An Experiment in Distributed Visual Attention [1]

P. Bachiller[1], P. Bustos[1], J. M. Cañas[2] and R. Royo[1]

[1] University of Extremadura, 10071, Cáceres, Spain

[2] University Rey Juan Carlos, 28933, Móstoles, Spain

Abstract. Attention mechanisms of biological vision have been applied to machine vision for several applications, like visual search and object detection. Most of the proposed models are centred on a unique way of attention, mainly stimulus-driven or bottom-up attention. We propose a visual attention system that integrates several attentional behaviours. To get a real-time implementation, we have designed a distributed software architecture that exhibits an efficient and flexible structure. We describe some implementation details and real experiments performed in a mobile robot endowed with a stereo vision head.

1 Introduction

The visual attention system in a mobile robot acts as a dynamical device that interacts with the environment to select what might be relevant to current active tasks. At the same time, it should maintain responsiveness to unforeseen events. More specifically, it should enclose the following functions [12]: selection of regions of interest in the visual field; selection of feature dimensions and values of interest; control of information flowing through the visual system; and shifting from one selected region to the next in time or the “where to look next” task.

Attention can be classified according to various aspects. In psychology, the terms generally used are active (voluntary) and passive (involuntary) attention [3]. From a stimulus point of view, there is overt or covert attention depending on the way the stimulus is attended [10]. Overt attention is the act of directing our eyes towards a stimulus source. Covert attention is the act of mentally focusing on a particular stimulus without any motor action. Attending to the mechanism that drives attentional control, there are two kinds of execution methods: one is bottom-up or stimulus-driven, which shifts attention to regions with visual features of potential importance; another is top-down or goal-directed, which uses knowledge of the visual features of the desired target to bias the search process.

In recent years, visual attention has taken an important place in robotics research. Most of the proposed models have focused on pure bottom-up [2][12] and some on

[1] This work was supported by the Department of Science and Technology of the Extremadura Government (grant 3PR05A044), by the Madrid Government (grant S-0505/DPI/0176) and, also, by the Spanish Ministry of Education and Science (grant DPI2004-07993-C03-01).

Bachiller P., Bustos P., M. Cañas J. and Royo R. (2007).
An Experiment in Distributed Visual Attention.
In *Proceedings of the 3rd International Workshop on Multi-Agent Robotic Systems*, pages 106-112

top-down [9] attention. There have been some efforts on combining both forms of attention by weighting bottom-up saliency maps with top-down information [8,11].

The approach proposed in this paper is mainly characterized by the integration of different attention categories at a single system endowed with a flexible and adaptable architecture. The proposed system is modelled as a collection of processes collaborating to fix a visual target and to choose the next one. The complexity of the resulting global system requires the use of distributed software engineering techniques. We use the Internet Communication Engine (Ice) middleware platform [1] and a custom component model specifically tailored to build distributed vision architectures.

2 Architecture of the Proposed System

The visual attention system proposed in this paper integrates several ways of attention to work successfully at different situations. It has been designed and tested on a mobile robot with a stereo vision head. The net of processes that compose the system are a set of Ice components collaborating to fix a visual target. As shown in figure 1, the elements in the architecture are roughly organized in two branches that converge in the lower part of the graph. From this point a closed-loop connection feeds back to the upper initial part. The two branches divide the visual function in a manner analogous to the “what” and “where” pathways proposed in neuroscience [7]. This division allows for a specialization of functions, dedicating specific resources to each branch and sharing what is common from lower-level processes. The “what” branch tries to find and track a specific target in the image using bottom-up computed regions of interest (ROI) and top-down specification of targets. The “where” branch extracts geometric information from stabilized ROI's and selects those regions that meet certain requirements, such as being on the floor plane, being too close or being in the current heading direction. The information from both branches has to be integrated solving an action selection problem. Given a current task or set of tasks, where to look next? This is accomplished by the lower component in the graph which outputs commands to the underlying motor system. In our system “looking to something” implies a 3D positioning of the robot with respect to the target, which we call a 3D saccadic movement. Following this reasoning, solving a generic navigation task such as going some-where following a predetermined set of (remembered) landmarks, amounts to generating the “correct” set of saccadic movements that will approach the current target while avoiding potential obstacles. This set cannot be computed a priori as long as it is the result of extended dynamical interactions between the robot and its environment. It is partially defined in the programmed code and partially selected from finding out how the outside world is. The whole system works as a complex mechanical device that is attracted towards some features and rejected from others. The specific interleaving between approaching and avoiding is given by an implicit time relation that links internal parameters and external geometry.

