7

Attentional Selection for Action in Mobile Robots

Pilar Bachiller, Pablo Bustos and Luis J. Manso
Universidad de Extremadura
Spain

1. Introduction

During the last few years attention has become an important issue in machine vision. Studies of attentional mechanisms in biological vision have inspired many computational models (Tsotsos et al., 1995; Itti & Koch, 2000; Frintrop et al., 2005; Torralba et al., 2006; Navalpakkan & Itti, 2006). Most of them follow the assumption of limited capacity associated to the role of attention from psychological proposals (Broadbent, 1958; Laberge, 1995). These theories hypothesize that the visual system has limited processing capacity and that attention acts as a filter selecting the information that should be processed. This assumption has been criticized by many authors who affirm that the human perceptual system processing capacity is enormous (Neumann et al., 1986; Allport, 1987). From this point of view, a stage selecting the information to be processed is not needed. Instead, they claim the role of attention from the perspective of selection for action (Allport, 1987). According to this new conception, the function of attention is to avoid behavioural disorganization by selecting the appropriate information to drive task execution. Such a notion of attention is very interesting in robotics, where the aim is to build autonomous robots that interact with complex environments, keeping multiple behavioural objectives. Attentional selection for action can guide robot behaviours by focusing on relevant visual targets while avoiding distracting elements. Moreover, it can be conceived as a coordination mechanism, since stimuli selection allows serializing the actions of, potentially, multiple active behaviours.

To exploit these ideas, a visual attention system based on the selection for action theory has been developed. The system is a central component of a control architecture from which complex behaviours emerge according to different attention-action links. It has been designed and tested on a mobile robot endowed with a stereo vision head. Figure 1 shows the proposed control model. Sensory-motor abilities of the robot are divided into two groups that lead to two subsystems: the visual attention system, which includes the mechanisms that give rise to the selection of visual information, and the set of high-level behaviours that use visual information to accomplish their goals. Both subsystems are connected to the motor control system, which is in charge of effectively executing motor responses generated by the other two subsystems.

Each high-level behaviour modulates the visual system in a specific way in order to get the necessary visual information. The incoming flow of information affects high-level

Source: Advances in Robotics, Automation and Control, Book edited by: Jesús Arámburo and Antonio Ramírez Treviño,
ISBN 78-953-7619-16-9, pp. 472, October 2008, I-Tech, Vienna, Austria

behaviours guiding their actions, so the attentional system also modulates the behavioural system. This double modulation results in two control loops running in parallel, affecting each other and working in cooperation to get a common objective. The control loop associated to the visual system controls ocular movements that lead to the fixation of a visual target. The other loop controls the sensory-motor activity of high-level behaviours.
The attentional nature of the proposed system favours the simultaneous execution of multiple behaviours. In our model, action selection is solved by selecting a visual target. Fixation of a focus of attention ensures that only actions compatible with the selected information take place, solving this way the problem of coordination among behaviours. The sensory selection matter can be treated easily in comparison to the action selection problem. Actions could be contradictory, making more difficult the process of selecting the most suitable one. However, stimuli are selected according to the relevance of their properties in the current situation. At any given time and for any situation, the stimulus having the most suitable properties can always be found. It results in a sensory selection that provides the necessary information to carry out the most appropriate action for that situation. Extending this idea, attention acts as a means of action serializing, which is produced by the ordered execution of actions associated to the sequence of selected stimuli. The specific interleaving among actions comes from a time relation that links internal system requirements and external world features.

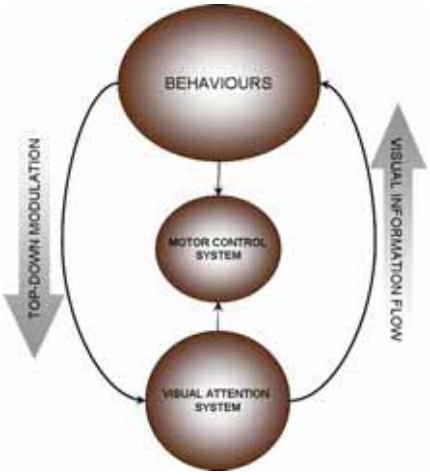


Fig. 1. Attention-based control model

2. The visual attention system

The proposed attention system has been modelled as a set of components collaborating to select, fix and track visual targets according to different task requirements. Components are organized in different processing levels as shown in figure 2. The low-level components are related to image acquisition, motor control, as well as computation and maintenance of regions of interest (ROI's). Components of intermediate level extract sets of ROI features related to "what" (appearance information) and "how" (spatial information) issues. These features are used by high-level components, called target selectors (TS), to drive attention

according to certain top-down behavioural specifications. Attention control is not centralized, but distributed among several target selectors. Each of them drives attention from different top-down specifications to focus on different types of visual targets. At any given time, overt attention is driven by one TS, while the rest attend covertly to their corresponding targets. The frequency of overt control of attention of each TS is modulated by high-level behavioural units according to their information requirements. The fixation of a selected target is accomplished by two independent camera movements: a saccadic and tracking movement in one of the cameras and a vergence movement in the other. This allows controlling attention from monocular information while keeping stable binocular fixation.

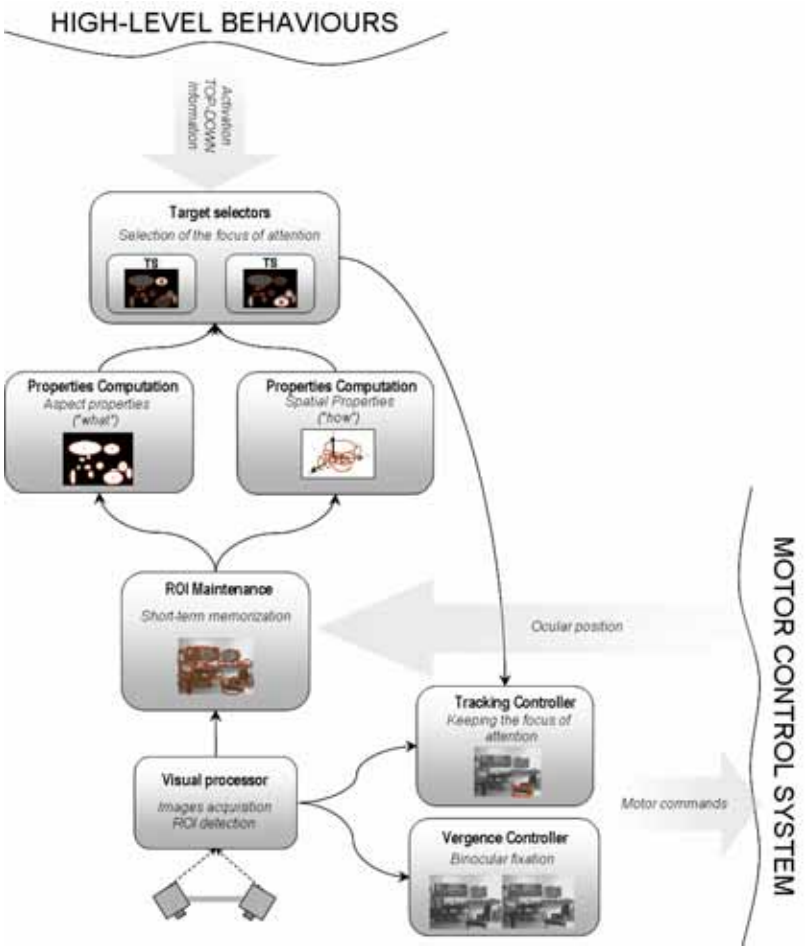


Fig. 1. Architecture of the proposed visual attention system

The whole process constitutes a perception-action cycle that begins with image acquisition and detection of regions of interest and provides a motor response as a result of the selection of a focus of attention. The rest of this section details the different processing stages of this cycle.

2.1 ROI detection and maintenance

In its first stage, system detects image regions that exhibit highly informative contents. To get stable results, the detection process must be invariant to different transformations, such as different scales or illumination changes. Several interest point detectors have been proposed in the literature (Lindeberg, 1998; Lowe, 1999; Mikolajczyk & Schmid, 2001). Among the existing approaches, the Harris-Laplace method, proposed by Mikolajczyk and Schmid (Mikolajczyk & Schmid, 2001), shows outstanding results regarding to important scale changes as well as different rotations, translations and changes of illumination. This method is highly suitable for our purpose. However, looking for a reduction in processing time, we propose a variant of the Harris-Laplace interest point detector. Our proposal consists in applying the Harris-Laplace method to the multiscale prism resulting from a centred section of the scale space representation (figure 3). This new representation exhibits a Cartesian topology that approaches the retinal structure of the human eye (Bandera & Scott, 1989).

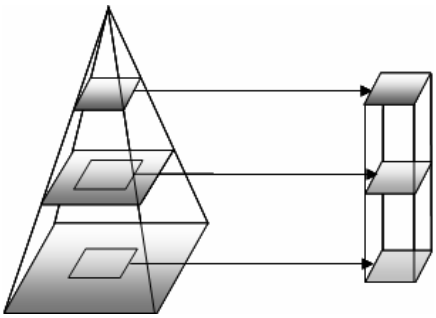


Fig. 3. Multiscale prism

The multiscale prism keeps an eccentricity-dependent resolution structure. With this representation, foveal areas are represented at every resolution level, whereas peripheral areas only appear at the higher levels. As a consequence, Harris-Laplace method detects regions of any size at the fovea and big-size regions at the periphery (figure 4). This provides a significant reduction of processing time at the expense of some loss of visual information. Nevertheless, non-detected regions correspond to less informative areas of the periphery. They can be recovered at any moment placing the fovea onto these areas through camera movements.

