Vol. 2, No. 2, Page 93 of 100

An Overt Visual Attention Mechanism Based on Saliency Dynamics

José M. Cañas*, Marta Martínez de la Casa and Teodoro González
Universidad Rey Juan Carlos, Departamento de Ing. Telemática y Tec. Electrónica, Spain

Received 1 January 2007; revised 2 February 2007, accepted 3 March 2007

Abstract
A visual overt attention mechanism is presented in this paper. Our algorithm chooses the next fixation point for a robot mobile camera in order to track several objects around the robot simultaneously, even if the whole set of them can not be covered by the same camera image. Our approach is based on two related measurements, liveliness and saliency, that dynamically evolve depending on the image observations and the camera movements. The attention is shared among exploring the surroundings for salient features, reobserving the tracked objects and other task-dependent points to look at. A winner-takes-all competition provides a flexible time sharing behavior with natural appearances of new objects, disappearances and inhibition of return. It also accepts top down features to look for and different priorities for them. Several experiments have been carried out with a real Pioneer robot endowed with mobile camera and are also described.

Keywords
Visual attention, robotics, saliency

1. INTRODUCTION

The use of cameras in robots is continuously growing. They can potentially provide the robot with much information about its environment and in the last years they have become a cheap sensor. Most service robots, including humanoid prototypes, are equipped with vision as it is the most promising technology for human-robot interaction. But dealing with the huge amount of data carried by video streams is not easy. The visual attention offers a solution for the processing bottleneck generated by such overwhelming source of raw data, especially convenient in machines with limited computational resources.

An attention mechanism of human vision system has been source of inspiration for machine visual systems, in order to sample data non uniformly and to utilize computational resources efficiently [2]. The performance of the artificial systems has been always compared to the performance of several animals, including humans, in simple visual search tasks. In last years, biological models are moving to the real-time arena and offer an impressive flexibility to deal simultaneously with generic stimulus and with task specific constraints [7,11]. Current trends in the design of visually guided autonomous robots urge for integration of recent advances in biological attention mechanisms.

Machine attention systems have been typically divided into *overt* and *covert* ones. The *covert attention mechanisms* [16,9,10] search inside the image flow for relevant areas for the task at hand, leaving out the rest. Search of autonomous vehicles in outdoor scenarios for military applications [8], and search for traffic signals inside the images from the on-board car cameras are just two sample applications.

The *overt attention systems* [17,8,14] use mobile cameras and cope with the problem of how to move them: looking for salient objects for the task at hand, tracking them, sampling the space around the robot, etc.. The saccadic eye movements observed in primates and humans are their animal counterpart. They have been used, for instance, to generate a natural interaction with humans in social robots like Kismet [3]. This active perception system can guide the camera to better perceive the

*José María Cañas Plaza, Universidad Rey Juan Carlos, c/ Tulipán s/n, 28933 Móstoles, Spain, Tel: +34 91644 74 68, josemaria.plaza@urjc.es

relevant objects in the surroundings. The use of camera motion to facilitate object recognition was pointed out by [2] and has been used, for instance, to discriminate between two shapes in the images [10].

Most successful systems define low level salient features like color, luminance gradient or movement [8]. Those features drive the robot attention following an autonomous dynamics in a close loop with the images. This way, the system is mainly bottom-up guided by the low level visual clues. One active research area is the top-down modulation of these systems, that is, how the current task of the robot or even the high levels of perception, like object recognition [15,12], can tune the attention system and maybe generate new focus of attention.

In our scenario, 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 [6]. This poses a problem when such objects do not lie completely into the cameras field of view. Some works use omni directional vision, and they have been successfully applied in problems like visual localization or soccer behaviors in RoboCup competition. Other approaches use a regular camera and an *overt attention mechanism* [9,18], which allows for rapid sampling of a very wide area of interest.

In this paper we report on an overt attention system for a mobile robot endowed with a pan-tilt camera, which can be oriented at will independently from the robot base. 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 explore the scene and answer to two questions: how many relevant colored objects are there around a robot? and, where are they located?. From an active vision viewpoint, this process of continuous trading between search and refresh can be seen as a situating process, in the sense of grounding visual objects to the external world [13].

