPI2001-0469-C03-03, and by the Spanish Junta de Extremadura under the project 2PR01A031

, other approach uses a regular camera and an *overt attention mechanism* [8, 4], which allows for rapid sampling of a very wide area of interest. The use of camera motion to facilitate object recognition was pointed out by [13], and has been used, for instance, to discriminate between two shapes in the images [12].

In this paper we report on an overt attention system for a mobile robot endowed with a pan-tilt camera. This system performs an early segmentation on color space to select a set of candidate objects. Each object enters a coupled dynamics of liveliness and saliency that drives the behavior of the system over time. The system will continuously answer to two questions: how many relevant colored objects are there around a robot? and, where are they located?.

2 Overt visual attention mechanism

The task of the overt attention mechanism is to set the target coordinates for the pantilt unit at every time in order to keep fresh and updated the scene representation. Such representation is the collection of relevant objects, their positions and visual size. We considered only the pink objects were relevant.

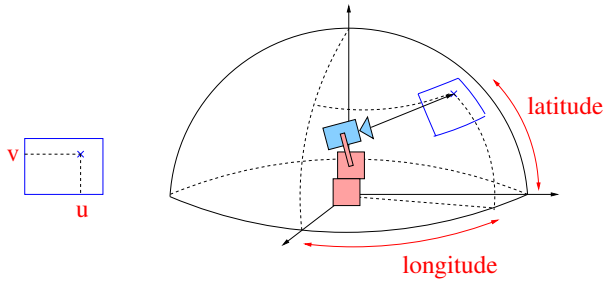

Fig. 1. Scene space and coordinates

The color images are cartesian ones, with two coordinates (u, v) per pixel. The pantilt unit is located aiming at $(pan, tilt)$ angles. We define a *scene space*, consisting of a sphere around the pantilt unit, accounting for all the possible camera orientations. Each pixel of the cartesian image projects into a scene pixel $(latitude, longitude)$ of the scene space (shown in figure 1), depending of its own position (u, v) and the current pantilt position. Each monocular image projects into a patch in such scene space, consisting of all the projected pixels. The projection equations incorporate the kinematics of the pantilt body and the pinhole model for the camera.

The attention mechanism designed follows the algorithm in figure 2. The pantilt is constantly moving from one fixation point to the next. No image is processed while the pantilt is moving, but once it has stopped at current fixation

```
initial sweep
loop
    move the pantilt to the next fixation point
    color filter of monocular image
    clustering of monocular objects
    project monocular image into the scene space
    matching with scene objects
    update liveliness
    update saliency
    choose the most salient fixation point
end_loop
```

Fig. 2. Pseudocode of our overt attention algorithm

point the monocular image is processed to update the scene representation, the next target point for the pantilt is computed and commanded.

The system starts with a complete sweep along the whole scene, fixating the pantilt unit at regular intervals horizontally and vertically. At every fixation point the monocular image is filtered searching for pink pixels, which are clustered together in pink objects, and projected into scene space. The color filtering is performed in HSV space, which is more robust to changes in illumination than RGB. A fast histogramic clustering algorithm is used [1]. This way the starting sweep builds the initial scene image, for instance that in figure 3 (left). For visualization the scene image is projected into the display using a polar transformation, where $\rho = latitude$ and $\theta = longitude$.

2.1 Liveliness dynamics

The attention mechanism is based on two related and concurrent dynamics: liveliness of objects and saliency of fixation points. Each object in the scene has a *liveliness*, meaning the confidence of such internal symbol being a proper representation of the real object. In general the objects will lose liveliness in time, but will gain it every time they are observed with the camera. The equations (1) and (2) describe the dynamics of the liveliness in the discrete time. Equation (1) is applied at each iteration. To avoid infinite values of liveliness, we introduced saturation: its values are bounded inside the [0,MAX_LIV] interval.

$$liv(object, t) = liv(object, t-1) - \Delta L_{time} \tag{1}$$

$$liv(object, t) = liv(object, t-1) + \Delta L_{observation} \tag{2}$$

There is a threshold, a minimum liveliness required for an object to be considered valid. Objects with liveliness below such threshold are simply discarded. This allows the system to forget objects that disappear from the scene or those not recently observed. In addition, in order to graphically show the effect of forgetting, pixels in the displayed scene gradually fade to white values (right side

of figure 3). So, areas of the scene which are not observed for a long time are displayed in white, while areas of the scene that have been recently visited show the fresh projected monocular patch.