Regions detected by the above process are integrated in a map built on the camera reference frame. This map serves as a short-term memory that maintains information about recently perceived regions of interest. Each region is indexed by its angular position. This allows keeping information about regions situated outside the current visual field, so attention can quickly shift to previously seen regions.

At every processing cycle, new detected regions are stored in this memory and the previously stored ones are updated according to the new visual information of the scene. The updating process consists in searching every stored region in the neighbourhood of its position inside the current scale space. The spatial search allows locating regions whose size changed due to, for example, a translation movement of the robot.

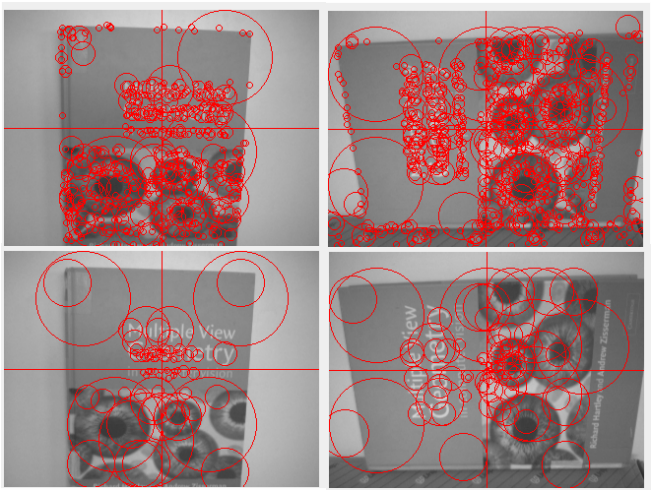


Fig. 4. Harris-Laplace regions from multiscale pyramid (upper images) and prism (lower images)

Each region maintains an attribute of permanence time that is increased or decreased depending on the success of its position updating. Regions whose permanence time is under a threshold are considered forgotten and are thus removed from the memory map. For remaining regions, other attributes, such as raw image window, 2D velocity and attention time, are updated. The group of attributes computed during this stage forms a first set of properties that are extended in later stages of the system. To get different pathways of processing that allow computing properties in a separate and independent way, regions keep an identification number. This identification value allows integrating subsets of separately computed properties that come from the same original region.

2.2 Computation of high-level properties

At this stage, processing flow is divided into two subsystems that compute high-level properties of each region of interest. These two subsystems are an analogous of the "*what*" and "*how*" systems proposed in neuroscience (Milner & Goodale, 1995). This division allows a specialization of functions, dedicating specific resources to each subsystem and sharing what is common from low-level processes. The "*what*" system computes properties related to the visual appearance of regions. The "*how*" system extracts spatial information of each visual region, such as position, orientation and movement.

The "*what*" system

Properties computed by the "*what*" system must provide region categorization from its visual appearance. To get consistent results in this categorization process, properties must be invariant to transformation of the point of view, changes of illumination, etc. Histogram-based descriptors, like SIFT (Lowe, 2004), RIFT and Spin (Lazebnik et al., 2005), provide stability against different image deformations. Experimental results obtained from several comparative studies (Mikolajczyk & Schmid, 2005) show better discriminative power of this kind of descriptors in relation to traditional descriptors, like banks of filters and differential invariants. Based on these results and considering a lower computational cost, we use RIFT

and Spin descriptors (Lazebnik et al., 2005) to represent aspect properties of regions. Both descriptors are invariant to scale and rotation changes and extract information about gradient orientations and image intensities, respectively.

The "*how*" system

The group of properties extracted by the second subsystem is composed of features of regions that describe their action possibilities. They must provide information about how to interact with each region. In our current implementation, properties computed by this component are: 3D position of the region computed using stereo matching; spatial motion from region tracking and 3D position; and planarity and plane orientation of the overall region gathered by homography estimation.

2.3 Selection of the focus of attention

The selection of the focus of attention is accomplished by multiple control components (target selectors) that individually determine the region to get the focus of attention following specific criteria. Distributed control of attention provides strong advantages compared to centralized control. Firstly, it provides a clearer and simpler design of the selection process. Secondly, it admits the coexistence of different types of visual targets, which is a key aspect of any task-driven attentional system. In this sense, we hypothesize that it is necessary to separate the way properties of regions are integrated for attentional selection according to behavioural objectives. For instance, in a generic navigation task such as going somewhere by following a predetermined set of landmarks, two types of visual targets can be distinguished: landmarks guiding navigation and potential obstacles. Integration of the different sets of properties characterizing both targets could turn any distracting area for the described task into a salient region, so defining global selection criteria including the two targets could be unfeasible. Moreover, although an effective integration of different kind of targets was achieved, the system can not guarantee to provide the necessary sequence of visual fixations that allow properly distributing attention time among them. If we put these ideas in the context of robot surveillance, a more important question emerges: how are reactive and deliberative abilities integrated? If attentional control is centralized, this issue becomes complicated. It is due to the fact that unexpected and foreseen things, such as moving and static objects, could be defined by contradictory properties. As a consequence, a centralized control strategy would cause either an excessive alerting state or a passive response to unsafe situations.

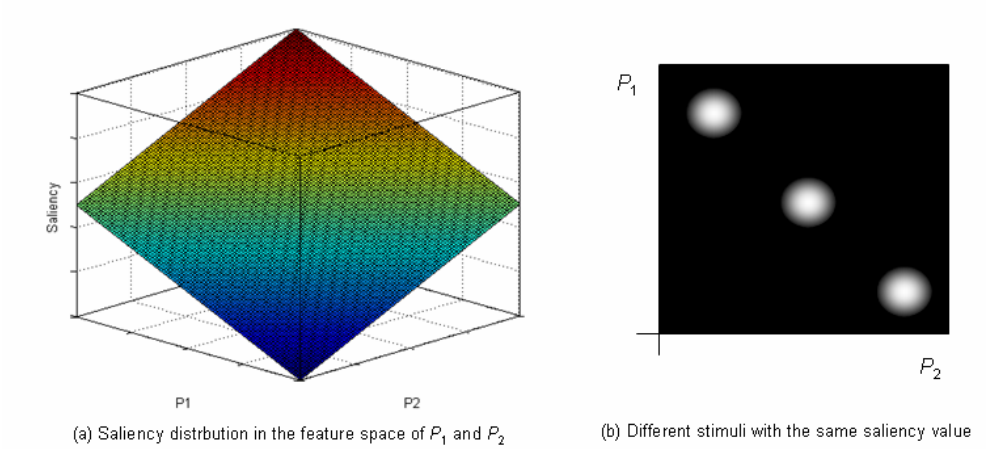
Individual selection process

Each target selector selects a focus of attention whose properties keep the greatest correspondence with its selection criteria. For this purpose, it computes a saliency map that represents the relevancy of each region according to its top-down specifications. This map acts as a control surface whose maxima match with candidate visual regions to get the focus of attention.

There are several alternatives for computing the saliency map. The most representative models that integrates top-down and bottom-up influences in attentional control employ a weighting process where features are modulated by multiplicative factors representing top-down specifications (Frintrop et al., 2005; Navalpakkan & Itti, 2006). This integration method presents some limitations that can be observed in figure 5. This figure shows the behaviour of a weighting method for the selection of a visual target on the basis of two properties (P_1 and P_2) . For any value of P_1 and P_2, saliency is computed as:

$$saliency = P_1 w_1 + P_2 w_2 \quad (1)$$

where w_1 and w_2 are top-down weights associated to P_1 and P_2, respectively.
Assuming that both properties are equally relevant in the selection process, weights w_1 and w_2 should have the same value to express the desired target description. Hence, there is no distinction between considering simultaneous or separate matching with several properties, which limits the definition of potential selection criteria. Moreover, as it can be observed in figure 5, the resulting saliency distribution in the feature space (a) would cause some ambiguity in the selection process. Thus, as shown in (b), any horizontal line across the plane in (a) gathers stimuli with equal saliency even though they correspond to very different candidate regions.



(a) Saliency distrbution in the feature space of P_1 and P_2 (b) Different stimuli with the same saliency value

Fig. 5. Saliency computation using a weighting method (see the text above for details)

The proposed selection process overcomes the limitations described above by using fuzzy techniques. Fuzzy logic (Zadeh, 1965) provides a framework of development that is very suitable for our purpose, mainly for three reasons:

- Properties used in the selection process present some degree of uncertainty and imprecision. Fuzzy logic allows dealing with this problem by treating each property as a linguistic variable.
- Selection criteria can be easily defined using fuzzy rules that associate matching of properties to saliency levels.
- Output of every rule can be fused together using a fuzzy inference process that integrates all selection criteria to provide a final saliency value.