Following this introduction, second section describes our attention mechanism, its dynamics of liveliness and saliency. Many experiments have been carried out on a real robot, testing the performance and behavior of the algorithm on a real setup. They are summarized in fourth section. Finally some brief conclusions end the paper.

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. For the sake of clarity, we will initially consider that only the pink objects are relevant.

The color images are Cartesian, 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 include 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

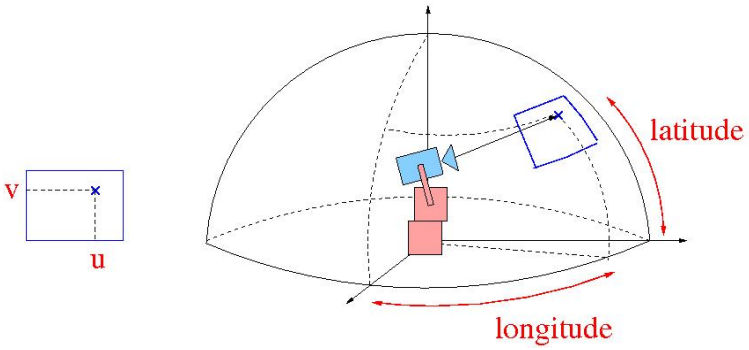

Figure 1. Scene space and coordinates.

```
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
     probabilistic insertion of exploration points
     choose the most salient fixation point
end_loop
```

Figure 2. Pseudocode of our overt attention algorithm.

image is processed while the pantilt is moving, but once it has stopped at current fixation point the monocular image is processed to update the scene representation, the next target point for the pantilt is computed and commanded. 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 HSI space, which is more robust to changes in illumination than RGB.A fast histogramic clustering algorithm [5] is used. For illustration, after an initial teleoperated seep of the camera the scene image built is shown in Figure 3 (left). The scene image is projected for visualization purposes into the display using a polar transformation, where ρ=*latitude* and θ=*longitude*.

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.

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

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

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,

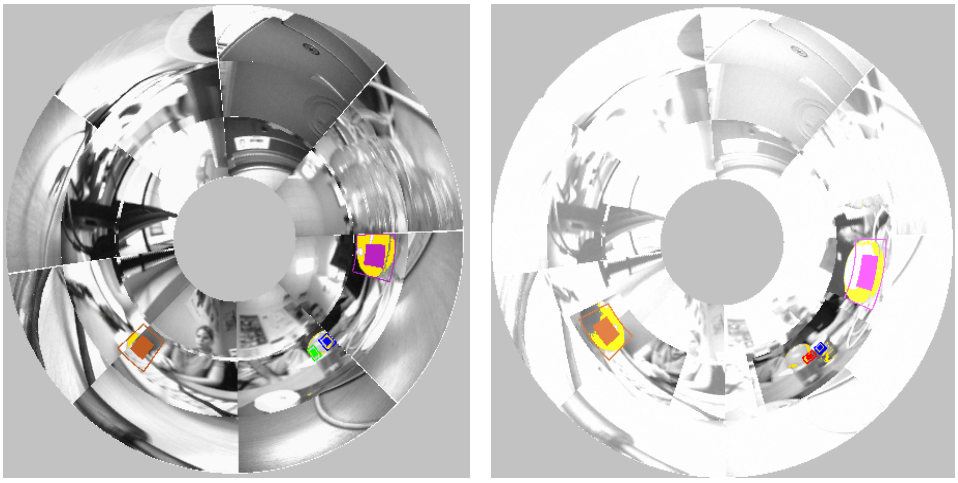

Figure 3. Scene image after the initial sweep and the forgetting mechanism.

while areas of the scene that have been recently visited show the fresh projected monocular patch.

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.

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.

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

$$sal(fixp,t) = 0 \quad (4)$$

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 [12], neither explicitly keeping a saccade history [18]. 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.

The algorithm allows the specification of several features to indicate relevant objects and it offers a different saliency slope for each of them. For instance, our system may be interested in pink and blue objects, and associate a different ΔS_{time} (3) for each type. This way, our attention mechanism allows several priorities: the higher the ΔS_{time} of a type, the faster saliency of those objects will grow up and gain the attention of the camera. Providing different slopes, the system will be more time looking at high priority objects, in mean values, than at low priority objects. These priorities provide a way to balance the importance of the respective objects for the current robot tasks.