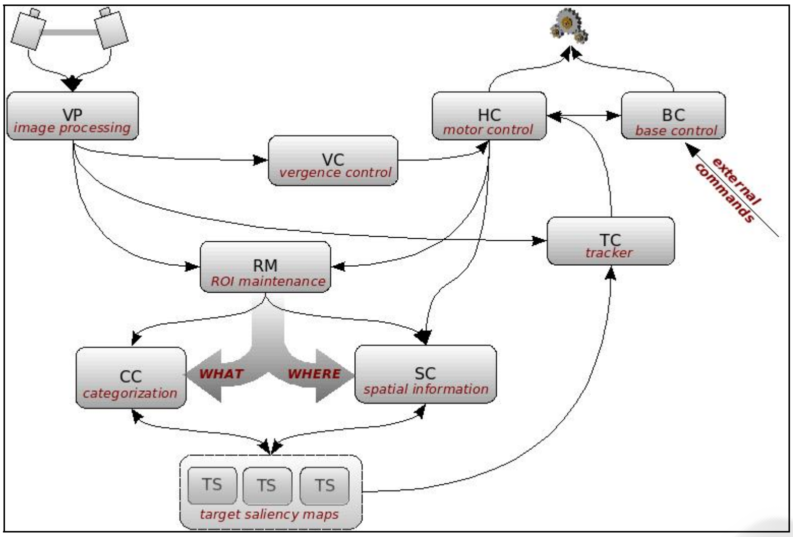

Fig. 1. Architecture of the visual attention system.

3 Components and Connections

Each component in the architecture is a C++ coded UNIX process using the Ice middleware. Similar to CORBA, each component supplies a public interface that can be used by other participants to call its methods remotely. We now describe the components depicted in figure 1 and its connections:

- **Visual processor** (VP): Acts as a vision server capturing images from the cameras and computing Harris-Laplace regions of interest at multiple spatial scales as described in Lowe[4]. For each set of ROI's, it fills a shared double buffer to supply quick responses to client requests.
- **Head controller** (HC): Head motor controller. Computes direct and inverse kinematics of the binocular head. It waits for client commands and passes them to a dedicated microcontroller that executes the PID loops. It keeps a copy of the state of the motors and of the head and can answer queries about them directly.
- **Base controller** (BC): Base motor controller. Computes direct and inverse base kinematics and waits for client commands that are passed to a dedicated microcontroller that executes the PID loops. It keeps a copy of the state of the motors and of the mobile base and can answer queries about them directly.
- **Vergence controller** (VC): Works independently to ensure the convergence of both cameras to the same spatial point. Vergence control is done by a multi scale cross correlation between the centre of the dominant camera and the epipolar homologous window in the other one. It looks for the maximum of the resulting structure (maximum of the image window at the whole scale-space) and performs the shift in the slave camera that leads to the convergence of both. To optimize resources the search is done in an increasingly wider window triggered by a failure in the direct matching of foveas. Vergence controller is a client of VP and HC.

- **ROI Maintenance component** (RM): Maintains a stable representation in time of recently perceived regions of interest. It provides a map of regions built in the camera reference system. It works as a short-term memory maintaining information about each region such as: raw image window, permanence time, attention time and update time. As VP does, this component always maintains an available ROI list that can be sent to its clients on demand.
- **Categorization component** (CC): Classifies regions of interest into known categories that can be used as landmarks. The current implementation uses Euclidian distance to compare SIFT[5] and RIFT[6] descriptors of candidate ROI's with previously stored examples. It accepts a target category so it tries to find a region compatible with the target and returns a list of candidates.
- **Spatial component** (SC): Compute spatial and geometric features of the last received ROI list and organize them in a head reference system. Properties computed by this component are: 3D position of the region using vergence and disparity; and planarity and plane orientation of the overall region by estimating the best homography. More useful properties will be implemented in near future to determine local shape with greater precision so more sophisticated hierarchical categorization can be accomplished in collaboration with CC.
- **Target selectors** (TS): Integrate local representations from CC and SC according to some criteria. Attending to their functionality, they request for specific patterns and features to CC and SC, respectively. CC provides a list of classified regions that are integrated with spatial information from SC to construct a final saliency map. This map is used to select a focus of attention that can be sent to the tracker component. This last action only takes place when the component is active. TS are specialized on a specific action. They are in communication with other components outside the attention system that activate them to take the attention control in order to carry out an action. An example of target selector component is the landmark selector, which is linked to a follow-landmark action. Another one is the obstacle selector, related to an action of avoiding-obstacles. The attention-action relationship is not a mandatory condition. A target selector can be linked to an idle behaviour that allows the system to maintain a pure bottom-up attention.
- **Tracker component** (TC): Receives the location of a region from the active TS and maintains the focus of attention on such region until another position is received. To achieve this goal, the TC implements a predictive tracking algorithm that combines the distance among RIFT descriptors of the regions and normalized correlation in YRGB space.