Following these ideas, the saliency map is obtained from a set of fuzzy rules that associate, for every visual region, the fulfilment of certain selection criteria (antecedents) to different saliency levels (consequents). Each premise of a rule antecedent is expressed by a linguistic label on a region property. These labels give rise to fuzzy sets defined in the domain of the properties that are relevant for the selection of the focus of attention. Thus, each selection criterion is expressed by a rule antecedent that determines matching between properties and associated fuzzy sets. The consequent part indicates the saliency value that is assigned to a region that fulfils that selection criterion. Since selection priority depends on the degree of fulfilment of antecedents and not directly on property values, we use a zero-order Takagi-

Sugeno model (Takagi & Sugeno, 1985). Hence, rule outputs are expressed through constant values that allow ordering regions according to the corresponding selection criteria. Thus, a rule of the proposed system is written as:

IF P_1 **is** A_i **AND** P_2 is B_j … **AND** P_n **is** X_k **THEN** *saliency* $= s_{ij..k}$

where P_1, P_2, …, P_n are properties of regions, A_i, B_j, …, X_k fuzzy sets associated to those properties and $s_{ij...k}$ a constant value expressing a saliency level.

To assign specific values to each rule output, two types of rules are distinguished in our system: exclusion rules and selection rules. Strictly negatives values (considering a 0 value as negative) are associated with exclusion rules, whose antecedents define regions that have no interest in the selection process and, therefore, should not be attended. Positive values are assigned to selection rules in such a way that rules that express greater fulfilment of top-down specification will take greater saliency values than those that define less important regions.

To illustrate these ideas, a fuzzy system for an obstacle selector is shown below. In this example, robot must detect potential obstacles situated in the trajectory to a target position. Three properties are considered for determining the obstacle quality of any visual region:

- *Relative Depth (RD)*: it is defined as the relation between region depth and target depth. Both depth values are computed at a reference system situated in the ground position of the robot with its Z axis fixed in the straight trajectory to the target. This relation quantifies the proximity degree of a region to the robot, to the target or to both of them. A value greater than 1 of this property can be considered as a region situated outside a potential trajectory to the target. Three fuzzy sets are considered for this property (figure 6): *NearR, NearT* and *Far*. *NearR* considers depths that are nearer to the robot than to the target, *NearT* is related to regions situated in a nearer depth to the target than to the robot and *Far* is associated to regions situated behind the target.
- *Deviation (Dv)*: it quantifies the proximity degree of a region to the straight line between the robot and the target. It is measured by the distance between the central point of the region and that straight line. This property is partitioned in three fuzzy sets (figure 7): *Low*, *Medium* and *High.*
- *Height (He)*: Properties described above consider ground parallel distances between a region and the robot. However, those regions situated high enough from the ground should not be selected as obstacles, since they are not interfering areas in the way to the target. For this reason, height of regions is included as the third property that allows defining selection criteria of our system. This property allows deciding whether to include or not regions as potential obstacles, so only two fuzzy sets are considered (figure 8): *Low* and *High.*

Once fuzzy partitions have been determined, the rule system is established by defining a fuzzy rule for each combination of fuzzy sets associated to the group of properties. For the current example, it results in a system with 18 rules, i.e., one rule for each combination of fuzzy sets of *He*, *Dv* and *RD*. Table 1 shows a simplified representation of the whole rule system. Each inner cell corresponds to a fuzzy rule whose antecedent is formed by the conjunction of three premises: the premise about *Height* (*He*) of the superior column, the premise about *Deviation* (*Dv*) of the inferior column and the premise about *Relative Depth* (*RD*) of the corresponding row. Inside each cell, the consequent value of the associated rule is shown. For instance, the first column of inner cells represents the following three rules:

IF *He* **is** *Low* **AND** *Dv* is *Low* **AND** *RD* **is** *NearR* **THEN** *saliency* $=s_{111}$
IF *He* **is** *Low* **AND** *Dv* is *Low* **AND** *RD* **is** *NearT* **THEN** *saliency* $=s_{112}$
IF *He* **is** *Low* **AND** *Dv* is *Low* **AND** *RD* **is** *Far* **THEN** *saliency* $=s_{113}$

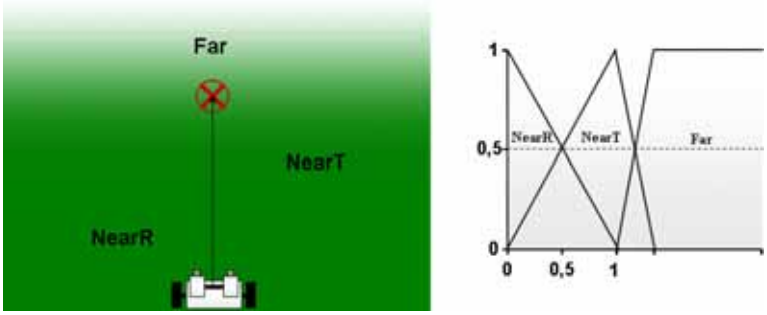


Fig. 6. Fuzzy sets of the *Relative Depth* property

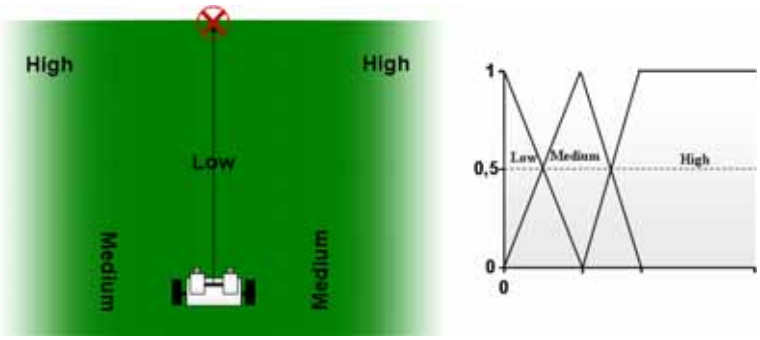


Fig. 7. Fuzzy sets of the *Deviation* property

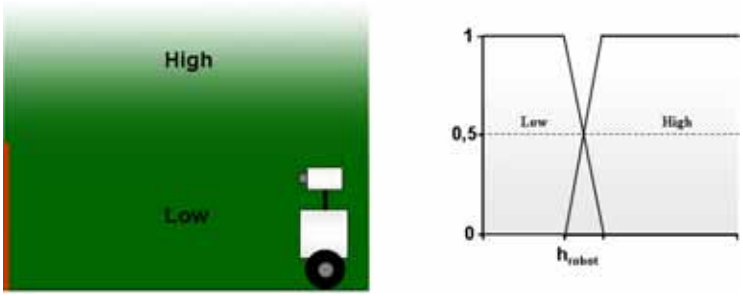


Fig. 8. Fuzzy sets of the *Height* property

Consequent values are set according to the excluding / selecting criteria of each rule. In the above example, every rule containing a premise with the form "*He* is *High*", "*Dv* is *High*" or "*RD* is *Far*" is considered an exclusion rule, since any region characterized by those properties does not interfere in the trajectory to the target. This group of regions should not

be selected as obstacles, which is expressed by assigning a negative saliency value to consequents of the corresponding rules (consequents in red of table 1). Remaining rules are selection rules whose positive output values must impose a saliency order. The specification of these output values is carried out by ordering rules from the most to the least adequate in the selection process. The most important regions for obstacle selection are those situated nearest to the robot. This implies that the maximum saliency value has to be associated to regions fulfilling the hypothesis "*He* is *Low* AND *Dv* is *Low* AND *RD* is *NearR*". On the other hand, regions at near positions to the straight line between the robot and the target are more important than those situated far away from the approaching line. It means that regions that meet the premise "*Dv* is *Low*" are prioritized over those that do not fulfil the premise. All these considerations allow imposing the appropriate order for consequents of selection rules, leading to the following output relations: $s_{111} > s_{112} > s_{121} > s_{122}$. According to these relations, rule consequents can be established. Also, other considerations, such as the relative relevance of one rule in relation to others, can be taken into account in order to provide additional running conditions.

		He					
		Low			*High*		
RD	*NearR*	s_{111}	s_{121}	s_{131}	s_{211}	s_{221}	s_{231}
	NearT	s_{112}	s_{122}	s_{132}	s_{212}	s_{222}	s_{232}
	Far	s_{113}	s_{123}	s_{133}	s_{213}	s_{223}	s_{233}
		Low	*Medium*	*High*	*Low*	*Medium*	*High*
		Dv					

Table 1. Tabular representation of the fuzzy rule system of an obstacle selector

Once the set of rules is defined, the system is operative to evaluate visual regions according to their properties. This evaluation results in a saliency map that stores the saliency value of each region. Through this map, system discards regions with negative saliency and orders remaining regions from the most to the least relevant. Final selection is achieved by choosing those regions whose saliency differ less than a certain percentage from the maximum saliency value. This makes possible to decide whether to obtain several candidate regions to get the focus of attention or to select only the most salient one.

Inhibition of return

When a target selector obtains several candidate regions where focus of attention should be directed to, a mechanism for distributing attention time among all of them must be included. The mechanism used is an inhibition of return (IR) process that is applied as the final selection step of each target selector. IR process allows displacing the focus of attention from currently attended region to the next most salient one throughout time. For this purpose, it maintains an inhibition map that represents the attentional fixation degree of each area of the visual field. According to both inhibition and saliency maps a winning region is finally selected to become the next focus of attention.

Global selection process

The simultaneous running of multiple target selectors requires including a global selector that decides which individually selected region gains the overt focus of attention at each moment.