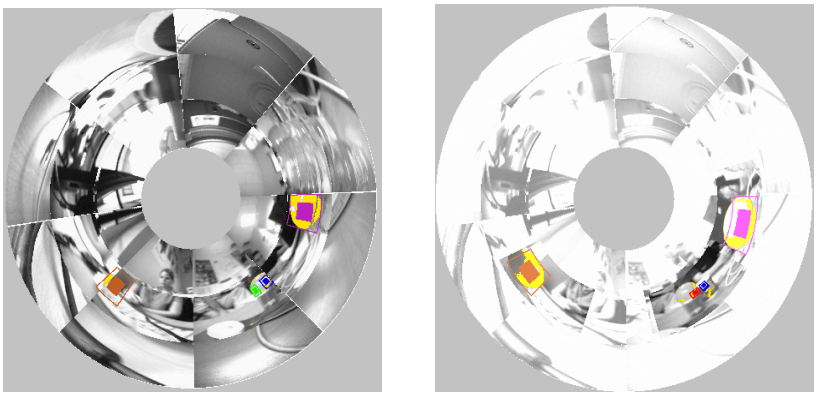

Fig. 3. Scene image after the initial sweep and the forgetting mechanism

Once the pantilt has stopped at the current fixation point, the monocular image is color filtered and its pink blobs clustered as local objects. Each local object is projected into the scene and matched against the current scene objects using a distance threshold. If the local blob projects close enough to an existing scene object then it is considered to be the new position of such object, which has moved a little bit and is properly tracked. In case of positive match the liveliness of such scene object will be increased following equation (2). Local objects without correspondence in the scene image are inserted as new scene objects, with a liveliness roughly above the validity threshold.

2.2 Saliency dynamics

The second dynamics of the attention mechanism is the saliency. The *fixation points* are a collection on possible target positions for the pantilt unit. In our mechanism, the center of each valid scene object is inserted as a fixation point. Each fixation point has a *saliency*, meaning how desirable such position is as target for the pantilt unit. In general the fixation points will increase saliency in time, but will reset it every time the camera is fixated at them. The equations (3) and (4) describe the dynamics of the saliency in the discrete time.

$$sal(fixp, t) = sal(fixp, t-1) + \Delta S_{time} \tag{3}$$

$$sal(fixp, t) = 0 \tag{4}$$

There is a winner-takes-all competition to gain the control of the pantilt motors. The fixation point with highest saliency is chosen as the next target for the pantilt movement.

To avoid being redirected immediately to a previously attended location the saliency of a given fixation point is reset each time the pantilt unit is fixated at it. This way inhibition of return (IOR) is achieved without adding a transient local inhibition activation [3], neither explicitely keeping a saccade history [4]. Our simple saliency dynamics keeps the pantilt away from recently visited locations. In the case of a single object the saliency is reset, but as long as it is the only fixation point, it will gain the pantilt control over and over again.

Because this IOR and that centers of valid scene objects are fixation points, the pantilt tend to jump among them, letting the objects to be periodically observed and to keep (or increase) their liveliness. Such behavior is flexible: new objects can be dynamically added or deleted to the lively objects list and they enter into or get out of the pantilt time-sharing. When a valid object disappears from scene, its liveliness will keep above the threshold for a while, the pantilt will insist on visiting its last location and on giving it a chance to be recovered again. After a while, its liveliness will fall below the threshold and it will be removed from the valid objects list, and so from the list of fixation points.