4 Interaction Dynamics

The architecture just described provides an attentional mechanism that can be incorporated in a wider network of components adding behaviour-based control, task selection, planning, topological maps and other abilities. In the experiment shown here, we use a couple of coordinating behaviours -approach and avoid- that activate the target selectors (TS's) of the attentional system. Together, they can be seen as a visual-goto-point (VGP) compound behaviour which is the basic serializing constituent of most complex navigation tasks. The attentional mechanism provided to VGP endows it with inner dynamic loops that take care of target detection, recognition, searching and tracking, lost target recovery and unexpected obstacle detection. In addition, these features are the result of parallel activities that get serialized to gain access to the orientable cameras.

When VGP activates, two activities take place simultaneously: a target landmark is downloaded to the attentional system through a TS, and another TS is activated to detect potential obstacles in front of the robot. Both TS's are coordinated in a very basic way by the VGP to achieve the current goal. The law to follow is hierarchical: “if there is free way, approach the target”. First, the landmark TS activates to search the target. Once it is fixated, the obstacle TS activates to locate potential hazards in the course towards the detected landmark. If the near space is free of obstacles, the cameras will search again the landmark and the base of the robot will reorient towards the gaze direction and start moving forward. Then again, the obstacle TS will use its short term spatial memory representation and covert attention capabilities to gaze towards any close enough obstacle in the way. If it happens, the base will reorient in a direction perpendicular to the pan angle in order to avoid the nearest obstacle. Once the danger is over, the landmark TS will regain control to relocate the target and establish a new heading direction. This alternating dynamics keeps going on until the goal landmark is within some specified distance and orientation.

Several kinds of attention can be observed in the last example. When the obstacle TS chooses a target to attend to, it performs overt attention on the obstacle so avoidance based on pan angle can take place. But when the landmark TS is waiting, it performs covert attention on the target so it can be quickly fixated again when it gets activated. In a similar way, we can speak about bottom-up and top-down attention. ROI's detected by the VP drive attention in a bottom-up way selecting those areas of the image most informative. From this set, a few are chosen attending to task dependent constraints such as target landmark or specific known geometric properties of obstacles.

5 Experiments

The visual attention system has been tested in a threaded mobile robot endowed with a stereo vision head. It encloses five degrees of freedom with digital PID controlled servos and two ISight Firewire cameras (figure 2). This type of robots has been developed in our Laboratory and is widely used in prototyping and algorithm testing.

The architecture of components just described runs on a local cluster of computers. Each process is an independent C++ application, which includes Ice objects and proxies to communicate with other processes. An Ice object is an entity used by a server to respond to client requests. An Ice proxy represents an Ice object local to the client that can communicate remotely to the server. Ice provides a remote method invocation capability that can use both TCP and UDP as the underlying protocol. In our current implementation, the network of processes is distributed among four physical processors – dual Opteron board for VP, HC and BC; 3GH-HT P4 for VC, TC and RM; and AMD64 dual core for CC, SC and TS- as seen by the Linux operating systems. The computers are locally linked by a 1 Gb ethernet switch providing enough bandwidth for real time communication among components.

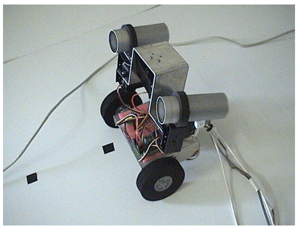

Fig. 2. Robot used in the experiments.