Exploration of the scene

While the pantilt is stably jumping among a certain set of valid scene objects, new objects out of current scope are also searched for. New locations are explored by periodically introducing pioneer fixation points into the list. Once the camera is looking at one of such points, in case of some relevant brand new objects really observed there their positions will be inserted as new fixation points. In case of no relevant object observed there, such fixation point will be simply removed from the list.

The exploration points are generated in two different fashions. First, (a) some of them are obtained sampling from a uniform distribution in space, performing a completely random search. Second, (b) other exploration points are generated following certain sweep pattern, like a systematic sweep to make sure that eventually the whole scene is explored. This pattern may be task-dependent as may concentrate the exploration points on different areas of the scene. For instance, a navigation application may lay exploration points in the space just in front of the robot, the closer the better, because further areas are less significant.

Two probability thresholds say whether any exploration point ((a) or (b) respectively) is inserted or not in the list in current iteration of the algorithm. This way, the amount of exploration points can be tuned from the upper cognitive levels.

3. EXPERIMENTS

A lot of experiments have been conducted on a real robot to validate our attention algorithm and test its performance.

The experimental setting includes an ActivMedia Pioneer endowed with a Directed Perception pantilt unit and Videre firewire camera (Figure 4). The pantilt unit accepts position commands through the serial port, and the camera provides a flow of 30 fps of 320x240 color images. In the experiments some pink and blue balls were located around the robot, as our focus here was the attention sharing, not the features to be tracked themselves.

Starting with a single object, the system is able to keep it tracked and to follow its (slow) movements around (Figure 4). Figure 5 shows the scene built when exploring the environment with a regular pattern and refreshing the single tracked object at the same time. Changes in illumination slightly move the visual center of the object. To avoid small pantilt oscillations a minimum distance is required in order to really command a new pantilt target.

This tracking could have been solved with classical closed loop techniques, but here we have solved it using exactly the same dynamics that will generate the tracking behavior for two, three and more objects, and the time sharing of the pantilt unit among them. In addition, one limitation of the current implementation is the maximum speed of the objects that the system can properly track. Only slow objects are successfully tracked. Faster hardware will for sure alleviate this limitation. For two objects the system reached a stable jumping loop. The saliency evolution for a scene with two objects can be seen at Figure 6 (left). The pattern shows a perfect alternating sharing of the pantilt device. Figure 6 (center) shows the liveliness evolution for such experiment, both objects were kept at high values as they are continuously observed. Figure 6 (right) displays how the liveliness of one of the pink balls lowed down when such ball disappeared from scene.

Figure 7 (left) shows an experiment with three pink balls around the robot. Figure 7 (center) displays the scene image for a similar situation, where the exploration was intentionally disabled for the sake of clarity. The pantilt continuously oscillated among the three pink balls, in a stable loop, jumping from one tracked object to another, in a round robin sequence. The visit pattern among the (numbered) balls was the following: 1-2-3-1-2-3-1-2-3. Only those three areas of the scene are continuously refreshed. Other areas are not visited by the pantilt and then they fade to white values.

To test the forgetting capability of our algorithm we hided the ball in the middle. The scene image evolved to 7 (right): after some iterations, the algorithm forgets the central ball and stops visiting its old location, changing to naturally jump among the two remaining ones. The central position gradually fades to white. The visit pattern was the following after the subtraction of the central ball (ball number 2): 1-2-3-1-2-3-1-3-1-3-1-3.

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 8 (left)) the perceived scene was that at the Figure 8 (center). Then the robot slowly moved forward 70cm, approaching to the objects, and the objects spread out in the scene image, as shown in figure 8 (right).

To probe the effect of different priorities, we populated the environment with two pink balls (numbered 2 and 4) and two blue objects (numbered 1 and 3), and assigned them different priorities. After several iterations the scene imaged evolved to Figure 9, where only the four relevant objects appeared. Nevertheless they were not visited equally. The visit pattern was: 1-2-4-2-4-3-1-2-4-2-4-3-1-2-4-2-4-3, with the attention algorithm showing a clear bias towards the high priority objects.