3 JDE perception and control architecture

The previous attention mechanism was designed inside a perception and control architecture named JDE [9]. In JDE stance, behavior is considered the close combination of perception and actuation in the same platform. Both are splitted in small units called schemas, which are organized in dynamic hierarchies. *Perceptive schemas* build some piece of information concerning the environment or the robot itself (stimulus), and *actuation schemas* make control decisions considering such information like speeding up the motors or moving some joint.

In JDE the perception is a dynamic collection of anchored stimuli [11], more than a complete reconstruction of 3D environment. The selective activation of some perceptive schemas and not others provide a general attention framework.

The described attention mechanism is inserted in JDE as a perceptive schema, which is in charge of build and anchor a scene representation. It consists of a set of pink objects in robot surroundings, where *objects* are defined as blobs of uniform color. As the single camera does not cover all the space around the robot, not all the relevant objects in the scene lies in the monocular camera. This perceptive schema involves *active perception*[14], as it has to move the pantilt unit in order to search for such objects and keep their internal representation fresh. The regular "perceive to actuate" is reversed here to "actuate to perceive".

The objects in the robot surroundings lead the camera movements, so the attention mechanism is bottom-up. The only top-down and task specific information provided is that relevant objects are the pink ones. This inclination towards pink objects is similar to the predisposition found by ethologysts in animals towards certain stimuli in different contexts [6].

4 Experiments

Some experiments[3] have been conducted on a real robot, a Pioneer endowed with a Directed Perception pantilt unit and Videre firewire camera (figure 4). The pantilt unit accepts position commands through the serial port.

Starting with a single object, the system is able to keep it tracked and to follow its (slow) movements around the environment (Figure 4). Such behavior could have been solved with classical tracking techniques, but here we have solved it using exactly the same dynamics that will generate tracking behavior for two, three... objects and the time sharing of the pantilt unit among them.

Fig. 4. The camera follows the objects when they move

For two objects the system reached a stable jumping loop. The saliency evolution for a scene with two objects can be seen at figure 5 (left). The pattern shows a perfect alternating sharing of the pantilt device. Figure 5 (center) shows the liveliness evolution for such experiment, both objects were kept at high values as they are continuously observed. Figure 5 (right) displays how the liveliness of one of the pink balls lowed down when such ball disappeared from scene.

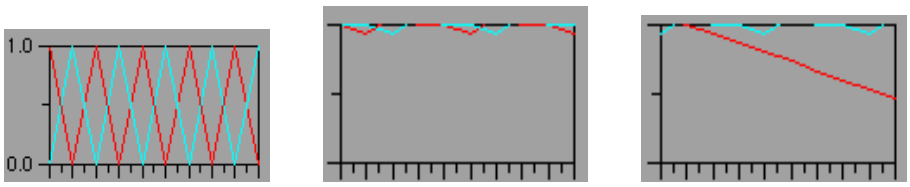

Fig. 5. Saliency (left), liveliness (center) evolution with two pink objects in the scene. In (right) one of them disappears from scene.

Figure 6 (left) shows an experiment with three pink balls around the robot. Figure 6 (right) displays the corresponding scene image. The pantilt continuously

[3] some videos are available at http://gsyc.escet.urjc.es/jmplaza/research-vision.html

oscillated among the three pink balls, in a stable loop, jumping from one tracked object to another, in a round robin sequence. Other areas are not visited by the pantilt and then they fade to white values.

Fig. 6. The robot has three pink balls around and pantilt oscillates among them

Our dynamics were able also to properly track several moving objects. We placed the robot in front of three relevant objects. At the initial position (figure 7 (left)) the perceived scene was that at the figure 7 (center). Then the robot was slowly pushed forward 70cm, approaching to the objects, and the objects spread out in the scene image, as shown in figure 7 (right).