Target selectors attend covertly to their selected regions. They request the global selector to take overt control of attention at a certain frequency that is modulated by high-level behavioural units. This frequency depends on the information requirements of the corresponding behaviour, so, at any moment, several target selectors could try to get the overt control of attention. To deal with this situation, global selector maintains a time stamp for each active target selector that indicates when to cede control to that individual selector. Every so often, the global selector analyses the time stamp of every target selector. The selector with the oldest mark is then chosen for driving the overt control of attention. If several selectors share the oldest time stamp, the one with the highest frequency acquires motor control. Frequencies of individual selectors can be interpreted as alerting levels that allow keeping a higher or lower attention degree on the corresponding target. In this sense, the described strategy gives priority to those selectors with the highest alerting level that require faster control responses.

2.4 Binocular fixation of the focus of attention

Overt control of attention is achieved by binocular fixation, centring the selected visual target in both images of the stereo pair. This process keeps attention focused on the target until the next one is selected.

Our proposal for binocular control is to divide the stereo fixation into two independent camera movements: a saccadic and tracking movement in one of the cameras and an asymmetric vergence movement in the other one. This separation results in a master-slave configuration that simplifies the global control. Moreover, it allows fixating visual targets with unknown spatial properties, so binocular focusing on the target can be achieved although attention is guided by monocular information (Enright, 1998).

Tracking control

In our master-slave configuration, the master or dominant camera is in charge of fixating and tracking the focus of attention. For this purpose, a hierarchical correlation-based method is employed using a multiscale prism representation of the target. This representation keeps information about target and its near areas, allowing false matching regions to be discarded according to target neighbourhood.

The proposed tracking method starts locating positions of the highest resolution level that exhibit high similarity to the target. If more than one position is found, a new comparison is done at the next resolution level. This comparison discards regions whose neighbouring areas do not match with proximal areas around target. If more than one similar region is found again, the comparison process is repeated at the next resolution level. Thus, the method proceeds by an ascendant search of the target position in the scale space until a winning region is obtained.

Figure 8 shows two frames of a sequence of tracking obtained while robot approaches a target. In this scene, three similar regions appear in the visual field. The one situated at the right constitutes the visual target. Images show the application of the hierarchical tracking method to both situations. For each level, regions of high similarity to the target are highlighted in green. Similar regions obtained from previous level that are discarded at a given level are marked in red. As it can be observed in this figure, the method employs an ascendant searching strategy that is maintained until a unique position of high similarity is found.

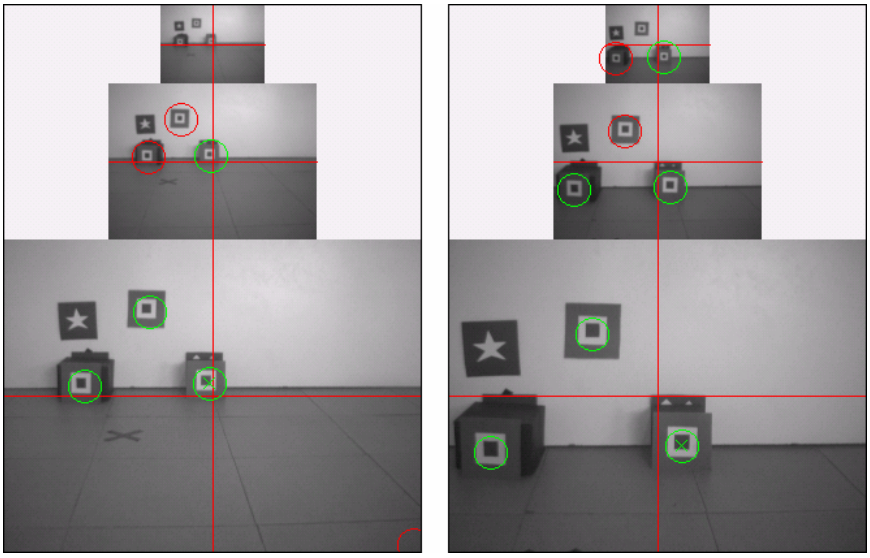


Fig. 8. Focus of attention tracking control

Vergence control

The function of an asymmetric vergence movement is to centre the visual target situated at the fovea of the dominant camera in the image of the slave camera. Hence, vergence control can be treated as a problem of maximization of the correlation coefficient measured between the central window of the dominant camera and the epipolar homologous window in the other one. Now then, the right size of that central window is dependent on the properties of the visual world, which makes it necessary to compute a new parameter: the size of the correlation window. To solve this problem, we introduce a hierarchical control method that carries out a descendent search in the scale space. This method is the inverse process used in tracking control. The reason for using a new strategy in vergence control is that the unknown size of vergence area impedes a reliable ascendant location of vergence position. In contrast, a descendent search provides both extension and position of the vergence area.

To deal with different potential extension of vergence area, the central image window of the dominant camera is represented at different resolution levels using a multiscale prism. Initial search is done at the lowest resolution level. It consists of locating the position inside the widest central image window that maximizes correlation with the corresponding image patch of the dominant camera. If a position of high similarity is found, the next level tries to find a more precise vergence position by searching in a neighbouring window of the previously obtained position. If no initial position can be found, the complete search is repeated at the lower level. This procedure is applied for each level until a final vergence position is obtained in the highest resolution level.

Figure 9 shows the proposed vergence method working in two real situations. Left images in (a) and (b) show the multiscale pyramidal representation of the image in the dominant camera. Central prism is delimited by the red squares of each level. Right images in (a) and (b) depict the multiscale representation of the image in the vergence camera. The search

window within each level is represented by a red rectangle. As shown in this figure, computing vergence position (marked with an "X") at every resolution level allows selecting the search window of the subsequent level. If a given level produces no result, the process maintains the width of the search window in the subsequent one, adapting its value to the corresponding resolution. This behaviour can be observed in the second level of image (b). Once the whole process is completed, vergence position is obtained in the highest resolution level.

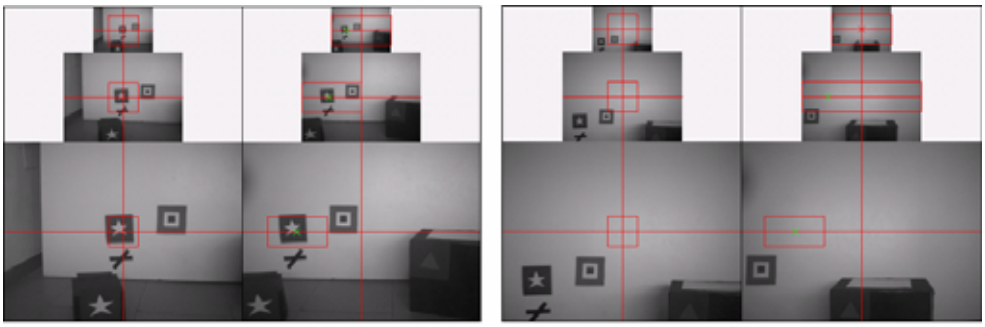


Fig. 9. Vergence control in real situations

3. Attention-based control

Intelligent robot control is carried out through pre-established links between the set of high-level behaviours and the visual attention system. Each high-level behaviour starts up an attentional control mechanism that provides a coherent visual behaviour according to its task-goals. This control mechanism is a specific target selector that responds to high-level activation and modulation by supplying the appropriate flow of visual information.

Links between attentional control components and high-level behaviours provide the necessary organization of the group of actions that solve a certain task. The attention system always guarantees the selection of a visual region that constitutes the unique visual input of every high-level behavioural unit. This sensory selection results in an effective execution of a subset of actions from the total of potential actions associated to active behaviours. The final sequence of actions, which is produced as result of the serialized attentional access to visual stimuli, forms a control dynamic that gives rise to the desired compound behaviour.

The proposed visual system models several kinds of attention that provide different high-level control dynamics:

- Bottom-up attention: selection of a focus of attention depends on properties of visual regions in the scene. If any region maintains properties that agree with selection criteria of active target selectors, the focus of attention shifts to that region. In this sense, the system provides a bottom-up control on non-searched stimuli, allowing the robot to react appropriately to unexpected events.
- Top-down attention: attention system is modulated from high-level processes through the activation of some target selectors that provide the appropriate visual behaviour for

the current situation. From this point of view, the acting context has a remarkable influence on the selection of a focus of attention, allowing global control to centre on task-relevant visual stimuli of the environment.

- Overt attention: visual attention ensures the selection of a focus of attention that acts as the unique source of visual information for high-level control components. This selection process allows high-level behaviours to produce appropriate reactions to the attended stimulus, avoiding any conflict among behaviours.
- Covert attention: coexistence of multiple target selectors allows keeping a "mental" focusing on several visual stimuli. It implies that it is possible to maintain an alerting state that allows displacing attention at any moment to any visual target for achieving different behavioural goals.

These control aspects can be observed in the following example of attention-based control for solving a task of navigation following a set of landmarks. The system is formed by the components and connections depicted in figure 10. Each behavioural component defines an attentional dynamic by activating and modulating a specific target selector. Three high-level behaviours take part in this system: *LANDMARK-GUIDED NAVIGATION*, *GO TO POINT* and *EXPLORE*. They modulate three attentional selectors of landmarks, obstacles and unattended regions, respectively.