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

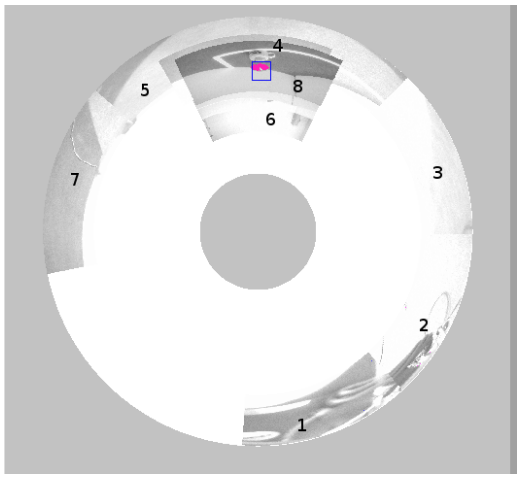

Figure 5. The attention is shared among exploration and tracking.

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 are 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.

4. CONCLUSIONS

A novel overt attention mechanism has been presented. It expands the visual scope of the robot single camera taking advantage of its movement. For instance, it can find how many pink colored objects there are around a robot and where they are the located, despite 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 positions of already known objects are candidate fixation points for the pantilt device. Random exploration points are also included as candidates. The liveliness of every object decreases with time, but increases when it appears in the monocular image. The saliency of the fixation points grows up with time, and is set to zero when the pantilt fixates at it. The pantilt movement is always set to the most salient fixation point among the candidates.

The algorithm accepts top-down modulation coming from high level cognitive processes. They can determine what low level features are relevant, they can assign different attention priorities to them, and they can set exploration patterns adjusted to the current robot task.

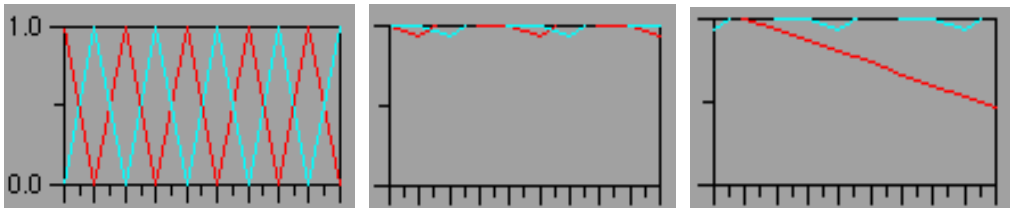

Figure 6. Saliency (left), liveliness (center) evolution with two pink objects in the scene. In (right) one of them disappears from the scene.

Figure 7. The robot has three pink balls around and pantilt oscillates among them.

Figure 8. When the robot approaches to the objects, they spread out in the image scene.

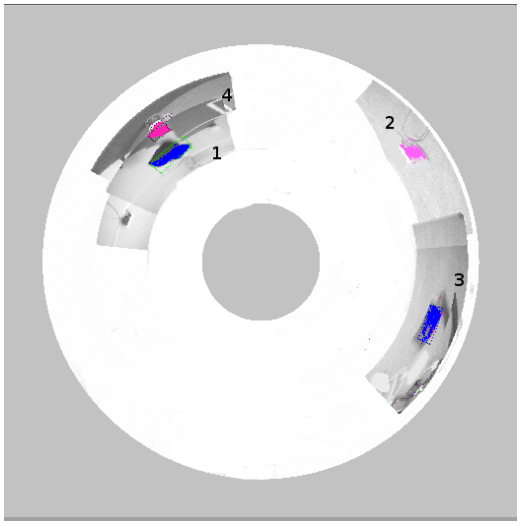

Figure 9. Objects with high priority are refreshed more often.

Such simple dynamics generate several interesting behaviors, as shown in the experiments with a real robot. First, it alternates the focus of attention among exploration and revisiting the relevant objects of the scene, regardless their amount, 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.

We are now working in extending the algorithm to stereo pairs and placing the focus points in 3D surroundings of the robot. Another future line is the insertion of new fixation points to test hypothesis coming from the perceptive analysis of the images. Conclusions are inserted at this point, before acknowledgements.