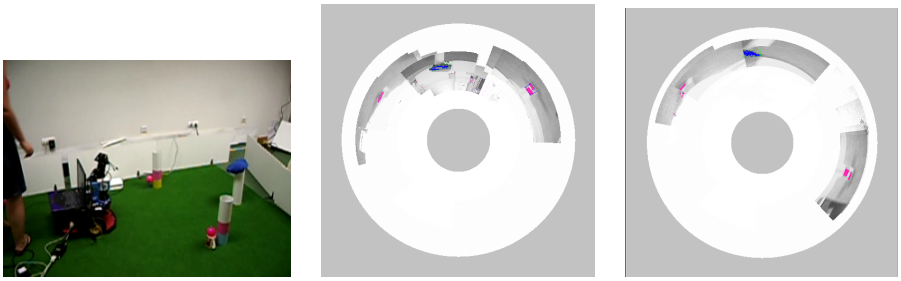

Fig. 7. When the robot approaches to the object, they spread out in the image scene

We found a limit in the number of objects the system can simultaneously track. In the case of few objects, they receive attention frequently enough to keep their liveliness at high values. When the number of objects increases, they tend to receive the attention of the camera at longer intervals, and their average liveliness decreases. There is a number of objects over which the camera movement is not fast and frequent enough to keep the liveliness of all of them above the liveliness threshold, and some of them are forgotten by accident. The particular limit depends on ΔL_{time} and $\Delta L_{observation}$ ratio.

5 Conclusions

A novel overt attention mechanism has been presented. The system finds how many pink colored objects there are around a robot and where they are the located, despited the limited field of view of its monocular camera. Our mechanism deals with valid objects and fixation points in the scene. It is based on two related dynamics: liveliness and saliency. The position of valid objects are the allowed fixation points for the pantilt device. The target of pantilt movements is always the most salient fixation point among the list.

Such simple dynamics generate several interesting behaviors, as shown in the experiments with a real robot. First, it alternates the focus of attention among the relevant objects of the scene, regardless their number, one, two, three... A limit in such number was also pointed out. Second, the system forgets positions of objects that disappear from the scene, with a certain tolerance to overcome spurious misses. Third, the system successfully tracks the relative movements of the objects, updating their position in the scene as they moved around.

References

1. Gómez, V., Cañas, J.M., San Martín, F., Matellán, V.: Vision based schemas for an autonomous robotic soccer player. Proc. of IV Workshop de Agentes Físicos WAF-2003 (2003) 109–120
2. Breazeal, C., Scassellati, B.: A context-dependent attention system for a social robot. Proc. of Int.J.Conf. on Artificial Intelligence, (1999) 1146-1151
3. Orabona, F., Metta, G., Sandini, G.: Object-based visual attention: a model for a behaving robot. Proc. of Int. Workshop on Attention and Performance in Computational Vision WAPCV-2005 (to appear)
4. Zaharescu, A., Rothenstein, A.L., Tsotsos, J.K.: Towards a biologically plausible active visual search model. Proc. of Int. Workshop on Attention and Performance in Computational Vision WAPCV-2004, Springer LNCS 3368 (2005) 133-147
5. Tsotsos, J.K., et.al.: Modeling visual attention via selective tuning. Artifical Intelligence 78, (1995) 507-545
6. Tinbergen, N.: The study of instinct. Clarendon University Press, Oxford UK, 1951
7. Itti, L.: Models of Botton-Up and Top-Down Visual Attention. PhD Dissertation, California Institute of Technology, (2000)
8. Itti, L., Koch, C.; Computational Modelling of Visual Attention. Nature Reviews Neuroscience 2, (2001) 194-203
9. Cañas, J.M.: Jerarquía dinmica de esquemas para la generación de comportamiento autónomo. PhD Dissertation, Universidad Politécnica de Madrid, (2003)
10. Navalpakkam, V., Itti, L.: Modelling the influence of task on attention. Vision Research 45, (2005) 205-231
11. Pylyshyn, Z.: Visual Indexes, preconceptual object and situated vision. Cognition 80, (2001) 127-158
12. Marocco, D., Floreano, D.: Active vision and feature selection in evolutionary behavioral systems. Proc. of Int. Conf. on Simulation of Adaptive Behavior (SAB-7), (2002) 247-255
13. Ballard, D.H.: Animate vision. Artificial Intelligence 48, (1991) 57-86
14. Bajcsy, R.: Active Perception. Proc. of the IEEE 76, (1988) 996-1005