Two situations are considered in this system. The first one occurs when the location of the landmark that the robot must approach is known. In this case, active behaviours are *LANDMARK-GUIDED NAVIGATION* and *GO TO POINT*. The first one activates a landmark selector and sends the aspect information of the landmark to be located. Attentional frequency on landmark is modulated by the *NAVIGATION* behaviour with an inversely proportional value to distance between landmark and robot. Thus, the less the distance to target position is, the greater the degree of attention on landmark is dedicated. When the *NAVIGATION* behaviour receives visual information that match up with the expected landmark, it sends to the *GO TO POINT* behaviour the target position in order to make the robot reach that position. To accomplish its goals, this last behaviour is connected to an obstacle selector, which allows shifting attention to regions situated close to the trajectory between the robot and the landmark. The *GO TO POINT* behaviour interprets incoming visual information as the nearest region that could interfere in the approach to landmark. Thus, it performs different actions according to the position of the attended region and the goal position. Firstly, if both positions are near enough, it is considered that the attended region is part of the target area and, consequently, it responds by approaching that region. However, if there is some distance between the target and the selected region, it is assumed that attention is focused on an obstacle that must be avoided. To get low reaction times in view of unsafe situations, obstacle selector is modulated with an activation frequency that is proportional to robot velocity.

The second situation that is considered arises when either *NAVIGATION* behaviour gets no initial information about landmark or when, after locating it, no new information is received for a long time. To deal with these situations, *NAVIGATION* behaviour deactivates *GO TO POINT* behaviour and activates *EXPLORATION*. The function of this new behaviour is to get visual access to regions of the environment that have not been recently attended. This function is accomplished through its connection to a selector of unattended regions of the

visual system. This selector maintains a purely bottom-up dynamic that, together with the inhibition of return mechanism, leads to the visual exploration of the scene. Simultaneous execution of landmark and unattended region selectors give rise to the localization of the landmark that is being sought as soon as it appears in the visual field. As a consequence, landmark gains the focus of attention and system returns to the first described situation. Once the goal position is reached, *NAVIGATION* behaviour reinitiates the whole process by sending to its target selector the description of a new landmark.

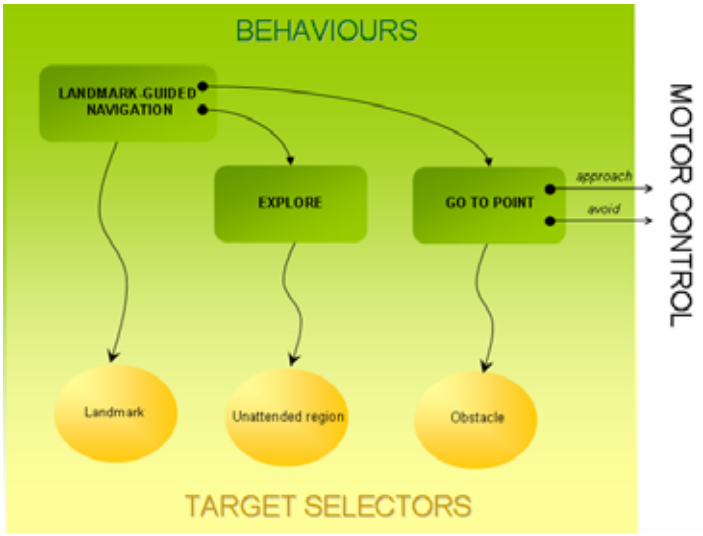


Fig. 10. Control components of a landmark-guided navigation task

The two situations described above are solved using a multi-target attentional control that exhibits several kinds of attention, providing different high-level control dynamics. On one hand, the system provides top-down attentional control on landmarks, allowing high-level processes to decide whether to approach a target position or to explore the environment to locate that position. When a landmark is located, the focus of attention alternates between landmark and potential obstacles, keeping overt control on one of the targets and covert control on the other one. This allows robot to react appropriately to obstacles while landmark position can be quickly recovered. In the exploring situation, searching is carried out through a bottom-up control where no target specifications are considered, so exploration depends exclusively on external world features. As in the previous situation, simultaneous covert control in landmark selector during exploration allows an immediate focusing on target once it is visualized.

4. Hardware and software architectures

Our attention-based control model has been tested in a five degrees of freedom mobile robot (three d.o.f. for the head and two for the base) that is endowed with a stereo vision head. It is equipped with digital PID controlled servos and two ISight Firewire cameras (Figure 11).

This model of robot has been developed in our Laboratory and is widely used for prototyping and algorithm testing.

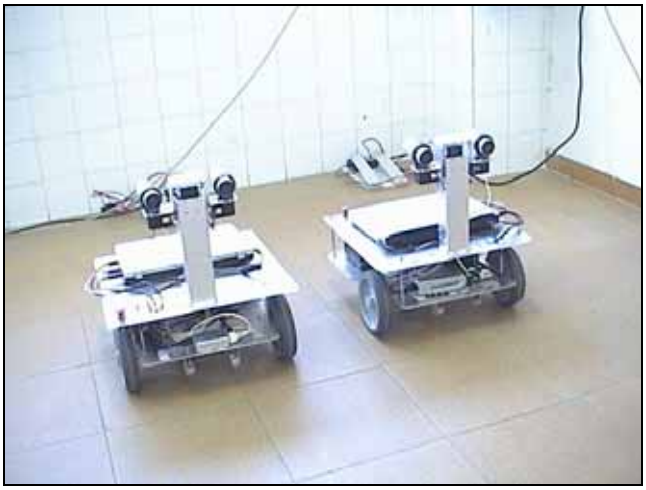


Fig. 11.Robots use for testing our attention-based control approach

To obtain a real-time modular implementation, we have designed a component-oriented software architecture that exhibits an efficient and flexible structure. Every component of the attentional system depicted in figure 2, as well as high-level behaviours and other support elements, has been implemented as independent software components following the principles of components-based software engineering (Szyperski, 1998; He et al., 2005). The Internet Communication Engine (Ice) middleware platform (ZeroC, 2007) has been used for connection among components. Thus, each component has been written as an independent C++ application that includes Ice objects and proxies to communicate with other components. The resulting architecture is characterized by a flexible structure, a high degree of reusability of its elements and real-time running capabilities (Bachiller et al., 2007).

5. Experiments

The attention-based control model has been tested through several real experiments using our testing mobile robot for solving autonomous navigation tasks. In this section, results of two experiments are presented. Several frames of the navigation sequences are shown. For every frame, two different views of the corresponding scene are presented. The first one is a view from the robot represented by the two images of the stereo pair. Through these images, the focus of attention that is selected at every moment can be appreciated. The second one is a general view obtained from a camera situated in the ceiling of the room. In this second group of images, the attended area is highlighted in red to facilitate the association between the two sets of images.

The control system employed in these experiments is the same that was described in section 3. Navigation is guided by a landmark or a sequence of landmarks that the robot must locate and approach while potential obstacles in the way are avoided.

5.1 Experiment 1: approaching a landmark with detection and avoidance of obstacles

In this first experiment, the robot must approach a landmark (star) situated in a wall in front of it. Two boxes stand in the way to the target position, so the robot must avoid them by going through the free space between them. On the floor, behind these two obstacles, several planar objects have been placed. The robot has to determine that these objects do not interfere in the trajectory to the landmark, since it can drive over them. Planarity and orientation properties of regions provide this distinction between obstacles and ground areas. Detection of ground areas is depicted in the sequence of images of the stereo pair by green regions. This process is only carried out for regions situated near to the focus of attention, so not all the planar regions placed on the floor are detected as ground regions.

Figures 12-15 show the result of this experiment. Initially (a), the robot is attending the landmark, causing the activation of the *GO TO POINT* behaviour. Consequently, attentional control responds by fixating the focus of attention on the first obstacle (b), which allows behavioural control to produce suitable avoiding actions (c-e). Proximity to the second obstacle gives rise to a shift of attention towards the corresponding area (f). This attentional change causes variations in the steering of the robot allowing the avoidance of this second obstacle (f and g). Once the robot is appropriately oriented, attention shifts again to regions near the target and new approaching actions follow (h). After some time, landmark selector recovers attentional control producing the updating of the target position (i). This updating causes new changes of attention that directs to proximal areas between the robot and the landmark. Specifically, attention focuses on plane regions of the floor (j), which allows high-level processes to detect them as ground regions. This detection gives rise to the correct treatment of the new situation that results in advancing movements towards target position (j-l). After several instants, attention is again centred on the landmark (m) making the robot to approach it in a more precise way (n-o) and finally reach the goal position (p).