ACKNOWLEDGEMENTS

This work has been funded by Spanish Ministerio de Ciencia y Tecnología, under the project DPI2004-07993-C03-01 and Comunidad de Madrid under the project RoboCity2030, S-0505/DPI/0176.

REFERENCES

[1] R. Bajcsy, “Active Perception,” Proc. of IEEE, Vol. 76, pp. 996-1005, 1988.

[2] D. H. Ballard, “Animate Vision,” Artificial Intelligence, Vol. 48, pp. 57-86, 1991.

[3] C. Breazeal and B. Scassellati, “A context-dependent attention system for a social robot,” Proc. of Int. J. Conf. on Artificial Intelligence, pp. 1146-1151, 1999.

[4] D. T. Cliff and J. Noble, “Knowledge-based vision and simple vision machines,” Philosophical Transactions of the Royal Society of London, Vol. 352, pp. 1165-1175, 1997.

[5] V. Gómez, J. M. Cañas, F. San Martín, and V. Matellán, “Vision based schemas for an autonomous robotic soccer player,” Proc. of IV Workshop de Agentes Físicos, pp. 109-120, 2003.

[6] J. M. Cañas, P. Díaz, P. Barrera, and V. Gómez, “"Visual memory for robot navigation using JDE architecture,” Proc. of VII Workshop de Agentes Físicos, pp. 91-98, 2006.

[7] L. Itti, "Models of bottom-up and top-down visual attention,” PhD dissertation, California Institute of Technology, 2000.

[8] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention,” Vision Research, Vol. 40, pp. 1489-1506, 2000.

[9] L. Itti and C. Koch, "Computational modelling of Visual Attention," Nature Reviews Neuroscience, Vol. 2, pp. 194-203, 2001.

[10] D. Marocco and D. Floreano, "Active vision and feature selection in evolutionary behavioral systems," Proc. of Int. Conf. on Simulation of Adaptive Behavior, pp. 247-255, 2002.

[11] V. Navalpakam and L. Itti, "Modelling the influence of task on attention," Vision Research, Vol. 45, pp. 205-231, 2005.

[12] F. Orabona, G. Metta, and G. Sandini, "Object-based visual attention: a model for a behaving robot," Proc. of Int. Workshop on Attention and Performance in Computational Vision WAPCV-2005, 2005.

[13] Z. Pylyshyn, "Visual Indexes, preconceptual object and situated vision," Cognition, Vol. 80, pp. 127--158, 2001.

[14] C. Soyer, H. Bozma, and H. Istefanopolus, "APES: attentively perceiving robot," Autonomous Robots, Vol. 20, pp. 61-80, 2006.

[15] Y. Sun and R. Fisher, "Object based visual attention for computer vision," Artificial Intelligence, Vol. 146, N. 1, pp. 77-123, 2003.

[16] J. K. Tsotsos et al, "Modeling visual attention via selective tuning," Artificial Intelligence, Vol. 78, pp. 507-545, 1995.

[17] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt visual attention for humanoid robot," Proc. of Int. Conf. on Intelligence in Robotics and Autonomous System (IROS-2001), pp. 2332-2337, 2001.

[18] A. Zaharescu, A. L. Rothenstein, and J. K. Tsotsos, "Towards a biologically plausible active visual search model," Proc. of Int. WoWorkshop on Attention and Performance in Computational Vision WAPCV-2004, pp. 133-147, 2004.

AUTHOR INFORMATION

José M. Cañas received the Telecommunication Engineer degree in 1995, and a PhD in Computer Science in 2003, both from the Universidad Politécnica de Madrid. He has researched in Carnegie Mellon University and the Instituto de Automática Industrial. Currently he is associate professor at Universidad Rey Juan Carlos, where leads the robotics group. Dr. Cañas' research interests include robotics and artificial vision.

Marta Martínez de la Casa received the Computer Science Engineer degree in 2005 from the Universidad Rey Juan Carlos. She is working as a Software Engineer at INDRA.

Teodoro González received the Telecommunication Engineer degree in 1991 from the Universidad Politécnica de Madrid. Since 1988, he has worked as technical director, system consultant, project manager, programmer and researcher in computer applications companies. Currently he is associate professor at the Universidad Rey Juan Carlos and teaches computer science at a High School.