We have designed a simple experiment for initial testing and validation of the proposed architecture. The robot has to localize and approach a landmark (star) in its near space avoiding an obstacle that blocks its heading direction. The running system incorporates all the components described before. As target selectors (TS's) we use landmark and obstacle selectors working in cooperation. Actions linked to selectors are approach and avoid, configuring a sort of visual goto-point.

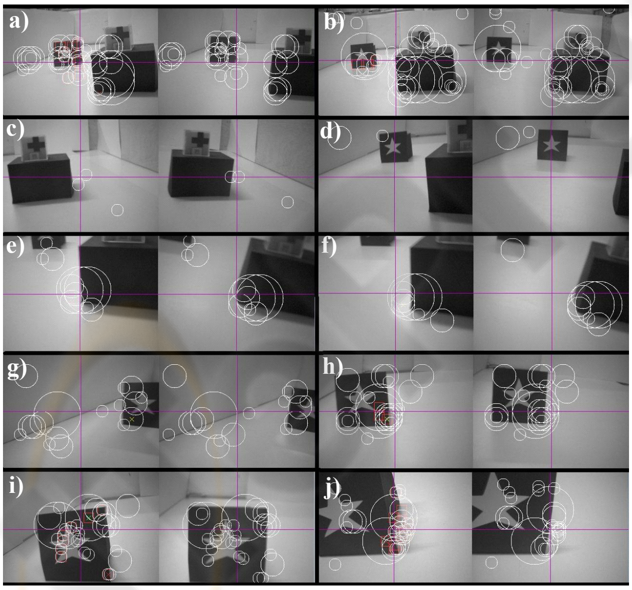

Fig. 3. Experiment of visual navigation avoiding obstacles.

Changes of attention alternating these two kinds of targets can be appreciated in the sequence above (figure 3). Initially, a) attention is fixated on the landmark and an action of approaching begins. Then, b) and c) frames, obstacles gain control of attention guiding the robot to avoid them. After several frames - d) - landmark is fixated again providing a new goal heading. To keep up with the new situation, the obstacle selector changes its focus of attention, e) and f). Once all the obstacles have been avoided g), attention is again centred on the landmark making the robot to approach it and finally reach the goal position, j).

6 Summary and Conclusions

In this paper, we have shown an experiment in distributed visual attention on a mobile robot with a stereoscopic head. Our goal has been to test the potential of combining ideas from visual neuroscience, distributed software engineering and robotics. We think that the proposed architecture will ease the way to model and implement more perceptual and cognitive capabilities in our robots. This first result shows that the complexity of distributed bottom-up and top-down attention can be integrated in a visual navigation framework to solve what we have called the visual-goto-point problem. Much work remains in this multidisciplinary area, but the possibilities offered by new multicore processors in conjunction with communications middleware will open new spaces for bio-inspired robotics modelling and building.

References

1. Internet Communication Engine. http://www.zeroc.com/ice.html
2. L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research, 40(10-12):1489–1506, May 2000.
3. W. James, The principles of psychology, New York: Holt, 1890, pp. 403-404
4. D. G. Lowe. Object recognition from local scale-invariant features. In ICCV, pages 1150–1157, 1999.
5. D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.
6. K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In ICCV01, pages 525–531, Vancouver, Canada, July 2001.
7. A.D. Milner and M.A. Goodale. The visual brain in action. Oxford UniversityPress. Oxford 1995
8. V. Navalpakkam, L. Itti, An Integrated Model of Top-down and Bottom-up Attention for Optimal Object Detection, In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-7, Jun 2006
9. R. P. Rao, G. Zelinsky, M. Hayhoe, and D. H. Ballard. Eye movements in iconic visual search. Vision Research, 42(11):1447–1463, Nov 2002.
10. A. Treisman and R. Paterson, Emergent features, attention and object perception, J. Exp. Psychol: Human Perception and Performance, 1984, 10:12-31.
11. A. Torralba et al. Contextual guidance of attention in natural scenes: The role of global features on object search. Psycological Review, October 2006
12. J.K. Tsotsos, et.al., Modeling visual attention via selective tuning, Arti. Intell., 1995, 78:507-545.