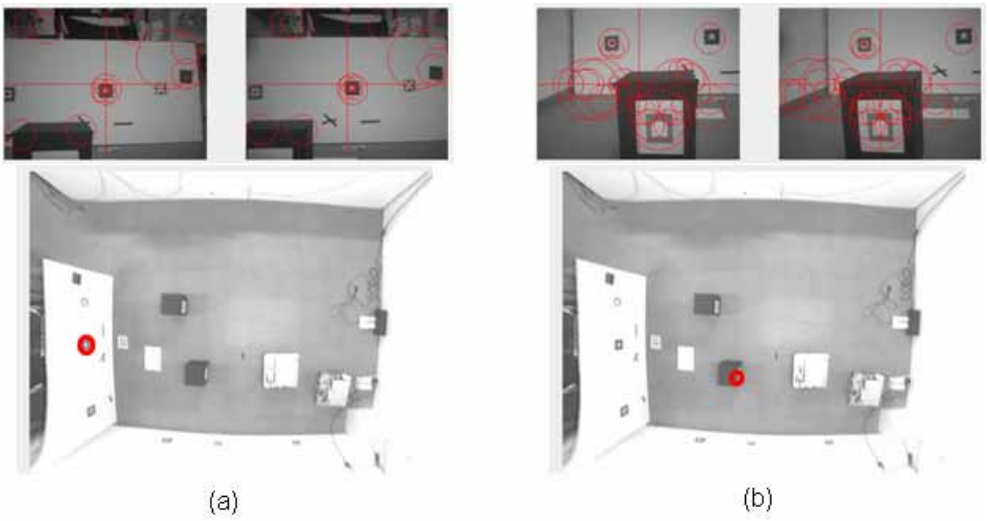


Fig. 12. Experiment of navigation avoiding obstacle (first part)

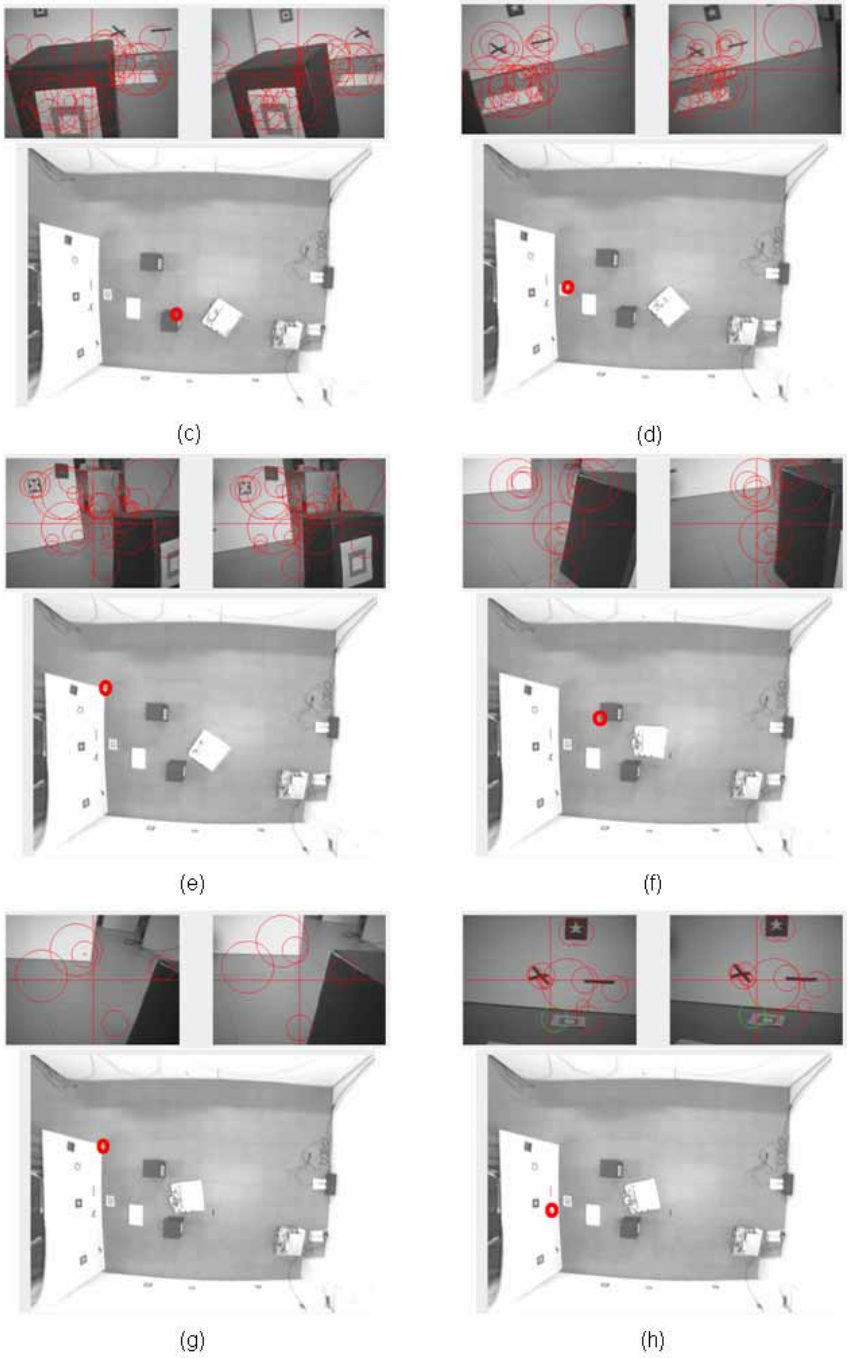


Fig. 13. Experiment of navigation avoiding obstacle (second part)

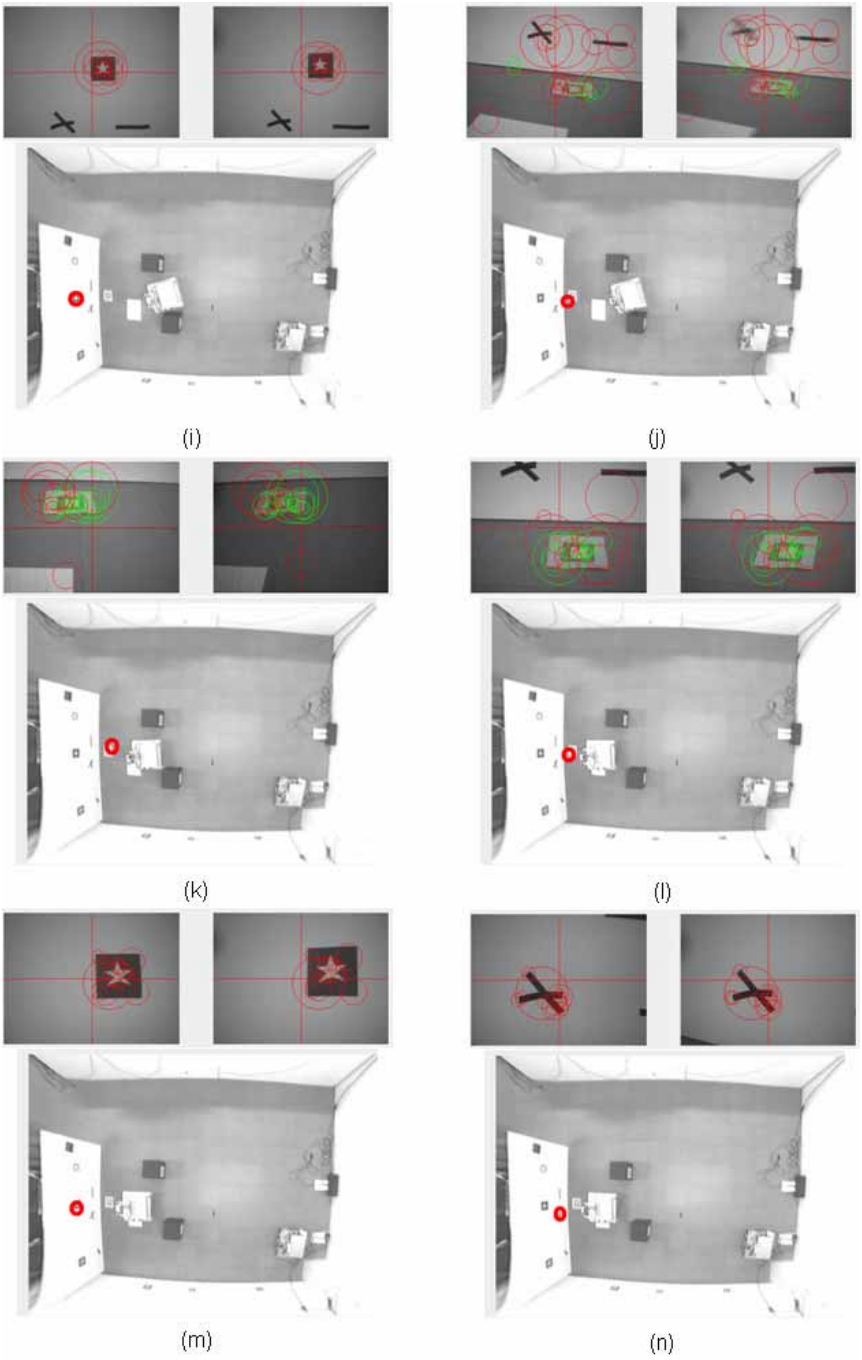


Fig. 14. Experiment of navigation avoiding obstacle (third part)

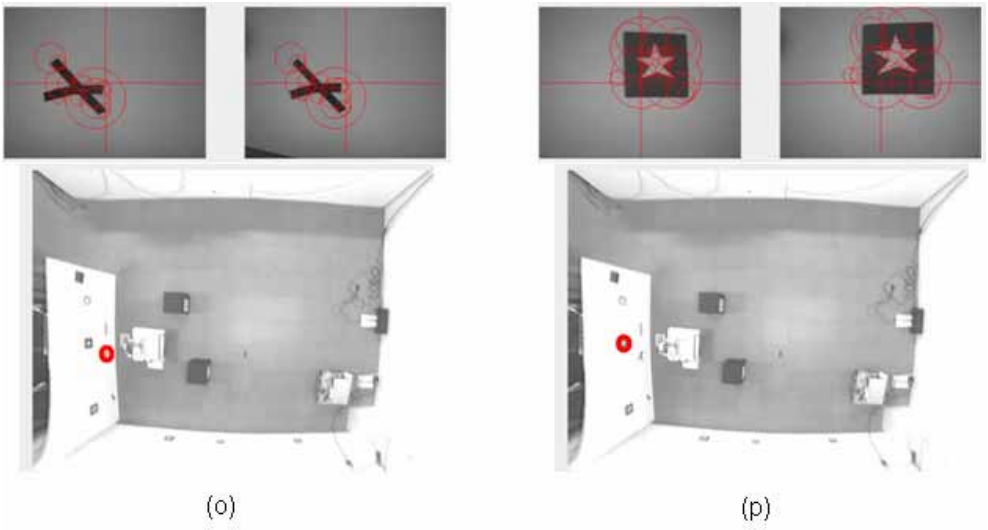


Fig. 15. Experiment of navigation avoiding obstacle (fourth part)

5.2 Experiment 2: navigation following a sequence of landmarks

The aim of this second experiment is to test the robot behaviour when navigation is guided by a sequence of landmarks.

The scenario of this experiment consists of two landmarks that the robot must sequentially reach. First landmark is formed by three concentric squares of different colours. The second one is the same landmark of the previous experiment. Figures 16-18 show results obtained in this test. At the beginning of the experiment (a), attention is fixated on the first landmark. Absence of obstacles allows a continuous advance towards the target position (b-d). Once this position is reached (d), the *NAVIGATION* behaviour reprograms landmark selector to locate the second landmark. Since there is no visual region that fulfils appearance properties of this new landmark, an exploratory behaviour is activated. Attentional control associated to this exploratory activity provides the visual access of the second landmark (g), once other visual areas have been explored (e-f). Since that moment, first landmark is conceived as an obstacle that robot must avoid for reaching the new target position. Obstacle selector provides this interpretation of the situation by centring attention on regions associated to the first landmark (h). Once the robot is appropriately oriented to avoid collisioning with the obstacle (i-j), attentional control focuses on regions situated near to the target. As a consequence, behavioural control produces approaching actions (k-l), which are interrupted when landmark selector recovers attentional control (m). After updating the target position, new approaching actions follow (n-o), allowing the robot to eventually reach the desired position (p).

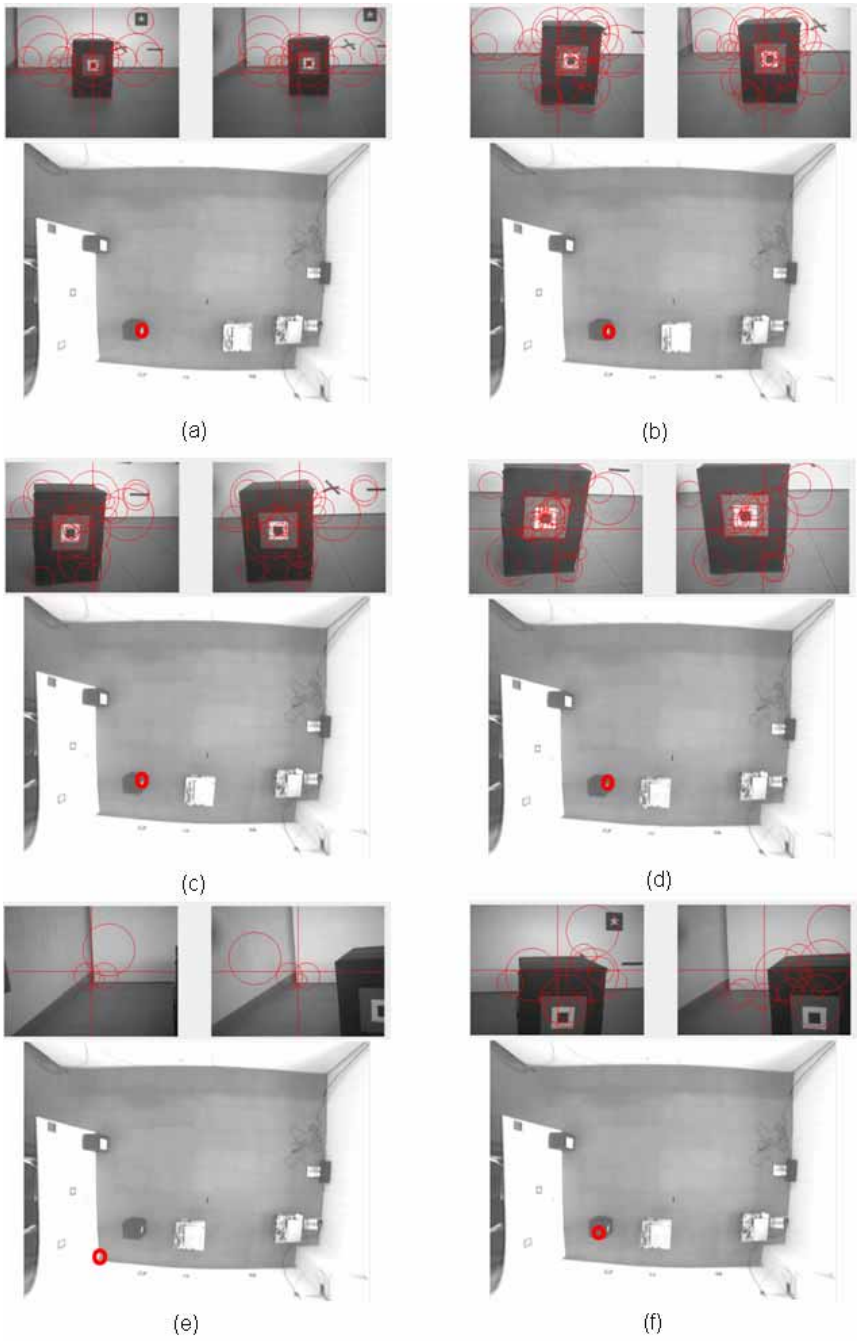


Fig. 16. Experiment of navigation following a sequence of landmarks (first part)

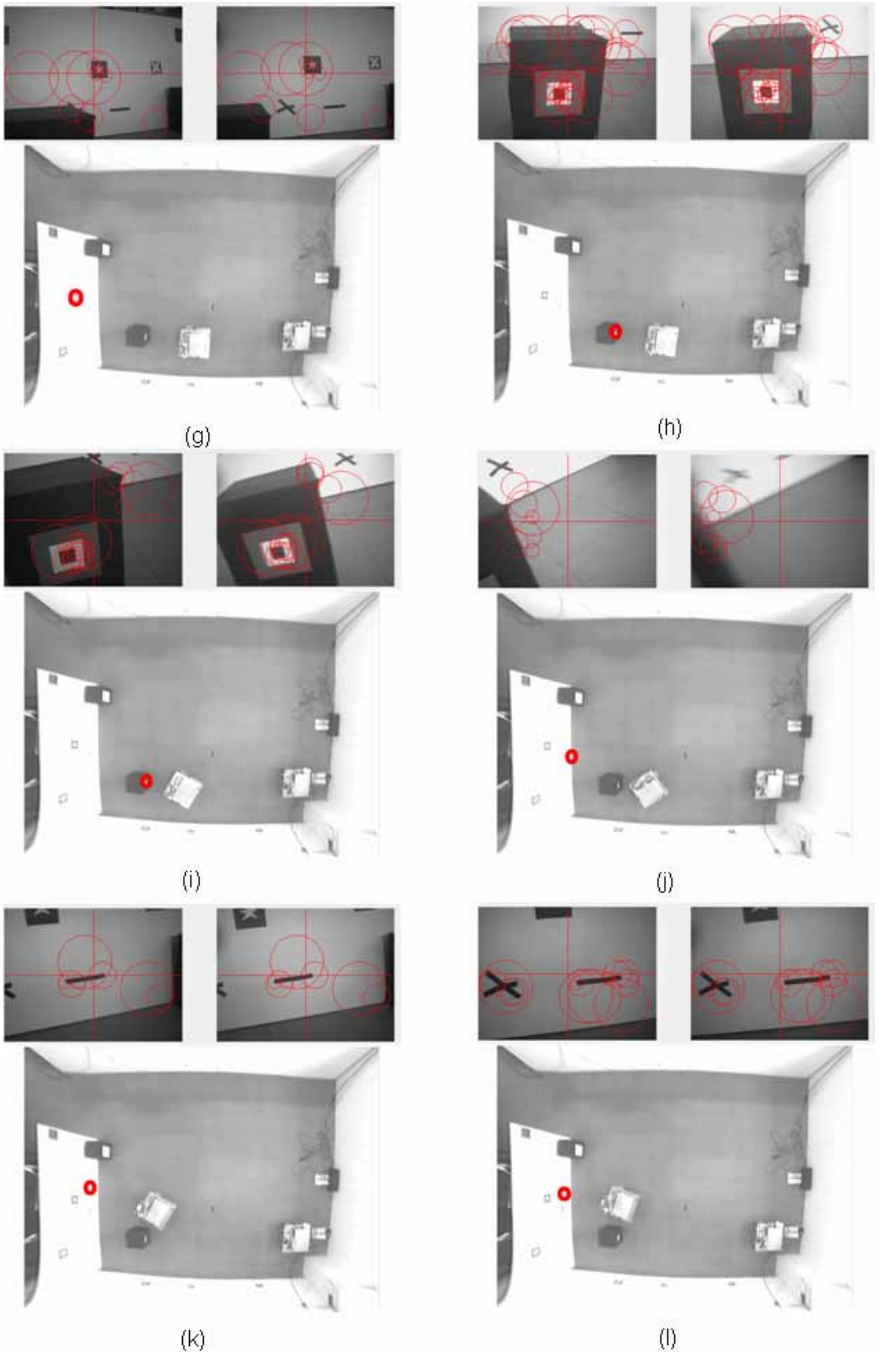


Fig. 17. Experiment of navigation following a sequence of landmarks (second part)

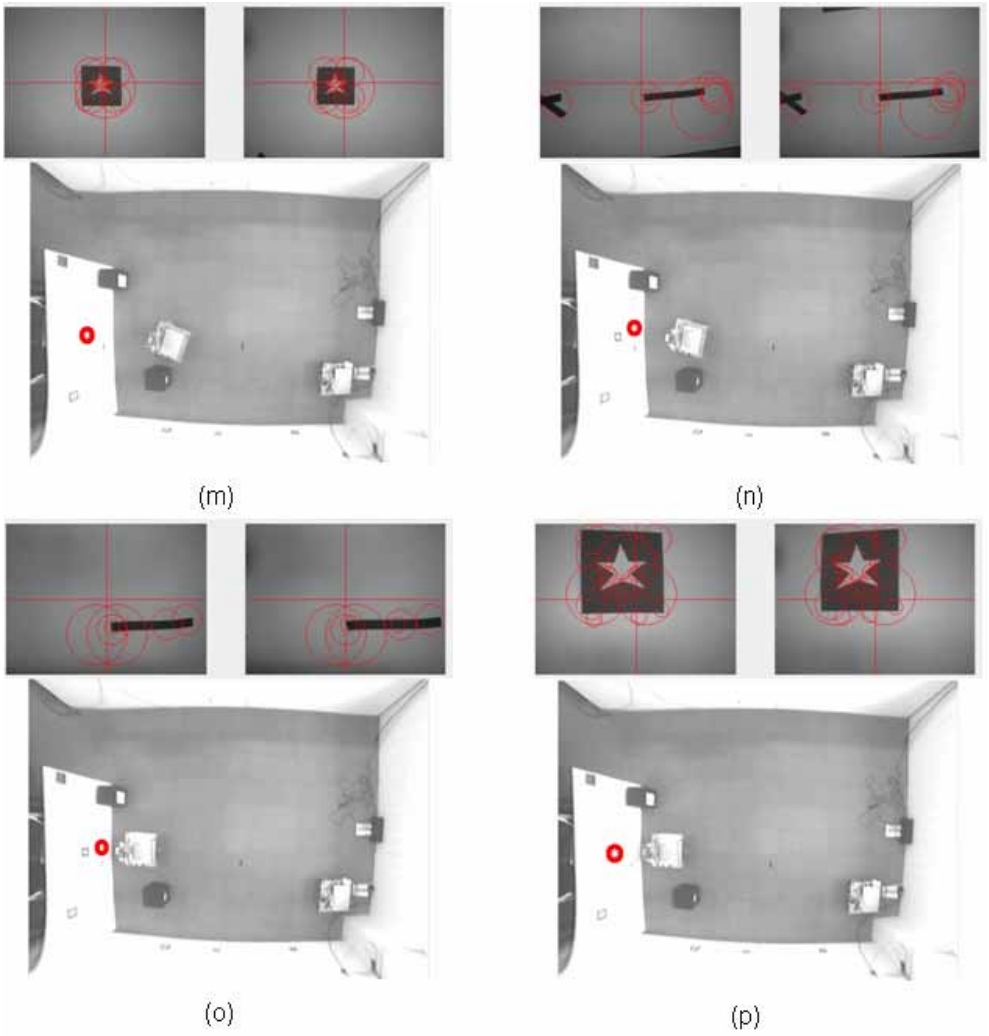


Fig. 18. Experiment of navigation following a sequence of landmarks (third part)

6. Conclusions

In this chapter, a novel computational model of visual attention based on the selection for action theory has been presented. In our system, attention is conceived as an intermediary between visual perception and action control, solving two fundamental behavioural questions:

- *Where to look?* Attention drives perceptual processes according to action requirements. It selects relevant visual information for action execution.
- *What to do?* Attentional selection limits potential actions than can be performed in a given situation. In this sense, actions are modulated according to perceptual result of attentional control.

These ideas give rise to what we have called attention-based control model. The proposed global system is a behavioural architecture that uses attention as a connection means between perception and action. In this model, behaviours are control units that link attention and action. They modulate the attention system according to their particular goals and generate actions consistent with the selected focus of attention.

Links between attention and action become effective through several properties of the visual attention system:

- **Top-down specifications**: attention is modulated from high-level behaviours according to their acting requirements. Hence, attentional selection is strongly influenced by action.
- **Distributed control**: attentional control is distributed between multiple individual control units that keep several visual targets simultaneously. This control scheme allows attention to be modulated from multiple behaviours with different information requirements.
- **Overt and covert attention**: the system provides overt and covert control of attention, allowing attentional selection to alternate among several visual targets according to external world features and internal robot requirements.

All these properties make possible to define proper interacting dynamics between behavioural and attentional controls that result in a global compound behaviour. Moreover, they provide the necessary coordination among individual behaviours that leads to an effective autonomous control.

The whole system has been designed as a component-based software architecture that provides a real-time modular implementation. Real experiments have been conducted in a mobile robot for solving navigation tasks. Results show good performance of the system in real situations, providing initial groundings for deeper explorations of links between attention and action in robot control.

7. References

Allport, A. (1987). Selection for action: some behavioural and neurophysiological considerations of attention and action, In: *Perspective on perception and action*, H. Heuer and A. Sanders (Ed.), Erlbaum

Bachiller, P.; Bustos, P.; Cañas, J.M. & Royo R. (2007). An experiment in distributed visual attention, *Proceedings of the 4th International Conference on Informatics in Control, Automation and Robotics*

Bandera, C. & Scott, P. (1989). Foveal machine vision systems, *IEEE International Conference on Systems, Man and Cybernetics*, pp. 596-599

Broadbent, D.E. (1958). *Perception and communication*, Pergamon Press, New York

Enright, J. (1998). Monocularly programmed human saccades during vergence changes?, *Journal of Physiology*, Vol. 512, pp. 235-250

Frintrop, S.; Backer, G. & Rome E. (2005). Goal-directed search with a top-down modulated computational attention system, *Lecture Notes in Computer Science*, W.G. Kropatsch, R. Sablatnig and A. Hanbury (Ed.), Vol. 3663, pp. 117-124, Springer

He, J.; Li, X. & Liu, Z. (2005). Component-based software engineering: the need to link methods and their theories, *Proceedings of ICTAC 2005, Lecture Notes in Computer Science*, Vol. 3722, pp. 70-95

Itti, L. & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention, *Vision Research*, Vol. 40, pp. 1489-1506

LaBerge, D. (1995). *Attentional processing*, Harvard University Press.

Lazebnik, K.; Schmid, C. & Ponce, J. (2005). A sparse texture representation using local affine regions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, pp. 1265-1278

Lindeberg, T. (1998). Feature detection with automatic scale selection, *International Journal of Computer Vision*, Vol. 30, No. 2, pp. 77-116

Lowe, D. (1999). Object recognition from local scale-invariant features, *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 2, pp. 1150-1157

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, Vol. 60, pp. 91-110

Mikolajczyk, K. & Schmid, C. (2001). Indexing based on scale invariant interest points, *Proceedings of the Eighth IEEE International Conference on Computer Vision*, Vol. 1, pp. 525-531

Mikolajczyk, K. & Schmid, C. (2005). A performance evaluation of local descriptors, IEEE Transactions on Pattern Analysis and Machine Intelligece, Vol. 27, pp. 1615-1630

Milner, A.D. & Goodale, M.A. (1995). *The visual brain in action*, Oxford University Press.

Navalpakkan, V. & Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed, *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, pp. 2049-2056.

Neumann, O.; van der Hejiden, A.H.C. & Allport, A. (1986). Visual selective attention: introductory remarks, *Psychological Research*, Vol. 48, pp. 185-188

Szyperski, C. (1998). *Component software: beyond object-oriented programming*, ACM Press / Addison-Wesley Publishing Co.

Takagi, T. & Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control, *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 15, pp.116-132

Torralba, A.; Oliva, A.; Castelhano, M. S. & Henderson, J. M. (2006). Contextual guidance of eyes movements and attention in real-world scenes: the role of global features in object search, *Psychological Review*, Vol. 113, No. 4, pp. 766-786

Tsotsos, J.; Culhane, S.M.; Winky, W.; Lay, Y.; Davis, N. & Nuflo, F. (1995). Modeling visual attention via selective tuning model, *Artificial Intelligence*, Vol. 78, No. 1-2, pp. 507-545

Zadeh, L. (1965). Fuzzy sets, *Information and Control*, Vol. 8, pp. 338-353
ZeroC (2007). Internet communication engine, URL: *http://www.zeroc.com/ice.html*